Bayes Rules – A Bayesian-Intuit Approach to Legal Evidence

Helena Miriam Likwornik

Doctor of Philosophy

Department of Philosophy
University of Toronto

2011

Abstract

The law too often avoids or misuses statistical evidence. The solution I propose is to accept Bayesianism as a normative framework and to encourage education about common conceptual errors involving statistical evidence as well as techniques to limit their occurrence. I address three types of objections to using Bayesianism in the legal context. The objection based on the irrelevance of statistical evidence is fundamentally incoherent in its failure to identify most evidence as statistical. The objection to the incompleteness of a Bayesian approach in accounting for non-truth-related values places legitimate limits on the use of Bayesianism in the law but in no way undermines it. Lastly, criticisms that rest on misunderstandings of the meaning and manipulation of statistical evidence are best addressed by accepting Bayesianism and presenting statistical evidence in ways that encourage correct understanding.
Acknowledgments

It is with great pleasure that I thank the many people who made this thesis possible.

I would like to thank the members of my thesis committee, Joseph Heath, Arthur Ripstein and Bruce Chapman for so many great suggestions, stimulating conversations and encouraging words. I feel incredibly fortunate to have had the benefit of their involvement.

The U of T philosophy department and law faculty have been great places to pursue my passions.

I have my parents to thank for instilling in me a love of knowledge and for their constant support.

I am extremely grateful to Andrew Ogilvie for his boundless patience, faith in me, and assistance with matters both big and small.

Any errors and inadequacies that remain are my own.
# Table of Contents

## Acknowledgments

Table of Contents

List of Tables

List of Figures

Chapter 1 Bayes Rules

1 Bayes Rules

1.1 Introduction

1.2 Statistical Evidence

1.3 A Brief Overview of the Debate

1.4 Why Bayesianism?

1.5 Bayesian Updating is What We Do (Most of the Time)

1.6 Why a Normative Model is Needed

1.7 The Problem: Case Illustrations of Trouble with Numbers

1.7.1 Statistical Evidence: Problematic Avoidance

1.7.2 Statistical Evidence: Misuse

1.8 Summary

Chapter 2 What is Legal Evidence? Constraints on Assertion and Justification

2 What is Legal Evidence?

2.1 Assertion and Justification: To Each Context, a Different Standard of Caution

2.2 Verdict Accuracy vs. Verdict Acceptability: Constraints on Assertion and Justification in the Legal Context

2.3 Is There a Unifying Theory of Legal Evidence?

2.4 Cohen’s Non-Mathematical Approach: A Critique
Chapter 3 Bayesianism: A Legitimate Epistemological Approach to Statistical Evidence

3.1 Introduction to the Theorem

3.2 The Derivation of Bayes’ Theorem

3.3 Interpreting Probabilities

3.4 It is Plain Silly to Ignore What We Know: The Role of Prior Probabilities

3.5 Encouraging Good Bayesian Reasoning

3.6 A Challenge to Bayesianism as a Normative Model

3.7 An Application to Medicine: Bayesian Analysis of Clinical Data

3.8 Two Applications to the Law: Identification Evidence and Base Rates

Chapter 4 The Probability “Paradoxes”: Need Legal Intuition and “Good” Bayesian Results Clash?

4.1 Description of the “Paradoxes”

4.1.1 Gatecrasher “Paradox”

4.1.2 Green Cab or Blue Cab “Paradox”

4.1.3 Todhunter/Lottery “Paradox”

4.2 Common Responses to the “Paradoxes”

4.3 Koehler and Shaviro’s Resolution to the Impasse

4.4 An Alternative Resolution Based on Corroborating Evidence

4.5 Simple Heuristics as a Possible Alternative

4.6 A Possible Place for Base Rates

Chapter 5 Medical Evidence: Constraints on Assertion and Justification

5.1 The Role of Priors in Medical Research: Some Examples
5.2 Medical Evidence: Constraints on Assertion and Justification in the Medical Context 124
  5.2.1 Medical Study Design: Randomization and the Importance of Priors .......... 126
  5.2.2 Theories about the Practice of Medicine .......................................... 135
  5.2.3 Actual Medical Practice ...................................................................... 145

Chapter 6 Lessons to Learn .............................................................................. 154
6 Lessons to Learn ......................................................................................... 154
  6.1 Methods to Adopt ................................................................................ 155
    6.1.1 The Power of Modes of Presentation ............................................. 156
  6.2 Approaches to Problems from Chapter 1: Addressing Common Fallacies .... 158
    6.2.1 Statistical Evidence: Addressing Problematic Avoidance ............... 159
    6.2.2 Remembering to “Think of the Opposite” ....................................... 162
  6.3 Implications .......................................................................................... 171
    6.3.1 Judges vs. Juries ........................................................................... 171
    6.3.2 Experts .......................................................................................... 172
    6.3.3 Identifying the True Locus of Danger ............................................. 177
  6.4 Recommendations ................................................................................. 183
    6.4.1 Clearly Define the Limits of Expert Evidence ................................. 186
    6.4.2 Consider Warnings Against Common Misunderstandings .............. 189
  6.5 Concluding Remarks ............................................................................. 190

Bibliography ..................................................................................................... 192
List of Tables

Table 1: Definition of Symbols for Breast Cancer Test Example ............................................ 69

Table 2: Sensitivity and Specificity .......................................................................................... 105

Table 3: Mapping the Case of Lucia de B. ............................................................................ 169
List of Figures

Figure 1: Normal Distribution Curve ................................................................. 19

Figure 2: Example of a Probability Pie ............................................................. 151

Figure 3: Probability Pie – Factors Affecting Global Warming ......................... 160

Figure 4: Probability Pie – Couples with and without All Traits ....................... 165
Chapter 1
Bayes Rules\(^1\)

For the rational study of the law the blackletter man may be the man of the present, but the man of the future is the man of statistics… \(^2\)

1 Bayes Rules

1.1 Introduction

The problem as I see it is this: statistical evidence is ever-more pervasive in our world and the law too often avoids or misuses it. The solution I offer to this problem involves both a diagnosis and proposed treatment plan. The diagnosis is the absence of a normative framework for integrating statistical evidence into legal reasoning. That is, no one model has achieved clear acceptance in establishing the standard for the correct manipulation of statistical evidence in the context of the law. The solution I suggest is as follows: (i) accept Bayesianism as a normative framework against which errors in considering statistical evidence may be evaluated; and (ii) encourage education within the legal community and amongst expert witnesses about common conceptual errors involving statistical evidence as well as techniques to limit their occurrence.

What follows is a presentation of support for this proposed solution. After primers on both the nature of legal evidence and Bayesianism, I tackle a series of “paradoxes” that are often used to dispute the relevance of Bayesianism to the law. I proceed to demonstrate that these paradoxes are far from inevitable and that Bayesianism and legal intuition can be aligned. This demonstration is essential to the proposal that Bayesianism can serve a useful role in the legal approach to statistical evidence; it is only if judges and juries can see and grasp the correctness of the Bayesian approach that it offers a plausible solution to the problem. A comparison is

---


\(^2\) Oliver Wendell Holmes Jr “The Path of the Law” (1897) 10 Harv L Rev 457. The term “blackletter man” is a reference to blackletter law, whereby the law is set down in specific concrete rules.
drawn to the medical community in which Bayesianism is accepted at a normative level while failures in application at the level of practice are acknowledged. The response to this in the medical community has been to increase education in numerical literacy for physicians. The law can learn something from this approach both in principle and at the level of practical tools that aid in understanding basic probability. Ultimately, I look at some specific applications to the law and conclude with recommendations both in relation to education and the presentation of evidence that can aid in putting statistical evidence to good legal use.

In the discussion that follows, formulas will be used to illustrate the mathematical support for defining certain errors as errors. That is, mathematics will play a normative role as arbiter of what constitutes the correct manipulation of numerical evidence. This is as it should be. Numbers are after all the mathematician’s domain. At the same time, no mathematical facility is required to understand the conceptual corrections that I will describe. To the contrary, I will try to show that while certain conceptual errors involving numerical evidence are common, with the right explanation, they can be corrected in a way that makes intuitive sense. That this is the case is essential to my project; it would be unreasonable to expect that judges, and even more so juries, become comfortable manipulating mathematical formulae. Instead, math will be used to justify applying the term “error” to certain common modes of thinking and to support alternative presentations of information that yield different – though still accessible – results. In this way, those with greater faith than mathematical ability may safely gloss over the math.

The structure of the chapters to come is as follows: this chapter will set out the problem, describing cases in which mathematical concepts were either avoided or misused. My hope is that the illustrations that follow will present a convincing case that the status quo cannot continue; some normative model is required. The second chapter discusses legal evidence more generally and provides a context in which to understand the role and use of statistical evidence in the legal decision-making process. Chapter three is devoted to the basics of Bayesianism and the idea that Bayesian mathematics provides a good candidate for a normative model against which the law can distinguish right from wrong when it comes to competing conceptual frameworks. Chapter four is devoted to fending off potential criticisms of the Bayesian normative model for the law: well-known “paradoxes” of probability that are designed to show that Bayesianism and sound legal intuition cannot coexist are described; they are then dispelled by demonstrating that properly understood, Bayesianism and intuition can align. Critics of Bayesianism generally
commit one of three errors: (i) they claim the irrelevance of statistical evidence at the same time that they presuppose it; (ii) they fail to separate the legitimate invocation of non-truth-related trial process values from errors in factual determination; or (iii) they misunderstand the correct application of statistical principles. Chapter five juxtaposes the medical profession with the law and encourages comparison: that is, while medical practitioners fall prey to the same kinds of conceptual failures we see in legal discourse, these disparities are not generally used to criticize Bayesianism as a relevant normative model but rather to inform physicians’ education. The final chapter applies the lessons learned in medicine to the law and suggests some ways in which we may improve the law’s approach to statistical evidence through education, evidence presentation, and jury charges. Other possible implications for the structure of the legal system are also discussed.

1.2 Statistical Evidence

Since the use of “statistical” evidence in the law is my topic, a few words should be said about what I mean by the term. There is a sense in which all evidence is, or at least could be, statistical in nature. That is, our world is fundamentally uncertain and any statement about it is inherently probabilistic. I may state that “there is a table in front of me” but there is some probability inherent in that statement. Even if I am awfully confident about this fact, if I am to be entirely accurate, I must admit that there is some possibility, no matter how remote, that I “see” a table that is not there. I may, for example, be hallucinating or experiencing an optical illusion. My statement becomes “statistical” the moment I report it in the following way: “I am 99.9% certain that there is a table in front of me.” For our purposes then, the term “statistical” evidence refers to any evidence that makes the possibility of error explicit. The consequence of such explicitness is that it allows for a methodical analysis and combination with other similarly explicit evidence. More will be said later about the rules for such methodical combination. At this point I will provide just one example to illustrate the point. If the statements “I am 99.9% certain that there is a table in front of me” and “I am 99.9% certain that it is blue” are independent, they should be multiplied. When 99.9% and 99.9% are multiplied, they generate a 99.8% probability that there is a blue table in front of me. This also makes intuitive sense: I should have slightly more doubt in stating that “there is a blue table in front of me” than in merely stating that “there is a table in front of me”. In claiming that “there is a blue table in front of me” I am making two statements about the world. Since each claim about the world has a
small possibility of error, the more I say, the greater are the chances that I will be wrong about something. Another way to think about it is this: if the bug screens on all my windows have a tiny tear, with each window that I open, the likelihood that some bug will get in somewhere increases.

Uncertainty about the world is not, however, the only way in which evidence is generally probabilistic. There are a number of ways in which this may be so. One is that evidence is always incomplete. The second and related reason is that evidence is commonly inconclusive; incomplete evidence does not necessitate that it will be inconclusive but inconclusive evidence is by definition incomplete. Evidence is also often ambiguous. For example, in the case of an alleged sexual assault, the complainant may appear upset upon encountering the accused. This evidence of distress is ambiguous: it could either support the claim of assault – it would be upsetting to confront your assailant – or the defense of false allegation, since it would also be upsetting to confront the person you have falsely accused who may be expected to react with anger or pleas for recantation. Bodies of evidence are often dissonant, as in the “he said”/“she said” accounts that are typical in sexual assault cases, in which the occurrence of sexual activity is often agreed upon while the accounts differ on the question of its nature: consensual versus forced. Furthermore, evidence tends to come from sources that lie on a continuum of credibility. The eyewitness testimony of a disinterested, respected citizen who observed an event from only a few feet away is considerably more credible than the testimony of a known criminal who has something either to gain or lose from the outcome of the trial.

In this way, all evidence is probabilistic, but only evidence that wears its risk of uncertainty on its sleeve is statistical in nature. My focus will be on statistical evidence, encouraging its proper use in legal decision making, though my arguments for accepting a Bayesian approach will draw on the probabilistic nature of evidence more generally.


1.3 A Brief Overview of the Debate

For well over thirty years now, there has been an ongoing debate about the use of overtly probabilistic evidence and methods in the law. The case of *People v. Collins* in the late 1960s, in which a U.S. appeal court overturned a criminal conviction citing flawed reasoning with probabilities, stimulated discussion about the role of statistical evidence in the law. This case will be discussed in greater detail below. Shortly thereafter, Michael Finkelstein and William Fairley published a seminal paper entitled “A Bayesian Approach to Identification Evidence”. The authors proposed an approach to calculating identification probabilities that was roundly criticized in Tribe’s piece, “Trial by Mathematics: Precision and Ritual in the Legal Process”. Tribe argued against any use of numerical probabilities in the law on the basis that, (a) so long as judges and jurors could be assumed innumerate, they should not be addressed in a language they do not understand; (b) mathematical arguments are likely to be prejudicial because quantifiable variables will wrongly be given pride of place over non-quantifiable factors; and (c) it is politically improper to quantify such things as the risk of convicting an innocent person. Tribe’s work was followed by an important paper by Richard Lempert in 1977 entitled “Modeling Relevance” arguing for the usefulness to legal decision makers of probabilistic analyses of evidential issues. According to Lempert, probability is entirely useful to the assessment of legal relevance: evidence is relevant if it allows us to adjust our probability assessment of a fact in question.

In 1982 a conference on probability and the law was held in Durham, England. This was partly in response to a book by L.J. Cohen published in the late 1970s that will be discussed in some detail in chapter two. In 1986 and 1990, Peter Tillers held conferences on probability and evidence in law at the Boston University and Cardozo Law Schools. Since then, the debate has

---


continued. Journals such as *Law and Probability* are devoted to the topic and regularly feature variations on the same theme: how and where does probability fit in the legal context?

My project can be situated against the background of this ongoing debate. While the focus of the literature on probability and the law is largely on modelling the legal decision-making process as a whole, however, my interest is in integrating probabilistic evidence in relation to particular findings of fact. That is, while much of the debate centers on whether a particular probability assignment can be made to the ultimate legal standards of proof on a “balance of probabilities” or “beyond a reasonable doubt” and whether the entire process of arriving at these ultimate determinations can be mathematically modelled, my view is that the role of probability in the law is best confined to parts of the fact finding process. That is, I in no way suggest that judges or juries should assign numerical values to all their relevant prior beliefs and proceed to repeatedly apply Bayes’ rule in the face of each new piece of evidence presented. This would be as undesirable as it is unfeasible. When an Australian judge attempted the feat, he reported an inability to confidently assign numerical values to his prior beliefs forcing him to work back from the plausibility of the results to “make up” numbers to assign to his prior beliefs.7 My proposal is limited to the suggestion that Bayesian methods should be used to inform an intuitive interpretation of statistical evidence in relation to certain findings of fact.

There is also more at stake in legal decisions than truth or verdict accuracy, and more involved in the legal decision-making process. The law seeks to promote societal goals that are distinct from fact finding and are important. To contemplate the prior probability of the accused’s guilt, for example, is inconsistent with the law’s mandated presumption of innocence.8

Closely related to this is a possible explanation for why we exclude unpalatable statistical evidence relating to race in the trial deliberation process. In “The Inevitable Efficiency of Racist

---


8 For more on this and other concerns about the use of Bayesian methods to fully model legal decision making, see Alex Stein, “Judicial Fact-Finding and the Bayesian Method: The Case for Deeper Scepticism About their Combination” (1996) 1 International Journal of Evidence & Proof 25.
Statistical Evidence\textsuperscript{9}, Amit Pundik discusses precisely this kind of statistical evidence. Pundik describes a hypothetical case in which a wealthy Jewish businessman is accused of tax fraud and denies the allegations. At trial, the prosecution wishes to put forward statistical evidence showing that the probability of a person committing tax fraud is doubled if he is Jewish. Pundik posits, and I think most would agree, that the use of such evidence is “obviously objectionable”. Pundik uses this hypothetical to argue for the deficiency of truth-efficiency theories of law to provide a complete picture. Instead, he suggests, corrective justice theories are better placed to explain the exclusion of such objectionable evidence.

This concern provides a perfect opportunity for me to limit the ambit of the discussion to come. The law cares about many factors apart from truth. Criminal law is shaped by at least three fundamental principles: (i) the error distribution principle which instructs that any errors should preferably benefit the accused; (ii) the error reduction principle which states the law’s preference that convictions track guilt and acquittals track innocence; and (iii) non-epistemic policy principles that are concerned neither with error distribution nor error reduction. Private law does not favour error distribution in one direction over the other, which informs the different standard of proof that applies (namely, the plaintiff succeeds where the evidence supports her \textit{on a balance of probabilities} as opposed to the standard of \textit{beyond a reasonable doubt} that applies in criminal law cases).

In terms of non-epistemic policy principles that are concerned neither with error distribution nor its reduction, the law is concerned with fairness of process and about efficiencies of time and cost. Getting at the truth and using a fair process to get there frequently conflict in the criminal context. For example, evidence inappropriately obtained by the police is routinely omitted from evidence even where there is no question that the consideration of such evidence would assist in getting at the truth of the matter. Why? The law cares about protecting individuals from the state as well as protecting them from one another. To allow inappropriately obtained evidence to be used to convict someone would create an incentive for police to engage in tactics that society cannot condone. In this way it is important to distinguish between what I will call “verdict

\footnotesize
accuracy” and “verdict acceptability”. Verdict accuracy reflects only one of the law’s goals: namely that legal outcomes track the truth of the matter. And this is of course where probability is instructive. But verdict acceptability encompasses all of the law’s goals and extends beyond verdict accuracy.

Treating accused as individuals and not attributing guilt by association is also central to verdict acceptability. In this, the law stands in stark contrast to insurance schemes. Many people find the use of statistical information to determine individual premium levels offensive. Why should a young man have to pay higher premiums just because he belongs to a higher-risk group? Shouldn’t he be judged on the basis of his actions and his choices? In fact, in City of Los Angeles Department of Water and Power v. Manhart10 the United States Supreme Court impugned the use of sex-segregated actuarial tables to determine contributions to a defined-benefit pension plan as discriminatory. In the context of private insurance, however, there is good reason to view risk classification in underwriting as acceptable: the point of insurance is not to socialize losses, but to socialize risk.11 Insurance should equalize expected benefit from participation in the plan, not expected outcome. And the only way to do this is on the basis of statistical generalizations. If people knew what their future losses would be, there would no point to insurance. And a private insurance scheme that sought to equalize outcomes would be unfeasible: those in the low expected loss group would never sign up for it. That is, if you make men and women pay equal insurance premiums for an annuity that will likely benefit women far more than men (because on average women live longer), then it is no longer worthwhile for men to participate in the scheme.12

But law is not like insurance. Both criminal law and torts are focused on wrongdoing. There is an important element of fault inherent in the process that is entirely absent from the insurance context. Given that the law seeks to punish and deter wrongdoing (and also to rehabilitate wrongdoers), it would be anathema to law’s goals to allow risk classification to affect

10 (1978) 435 U S 702.
12 Ibid.
determinations of fault. Just as we must look to the purposes of insurance to understand why classification of risk in that context makes sense, so too must we consider the goals of law in evaluating risk assessment here. Law aims neither to socialize losses nor to socialize risk but rather to allocate fault. If risk classification were relevant to determinations of fault, then those with high risk profiles, that is, those belonging to groups most likely to commit wrongs, would actually be subject to a lesser deterrent effect. They could then reason that faced with a high likelihood of punishment for wrongs, they may as well reap the benefits of committing them! So while it makes sense that insurance arrangements run contrary to our intuitions about responsibility and desert – they are, after all, only mitigating the effects of bad luck, not redressing them – the law remains faithful to these values.

For this reason, statistical generalizations about fault have no place in the law. As we will see from some of the examples that follow, such statistical generalizations would be of very limited relevance even if they were considered by decision makers, but the reason for their categorical exclusion goes beyond limited relevance and relates to fundamental philosophical commitments to the idea of free will.

To understand the rationale behind excluding statistical evidence relating to fault is also to see that there is no need to abhor all statistical generalizations. They have a very useful role to play in many factual determinations. Where probabilistic and legal assessments part ways, it should be clear that legitimate considerations of verdict acceptability are in play. My goal is to shed light on the dark corners where verdict acceptability is invoked to obscure blatant computational errors. Non-epistemic considerations should be made explicit and contemplated separately from evaluations of truth. It is one thing to reject illegally obtained evidence on the basis that we wish to discourage police violations; it is quite another to reject it based on some false computation that is embraced as correct because it supports the desired result. This would be a little like congratulating a child who sets out to give his sister less than her fair share of the candies but through faulty arithmetic winds up giving her an equitable share: “Good job Johnny! 30 divided by 2 minus 5 candies is 15!” If an equal sharing is desired, then this should be made explicit: “No Johnny. 30 divided by 2 minus 5 candies would only be 10, but you should still give your sister 15.”
A vivid illustration of the limitations of truth in the legal decision-making process is provided in a reporter’s account of attending the hearing in the case of Eric Leonard, a twenty-two-year-old man accused of serial murder in California. In this particular preliminary motion, the issue was whether there should be a change of venue for Leonard’s trial:

The defendant raised his hand, waiving it ever so slightly. “I am guilty,” Leonard said. Immediately…the courtroom fell silent…When it appeared that Leonard was about to repeat his statement,…Assistant Public Defender Caroline Lange swung her arm around Leonard and whispered frantically in his right ear. His other attorney…spoke into the left ear. Before anything else could happen, Judge Thomas M. Cecil – noticeably jarred – said, “Let’s take a break and go off the record.” He quickly left the bench, and proceedings were adjourned for the day. Neither Justice Cecil, Deputy District Attorney John O’Mara nor the defense attorney would comment on Leonard’s unsolicited remark.13

This particular incident occurred in California, but a similar response could be expected in Canada. It may seem surprising that neither the judge nor the prosecutor treated an accused serial killer’s voluntary courtroom confession as a welcome discovery of the truth or of the defendant’s capacity for remorse. Robert A. Kagan uses the incident to illustrate, “how deeply the American legal profession is wedded to the ideology of adversarial legalism.”14 It also suggests how deeply the Western legal system is wedded to due process and not just to the truth (assuming, that is, that you believe a voluntary confession of this nature is very likely to reflect the truth).

Whether this is as it should be is a subject that Kagan explores but that is well beyond the scope of my project. I only wish to point out that embracing a mathematically sound approach to statistical evidence in no way necessitates a commitment to the idea that all statistical evidence

14 Ibid. at 243-44.
will be admissible, or that where it is admissible, it will determine the ultimate legal issues in the case. My position is far more modest: where factual truth is at issue, and there are no competing policy considerations, we should present and use statistical evidence appropriately. The problematic avoidance of statistical evidence to which I have referred does not extend to cases in which the law takes a principled position against its use for reasons such as a commitment to the respect for individual autonomy (a consideration that may help to explain our refusal to use race-based statistics in assessing criminal responsibility). More will be said about this in the next chapter. Instead, my focus is on the avoidance of statistical evidence for reasons of confusion or fear about how to integrate this otherwise useful information into the decision-making process.

Where the law chooses to exclude statistical evidence, it ought to be clear about the policy reasons informing the exclusion. In the general confusion that surrounds statistical evidence, however, the distinction between non-truth-related policy concerns and relevance to the fact-of-the-matter is too often obscured.

An example of this unfortunate obfuscation occurred in the decision of a panel of the Court of Appeals for the Second Circuit in United States v. Shonubi, 103 F.3d 1085 (2d Cir., 1997). In that case, Charles O. Shonubi, a resident of the United States but a Nigerian citizen, was arrested at John F. Kennedy International Airport while allegedly trying to import heroin. At the time of his arrest there were 427.4 grams of heroin in his stomach. At trial, the judge found that Shonubi had smuggled heroin into the country undetected seven other times. What was at issue for the purposes of sentencing was the total quantity of heroin that Shonubi had smuggled in over the course of his eight smuggling ventures. The length of sentence would vary in accordance with this total quantity. To resolve this question, the sentencing judge took the amount that Shonubi had smuggled at the time of his arrest, namely 427.4 grams, and multiplied it by eight – the total number of smuggling trips he had made.

On appeal, the Court of Appeals overturned the sentencing judge’s decision, in part because they held that there was no “specific evidence” in the record to support the factual finding of the total

---

15 For a more far-reaching challenge to the primacy of the particular, see Frederick Schauer, Profiles Probabilities and Stereotypes (Cambridge: Harvard University Press, 2003).
amount of heroin smuggled. When the sentencing issue was reopened, the prosecution presented data collected by the U.S. Customs Service relating to other Nigerian drug smugglers who had used the same smuggling method during the same time period. On the basis of this evidence, the sentencing judge concluded that the total amount of heroin smuggled during Shonubi’s eight trips was between 1000 and 3000 grams. Shonubi would only have been eligible for the lighter sentence had he smuggled less than 1000 grams in total.

Shonubi appealed once again. Once again the Appeals Court overturned the sentencing judge. The Appeals Court reiterated that any finding about the quantity of heroin that Shonubi had imported during his other trips had to be supported by “specific evidence”. And this time, the court went further to assert that the customs service data did not qualify as specific evidence.

In an article concerning the court’s treatment of this case, Peter Tillers rightly criticizes the appellate court’s reliance on the idea of evidence “specificity” in overturning the decision below. This is because to reject all non-specific evidence is to reject the whole class of information that is gained by gathering statistics for the purpose of drawing inferences about unknown matters. As Tillers states,

[i]n principle, statistics, when offered as a basis for inferences about unknown matters, are never “specific” to the matter about which an inference is to be made. The validity of drawing inferences on the basis of statistics, thus, never rests on the “specificity” of the statistical evidence to the unknown matter in question. What assures the potential inferential value of the statistics one chooses to use is, rather, the perceived similarity of the known or observed events or matters observed to the unknown or unobserved matter or event about which a conclusion is sought to be reached.

This apparent clash between legal reasoning and sound statistical reasoning is what I hope to limit. That is, my view would prevail so long as the appellate court did not use lack of specificity as the basis for its rejection of the customs evidence. By resorting to an argument based on lack of specificity, the Appeals Court failed to understand the essential nature of statistical evidence, which indeed provides useful information about our uncertain world even though it is not specific to a particular unknown. There may well be legitimate reasons to reject the customs evidence in sentencing Shonubi. For example, the law may consider it an affront to individual autonomy to use any statistical evidence pertaining human behaviours: that is, the law may wish to draw a blanket prohibition on any inferences from the behaviour of a group to the behaviour of an individual. But to criticize the customs evidence on this basis is not to take issue with the evidence’s lack of specificity, but rather with drawing inferences from the behaviour of groups to the behaviour of an individual. An example of equally non-specific evidence that would not offend individual autonomy in this way would be statistical evidence about human physical limitations related to transporting drugs. If there were statistical evidence before the court stating that 99% of people cannot physically transport more than a certain quantity of heroin at one time, this evidence would be no more specific, but would likely prove unobjectionable. That is, there would be no policy reason to object to an inference from non-specific evidence about the physical limitations of the human gastro-intestinal track in general to the physical limitations of a particular human. The law views behaviours differently.

What I wish to stress is that the weakness in the Appeals Court’s reasons in Shonubi was its failure to articulate the truly objectionable aspect of the use of the customs evidence. The court’s criticism of the evidence as non-specific was, ironically, insufficiently specific. Instead, they should have taken care to distinguish between objectionable and non-objectionable types of statistical evidence.

My response to both Pundik’s hypothetical case of racial evidence and the real case of Shonubi will hopefully assuage concerns that radical implications arise from my view. In fact, the only necessarily reformist implication of the position I put forward is that the law should be more

---

17 For a challenge to the legal system’s reliance on the idea of free will, see Kelly Burns & Antoine Bechara, “Decision making and free will: a neuroscience perspective” (2007) 25:2 Behavioural sciences & the Law 263.
receptive to many types of statistical evidence. Concerns beyond verdict accuracy currently exist within the law and are unaffected by my proposals. My objective is to show that where we are interested in the truth, and in the absence of other countervailing considerations, the law can and should do a better job of using statistical evidence.

I am also unconcerned with commonly debated questions of whether subjective probabilities (also known as personal probability estimates) of factual propositions can be used by legal decision makers, and whether it is possible to define the underlying purposes of trials. It seems to me that subjective probabilities are necessarily used by legal decision makers except where expressly forbidden. One such prohibition has already been alluded to, namely the presumption of innocence. Jurors are instructed to check some of their prior beliefs at the courtroom door, while other prior beliefs inevitably play a role in the decision-making process. In a “challenge for cause”, jurors are routinely asked whether the fact that the accused is a black man will affect their ability to judge the evidence impartially. This question amounts to an inquiry into the juror’s prior beliefs about black men and the commission of crimes. Built into the challenge for cause process is the belief that some subjective priors impede fair process goals. The challenge for cause makes it clear that it is unacceptable to import prior probability assessments of guilt based on race into the decision-making process. At the same time, when it comes to assessing any number of other facts, decision makers will necessarily import their subjective understanding of the world into an assessment of new evidence. If a complainant’s account of an assault includes the assertion that she bit the hand of the accused as hard as she could, jurors will expect pictures of the accused’s hands shortly after the assault to reveal marks. Why? Because they have a prior belief that human bites mark flesh. Since jurors must import personal estimates of probabilities into legal decision making, the question of whether they can is of little interest.

Lastly, while I allude to the underlying purposes of the trial in explaining the need for considerations beyond verdict accuracy, my aim is neither to define nor justify these purposes. My goals are to encourage the acceptance of probabilistic evidence as relevant to legal determinations of fact, to highlight common errors in the use of such evidence, and to suggest ways to limit their occurrence.

These goals are of increasing importance since statistical evidence has become increasingly relevant to a growing number of legal contexts. This is in part because new technologies have
greatly simplified the collection and analysis of large bodies of data. For example, explicitly statistical evidence is now often central to cases involving the fairness of tax assessments, the safety and efficacy of drugs and chemicals, charges of discrimination in employment and housing, the use of blood tests to determine paternity, the accuracy of forensic tests based on blood, semen or hair samples, and trademark infringement.\(^{18}\)

My focus will be on the types of cases in which findings of fact play a pivotal role. The law distinguishes between questions of fact and questions of law, acknowledging a middle ground of mixed fact and law. Factual questions are questions about what happened in the world, while legal issues concern the legal principles that are applied to such facts. A typical factual question in a criminal case may be whether the accused was at the victim’s house on the night she died. This is a question about which there is a truth of the matter. Either the accused was there or he was not. End of story. In contrast, the question of whether the accused’s belief that he was in danger is sufficient grounds for self-defence is legal in nature. The answer to this question may change from time to time or place to place and is open to debate about the relative merits of differing responses. And the question of whether the accused acted in self-defence would be a question of mixed fact and law since it involves the application of a legal principle to the specific facts of the case. Statistical evidence may be invoked in relation to any of these questions. My focus, however, is on the use of statistical evidence where the question at issue is purely one of fact.

1.4 Why Bayesianism?

Given a focus on findings of fact, the question remains as to why a Bayesian normative approach to statistical evidence should be accepted. Discovered by English clergyman Thomas Bayes, Bayes’ rule establishes a relationship between prior and conditional probabilities that can be used to model rational learning. Put simply, this model of learning looks like this: initial beliefs + recent objective data = a new and improved belief.\(^{19}\) Expressed in this way, there is little controversial about the acceptance of the rule and its application. French mathematician Pierre-


Simon Laplace built on the rule by recalculating the equation with every new piece of data. Using this approach he could distinguish highly probable hypotheses from those that were not. While the main criticism levelled at Bayesian methods relates to their subjectivity – everyone starts with his or her own subjective initial beliefs – the success of Bayesian methods is apparent. Bayesian computer analyses were effectively used by U.S. military intelligence to predict the most probable paths of Soviet nuclear submarines in the 1970s. Today, Bayesian analyses are successful in sorting spam from e-mail, assessing medical and security risks, and decoding DNA.\(^\text{20}\) The phenomenal successes of Google’s translator and spell-checker are also attributable to the application of Bayes’ rule: they both employ the principle of repeatedly updating knowledge on the basis of vast quantities of new objective data. This process generates highly accurate translations and spelling corrections. How do I know? I “googled” it.\(^\text{21}\) By drawing both upon the vast pool of search terms entered by Google users across the planet, as well as the results of such searches, Google generally “knows” what word you’re trying to spell; this approach far exceeds the success rate of spell-checkers based on a pre-established lexicon. Similarly, Google’s translation program works by analyzing millions of documents that have already been translated by human translators, identifying patterns, and repeating this process billions of times. It is this repeated updating using the combination of Bayes’ rule with the vast quantities of available source data that accounts for the program’s impressive results.\(^\text{22}\) In this way, it is the indisputable real-world success of the Bayesian approach that best responds to the question “why Bayesianism”.

### 1.5 Bayesian Updating is What We Do (Most of the Time)

In addition to the real world success of Bayesian methods, it is also true that considering prior probabilities comes naturally to most of us most of the time. At the same time, accepting Bayesianism as a formal normative framework allows us to identify the penumbral cases in which mistakes are made, and in this way opens the door to correcting such errors.


For example, in contemplating circumstantial evidence, prior probabilities are implicitly accounted for. In the law of negligence, for example, the legal principle of *res ipsa loquitur*, literally translated as “the thing speaks for itself” necessarily invokes prior assessments about the world. *Res ipsa loquitur* allows legal decision makers to surmise that the most probable cause of an accident was negligence even where there is no direct evidence to speak to the negligent act. The principle was first defined in 1863 in a case in which a barrel of flour rolled out of a warehouse window and fell on a pedestrian walking on the sidewalk below. It was deemed reasonable to conclude that under the circumstances, the defendant was probably culpable since no other explanation was likely. This conclusion was based on the *prior* knowledge that barrels of flour do not fall out of windows unless they are dropped or insecurely fastened – both of which would amount to negligence on the part of the defendant in that case. The principle is now regularly applied to cases involving injured passengers and railroads or airlines where in the case of injury, plaintiffs often argue that in the absence of an alternative explanation, the burden lies with the defendant to dispel what is (now) a prior probability of causation. This is a tactical maneuver that effectively creates the impression of a prior probability of causation which may appear to shift the burden of proof from the plaintiff to the defendant.

In this way, common sense or “gut feelings” are often Bayesian gut feelings. What we are interested in are the outlier cases in which evidence presented in a particular way elicits a reaction at odds with the correct Bayesian result. These sorts of errors can be identified once Bayesianism is accepted as a suitable normative model that is actually consistent with legal intuition on most occasions.

### 1.6 Why a Normative Model is Needed

If our gut feelings are often consistent with sound Bayesian reasoning, then where is the problem in the first place? The answer is this: it is important to be able to judge where and how thinking about the facts in a legal case has gone awry. Sometimes we think about things in the right way and other times we get it wrong. There are certain well-known biases in our thinking that can adversely affect our ability to think about things correctly especially when presented in ways that take advantage of these biases.
The term “bias” is used here to refer to systematic errors that people make in choosing actions and in estimating probabilities. But errors can only be identified as such where there is a more objective assessment of rationality and probability that is independent of common sense or gut feelings. An objective assessment that sets a standard for evaluation is called “normative”.

If my reader is suspicious of the assertion that sometimes common sense just gets it wrong, let’s consider a couple of examples.

Suppose there are two contraceptive methods available, each of which is known to be 90% reliable. You decide to use both to make doubly sure that pregnancy is averted. It then occurs to you that the probability of success of both contraceptive methods is $0.9 \times 0.9 = 0.81$ or only 81%. Somehow it looks as though in making doubly sure to avoid pregnancy, you’ve actually reduced your chances of effective contraception! But to reason in this way is to err in understanding how probabilities work. While it is true that it is appropriate to multiply two independent probabilities to arrive at the probability that both independent events will occur, you cannot simply multiply probabilities when the two are not independent. Since both methods of birth control affect the same outcome, namely pregnancy, they are not independent. Instead, the second method of birth control will prevent pregnancy in 90% of the 10% of cases where the first method has failed. 90% of 10% is equal to 9%. This means that using both methods has actually increased the efficacy of your birth control from 90% to 99%.

Another way to think of this example is that while the two outcomes, namely pregnancy, are not independent, the probabilities of failure for each method of contraception are. That is, the 10% failure rate for one method of birth control is independent of the 10% failure rate for the second method: there is no reason to think that there is any relationship between, say, the failure of the birth control pill and the failure of a condom. Given this independence, you can multiply the probabilities of failure to generate a failure rate of $0.1 \times 0.1 = 0.01$. By using two methods of

---


24 I owe this example to Professor Ronald deSousa, Professor Emeritus at the University of Toronto.
birth control, each with 90% reliability, the probability of failure is reduced from 10% to 1%, generating the same, correct, result of a combined efficacy of 99%.

This example illustrates at least two things: (i) a basic understanding of probability is essential to comprehending our everyday world; and (ii) if probability theory is accepted as the normative model, errors can be identified and the correct result can often be explained in a way that becomes intuitively accessible.

I will turn to one more example before moving on. Imagine that you have an avid interest in politics (this will require more imagination for some than for others). In advance of an election, you read the results of a poll and learn that your preferred political candidate, Ernie, is leading Bert by 51 percent to 49 percent with a margin of error of 2 percent. Common sense may lead you to spend election-day on tenterhooks, earnestly believing that this is one extremely close race. After all, the difference between 49 and 51 is only 2%, so it may seem as though reporting a 51% lead with a 2% margin of error means that Bert and Ernie are in a dead heat. But, a 2% margin of error is equivalent to a 1% standard deviation and a so-called “normal” distribution curve looks like the one shown in Figure 1:

![Normal Distribution Curve](image)

Figure 1: Normal Distribution Curve
Assuming the poll results are normally distributed, a result of “51%” (which places “51” at the highest point of the curve) with a 1% standard deviation translates into a 68.2% ($34.1\% + 34.1\%$) chance that Ernie is within 1% of 51% in either direction, and only about a 16% chance that he falls below that. Viewed in this way, according to the poll, Ernie’s chances of a win are 84%! So you can put your feet up and relax. This is no dead heat!^{25}

The above two examples illustrate the need for some standard against which erroneous reasoning can be judged as such. Sometimes we do just get it wrong. It behooves us to accept a structure that can reveal this.

1.7 The Problem: Case Illustrations of Trouble with Numbers

In this section, I describe a number of real cases from a variety of countries that illustrate the perils of the current approach to statistical evidence. While my ultimate focus is on the Canadian legal system and its rules of evidence, the cases described do not depend in any relevant way on differences between the legal systems of Canada and the countries in which they occurred.

Under the current approach to statistical evidence, I contend that relevant and useful statistical evidence is too often avoided or misused.

1.7.1 Statistical Evidence: Problematic Avoidance

1.7.1.1 Statistical Environmental Evidence Avoided

The issue of statistical background rates arose in the case of Massachusetts v. EPA.^{26} The treatment by the United States’ Supreme Court of background rates in that case was discussed in an article by Aaron Taggart and Wayne Blackmon.^{27} The particulars of the case were as follows: petitioner states sued to force the United States Environmental Protection Agency (the “EPA”) to regulate emissions from automobiles. To succeed, the petitioners were required to prove harm. The harm contemplated here was global warming. For the EPA’s acts or inaction

^{25} Example adapted from one provided in Ian Ayres, Super Crunchers (Toronto: Bantam Books, 2007) at 202-204.

^{26} (2007) 127 S Ct 1438.

to have caused an increase in global warming, however, would require consideration of the background rate of global warming \textit{unrelated} to the EPA’s actions. That is, to prove their case, the petitioners would have to distinguish between, on the one hand global warming associated with man-made pollution in the United States, for which the EPA could potentially be held responsible, and on the other hand, global warming due to other causes. That is, the likelihood of the hypothesis that EPA regulation of man-made pollution caused or contributed to global warming (the probability that would inform the ultimate determination with respect to the role and any legal responsibility of the EPA), depended on the probability of global warming \textit{absent any human intervention}.

The case went all the way to the Supreme Court, but at each level the courts did not squarely address the problem posed by the need to address this background rate. The focus instead was on the EPA’s exercise of discretion and the case was ultimately decided in the EPA’s favour on the basis of standing. The case provides just one example of the ways in which statistical evidence, and background rates in particular, are relevant to judicial determinations. It also exemplifies the judicial system’s reluctance to directly face the relevance of background rates.

1.7.1.2 Statistical Evidence on Seatbelt Efficacy Avoided

In addition to the problematic avoidance of statistical evidence that is clearly relevant to an issue for determination, statistical evidence is also sometimes improperly excluded when it is relevant and could be helpful.

For example, in \textit{Beazley v. Suzuki Motor Corporation}\textsuperscript{28} the British Columbia trial court excluded evidence relating to the statistical link between seatbelt use and injuries in motor vehicle accidents in a case concerning contributory negligence.

The case involved a single car rollover accident in which the three passengers suffered serious injuries and commenced actions against the driver of the car. One of the passengers was suing the driver for negligence, alleging that the harm she suffered as a result of the accident was due to the driver’s negligence. At trial, one of the issues was to be whether the defendant could raise

\textsuperscript{28} 2010 BCSC 480 (CanLII).
a defence of contributory negligence, since the passenger was not wearing a seatbelt at the time of the accident. If the driver could prove on a balance of probabilities that the use of a functioning seatbelt would have minimized or avoided the plaintiff’s injuries, the award of damages could be reduced or avoided altogether.

The defendant sought to introduce expert evidence in the form of a report showing the statistical link between seatbelt use and injuries in motor vehicle accidents. The report reviewed statistical data and generated an assessment of the relative risk to belted and unbelted occupants in rollover accidents. The conclusion of the report was that an unbelted occupant in a light truck rollover crash is about six times more likely to be seriously injured than is a belt occupant. Similarly, an unbelted occupant is about eight times more likely to sustain serious head/face/neck injuries than a belt occupant. Overall, the statistical data suggested that seatbelts are highly effective in reducing complete ejection in vehicle rollover accidents.

In considering whether to admit the expert evidence in relation to seatbelt efficacy, the judge turned to the test for the admission of expert evidence, which depends on the application of the criteria of (i) relevance; (ii) necessity in assisting the trier of fact; (iii) the absence of any exclusionary rule; and (iv) a properly qualified expert. He concluded that the statistical evidence should be excluded. The trial judge’s reasons for this were as follows:

It has been long recognized in British Columbia that a party who fails to use an available seatbelt and sustains injuries more severe than if the seatbelt had been worn will be found to be contributory negligent: Yuan et al. v. Farstad (1967), 66 D.L.R. (2d) 295 (B.C.S.C.); Gagnon v. Beaulieu, [1977] 1 W.W.R. 702 (B.C.S.C.).

While there appears to have been statistical evidence led in Yuan and in Gagnon, subsequent cases have held that such evidence is not necessary. In Lakhani (Guardian ad litem of) v. Samson, [1982] B.C.J. No. 397 (S.C.) McEachern C.J.S.C. (as he then was) noted at para. 3:

I reject the suggestion that engineering evidence is required in these cases. The court is not required to leave its common sense in the hall outside the courtroom, and the evidence is clear that upon impact in both cases the Plaintiff's upper body was flung or thrown forward striking the dashboard or the steering wheel. And common sense tells me that the restraint of a shoulder harness would have prevented that, and therefore some of the injury from having occurred.  

This rejection of statistical evidence appears to be based on a misunderstanding: while it may well be true that common sense is sufficient to establish that a seatbelt would have reduced the damage suffered by the plaintiff, surely common sense is insufficient to quantify either of these factors. And quantification becomes relevant in cases where contributory negligence is alleged. According to s.3 of the Ontario Negligence Act, R.S.O. 1990, c. N.1, 

\[
\text{in any action for damages that is founded upon the fault or negligence of the defendant if fault or negligence is found on the part of the plaintiff that contributed to the damages, the court shall apportion the damages in proportion to the degree of fault or negligence found against the parties respectively.}
\]

Once a finding of negligence by both parties has been made, the court will attempt to quantify and apportion the damages in proportion to the degree of fault. In the 2005 case of *Snushall v. Fulsang*\(^{31}\) the Ontario Court of Appeal embraced the guidelines laid out by Lord Denning in *Froom v. Butcher*\(^{32}\), in stating that the apportionment to the plaintiff for not wearing a seatbelt should not exceed 25 per cent. In that case, the court emphasized that while wearing a seatbelt may have reduced or even prevented injury, it was still not a cause of the accident. In *Froom v. Butcher*, Lord Denning proposed that the normal apportionment when failure to wear a seat belt made a considerable difference in the severity of the injuries suffered by the plaintiff versus

\(^{30}\) *Beazley v Suzuki Motor Corporation*, 2010 BCSC 480 (CanLII) at paras 29-30.  
\(^{31}\) *Snushall v Fulsang*, [2005] 78 OR (3d) 142.  
\(^{32}\) *Froom v Butcher*, [1975] 3 All E R 520.
when it would be prevented the injuries entirely should be 15 and 25 percent respectively. When seatbelt use would have made no difference, there should be no apportionment to the plaintiff.

In order for the court to assess where on the spectrum of “no difference” to “complete prevention of the injuries” this particular plaintiff’s failure to wear a seatbelt lies, the engineering evidence on offer would have been instructive. The exclusion of an expert report may be justifiable on the basis of efficiency or cost-savings, but not on the basis that it adds nothing to common sense.

1.7.2 Statistical Evidence: Misuse

One reason for the problematic avoidance of otherwise helpful statistical evidence is the frequency with which statistical evidence is misused when it is presented. While avoidance is one response to this problem, I hope to show that this is a high and unnecessary price to pay. Instead of avoiding statistical evidence altogether, the law needs to develop a viable approach to its integration. This is a two-step process of embracing a normative model with which to evaluate the correctness of reasoning with statistical evidence and then developing a mode of presentation that promotes sound reasoning to correct conclusions.

One of the most common errors made in reasoning both with explicitly statistical evidence as well as when the probabilities inherent in the evidence are only implied, is what is sometimes referred to as “failing to think of the opposite” or failing to consider the probability of the data in the case that the hypothesis under consideration is not true.

The so-called “prosecutor’s fallacy” is a subset of this type of error. The term “prosecutor’s fallacy” is invoked in relation to the improper equation of the low probability of the data under consideration with the low probability of innocence. For example, if a lottery winner is accused of cheating, then to put forward the very small probability of winning the lottery without cheating and equate it with the chance of innocence would be an example of the prosecutor’s fallacy. The fault with this line of reasoning is that it fails to consider the opposite: namely, the probability that someone will claim a win where there has been no cheating. This probability is extremely high; most lottery jackpots do not go unclaimed. This is so despite the vanishingly small chances of any particular person winning.
Along similar lines, if a mother is convicted of murdering two of her infant children on the basis that the probability of two accidental deaths in one household is extremely low, this would also be an example of the prosecutor’s fallacy and failing to think of the opposite. Though multiple accidental deaths are rare, so are the opposite: multiple murders. Given only the facts of the deaths as evidence, the unlikelihood of accident is only of any interest in direct comparison to unlikelihood of murder.

**People v. Collins**

The prosecutor’s fallacy was one of a number of errors that occurred in the presentation of statistical evidence in *People v Collins.* In that case a woman was robbed in Los Angeles and a husband and wife were subsequently arrested. The victim recalled only that she saw a young woman running from the scene after she, the victim, had been pushed down. The victim described the offending woman as around 145 pounds with hair “between a dark blond and a light blond”. Bystanders reported sounds of crying and screaming coming from an alley and observed a woman running and then entering a partly yellow car driven by a black male with a mustache and a beard. The prosecution rolled out an expert to testify that since the defendants were “a Caucasian woman with a blond ponytail…[and] a Negro with a beard and mustache” who drove a partly yellow automobile, this information was sufficient to convict the defendant couple. The “expert” identified the following characteristics as relevant: partly yellow automobile, man with mustache, negro man with beard, girl with ponytail, girl with blond hair, and interracial couple in car. He then assigned an individual probability to each of these characteristics: 1/10, 1/4, 1/10, 1/10 1/3, and 1/1,000 respectively. Applying the multiplication rule, he multiplied all the probabilities to conclude that the chance of a couple fitting all the distinctive characteristics was 1 in 12 million. The expert then equated this probability with the couple’s chance of innocence, inviting the jurors to supply their own individual probabilities while at the same time suggesting that he had used “conservative” estimates. The jury convicted.

---

33 (1968) 66 Cal Rptr 497.
The decision was overturned on appeal, but is often trotted out as an example of the dangers inherent in the use of statistical evidence. There are many ways to understand how the expert (and the jury in relying upon him) went wrong in this case.

To begin with, the expert witness treated all of the characteristics as independent without providing any reason for this strong assumption. Most men with beards, for example, also have moustaches. In this way, the characteristic of having a moustache is not independent of the characteristic of having a beard. As has already been explained, the justified application of the multiplication rule to a set of probabilities presumes the independence of the variables involved. If that presumption is mistaken, then the resulting probability will also be grossly mistaken. In the extreme, it would quite obviously be mistaken to reason as follows: according to the victim, the culprit was an East Asian male with dark hair and dark eyes who was also less than six feet tall. If you consider that 15% of the population is East Asian, 35% of people have dark hair, 35% have dark eyes and 60% of the male population is less than six feet tall, by application of multiplication rule, the chance of finding someone with all of these characteristics is only about 1%! The calculation has gone horribly awry because the factors under consideration are not independent: most East Asians have dark hair, dark eyes and are less than six feet tall.

In addition, just as I have done in my example, the expert provided no support for the individual probabilities that he presented and upon which the ultimate probability depended. That is, we have no reason to believe that 1/10, 1/4, 1/10, 1/10, 1/3, and 1/1,000 actually reflect the likelihood of any of the characteristics in question.

In addition to using unsubstantiated probabilities and treating non-independent factors as independent, the expert witness committed the prosecutor’s fallacy by mistakenly equating the probability that a couple had all of the characteristics in question with the chances of the defendant couple’s innocence. These concerns were raised by the Court of Appeal and resulted in the overturning of the decision at trial.

---

34 If the probabilities used in this example are accurate it is only by accident.

35 People v Collins, (1968) 66 Cal Rptr 497.
Equating the probability that a couple would have all of the characteristics in question to the chances of innocence is a classic example of the prosecutor’s fallacy. In this case, the fallacy arose from treating the chance that any randomly selected couple would meet the description of the culprits as equal to the chance that this particular couple, already sitting before the court because they possessed these characteristics, committed the crime. Put in very simple terms, since the police already went out to apprehend a couple somewhere in the U.S. with the described traits, the question becomes whether it is this pony-tailed and bearded interracial couple with a yellow car that committed the crime, or one of the other couples in the country meeting the same description. And since there were at least 60 million couples in the U.S. at the time, one could expect around 5 instantiations of an event occurring with a probability of 1 in 12 million. And 1 in 5 is by no means sufficient probability for a finding of guilt beyond a reasonable doubt.

People v. Collins illustrates a number of errors that can be made in the presentation of statistical evidence. It also plays a significant role in the debate about statistical evidence in the law since it illustrates the dangers inherent in the use of such evidence. Others in the debate conclude from this that avoidance is the best response. It is on this point that I beg to differ.

R. v. Johnson

The error of failing to think of the opposite can also arise in the context of DNA evidence. In August 1995, a seven-year-old girl (the “complainant”) went to her friend’s home for a sleepover. During the night, she was taken from the home, raped and abandoned near a shed. After the assault, the complainant went to a nearby neighbour who took her in, wrapped her in a quilt, and called the police and the girl’s family. At the time, the complainant stated that her attacker was a white teenager with brown curly hair and described how he was dressed.

The police conducted a two-year investigation without success. They obtained DNA from semen found on the complainant’s pyjamas and on the quilt (the “sample DNA”). The police also obtained DNA from several male suspects who were at, or associated with, the residence where the sleepover had taken place. None of those suspects matched the sample DNA.

36 R v Johnson, 2010 ABCA 230 (CanLII).
Five years later, the National DNA Data Bank of Canada (the “Data Bank”) was developed. The Data Bank contains both genetic profiles drawn from unknown individuals and those drawn from individuals convicted of particular crimes. The sample DNA was entered into the database and was found to match Neil Johnson’s genetic profile. Johnson was charged and tried before a jury in 2008. This was 13 years after the incident. At trial, the Crown called witnesses, entered into evidence a video-taped interview with the complainant taken at the time of the incident, and presented the DNA evidence along with expert testimony that the probability of “selecting an unrelated individual at random from the Canadian Caucasian population with the same profile [as the sample DNA] was 1 in 890 billion”. The expert went on to testify that if you eliminate an identical twin and a full brother from the analysis, it was within scientific certainty that Johnson deposited the sample DNA. At the time of the incident, Johnson was 18 years old. He had no twin or full brothers, but did have several other male blood relatives.

Johnson’s defence at trial was that there was no evidence about the likelihood of a random match with a blood relative, other than that it would be more likely than 1 in 890 billion. Therefore, he submitted that there was a small pool of people who could have produced the sample DNA and that he was only one member of that small pool.

On appeal to the Alberta Court of Appeal, the appellant argued, among other things, that the jury should have been specifically instructed about the prosecutor’s fallacy, namely that they should have been warned that the DNA evidence spoke to the chances of a random match and did not directly measure the probability of the accused’s guilt. In particular, the defence argued that the jury should have been provided with evidence as to the chances of a false match with some other male relative. The appeal was dismissed.

In effect, the defence’s argument on appeal was that the jury should have been instructed to “think of the opposite”. Was the DNA evidence here properly presented and understood? The final chapter of this dissertation will answer this question in light of all that has preceded it.

37 Ibid.
The Case of Lucia de B

In March 2003, a nurse named Lucia de B. (“Lucia”) was sentenced to life in prison for the murder or attempted murder of a number of patients in two hospitals where she had worked. At issue in that case was not whether Lucia had the requisite criminal intent for the alleged crimes but rather whether Lucia’s presence on the ward was causally related to the incidents. Were her actions the cause of the deaths? The two possibilities that the court considered were that either Lucia’s presence was causally related to the deaths, or the coexistence of her presence and a disproportionate number of deaths was a coincidence. At one of the hospitals, an unusually high proportion of deaths or near-deaths (“incidents”) occurred during her shifts. That is, at one hospital where Lucia worked 142 out of the hospital’s total annual number of 1029 shifts (i.e. Lucia worked 13.8% of all the shifts that year), all 8 of the incidents that occurred in that hospital (i.e. 100% of the incidents) took place during one of Lucia’s shifts. At another hospital where she worked, in one ward, 1 out of a total of 5 incidents in that ward occurred on Lucia’s watch where she worked 1 of 336 shifts. In another ward, 5 out of a total of 14 incidents occurred during Lucia’s 58 (out of a total of 339) shifts. The main issue at trial was whether the disproportionate number of incidents during Lucia’s shifts was merely coincidental.

At trial, a witness for the prosecution, Dr. Elffers, provided the following evidence: if we assume that (i) the probability that Lucia experienced an incident during a shift was the same as the corresponding probability for any other nurse (i.e. Lucia didn’t cause the incidents); and (ii) that the occurrences of incidents are independent for different shifts; then, the probability that Lucia experienced the incidents as a random matter was less than 1 in 342 million.

Based on this statistical evidence, the court reasoned as follows: Dr. Elffers concludes that the probability that a nurse coincidentally experienced as many incidents as the suspect is less than 1 in 342 million. The probability that one out of 27 nurses would coincidentally experience 8 incidents in 142 out of a total of 1029 shifts…is less than 1 in 300,000. From here the court reasoned as follows:

The court is of the opinion that the probabilistic calculations given by Dr. H. Elffers in his report of May 29, 2002, entail that it must be considered extremely improbable that the suspect experienced all incidents mentioned in the indictment coincidentally. These calculations
consequently show that it is highly probable that there is a connection between the presence of the suspect and the occurrence of an incident.\textsuperscript{38}

On this basis, Lucia was convicted of murder and attempted murder and sentenced to life in prison.\textsuperscript{39} Again the question is whether the statistical evidence was properly presented and understood.

It is clear that in the case of Lucia de B. the statistical evidence was misused. Instead of considering the relative likelihood of two improbable explanations, namely coincidence versus wrongdoing, the improbability of coincidental deaths alone was considered and erroneously equated with the improbability of innocence.

Examples such as the ones above are sometimes used to show that statistical evidence should be kept out of the courtroom on the basis that it mystifies decision makers and obscures the proper standard of proof. On the contrary, I wish to show that mistakes such as the ones alluded to above are to be avoided not by eschewing statistical evidence but by adopting a basic conceptual framework to understand it and helpful strategies to present it. A thorough consideration of how the evidence in the three cases described above could better have been presented using the Bayesian-intuit approach is found in the final chapter.

\section*{1.8 Summary}

The purpose of the analysis to come is neither to criticize the existing legal framework nor to praise it. Rather, within existing constraints, my aim is to show that Bayesianism sufficiently conforms to current legal intuition and may serve as a normative model and elucidate discrete areas in need of change.

\textsuperscript{38} R Meester et al, “On the (ab)use of statistics in the legal case against nurse Lucia de B.” (2006) 5 Law, Probability and Risk 233. The original Dutch version of the decision can be found at www.rechtspraak.nl.

\textsuperscript{39} The Ontario case against nurse Susan Nelles at the Hospital for Sick Children featured similar facts. In that case, the charges against Ms. Nelles were eventually dropped after a lengthy preliminary inquiry into whether there was sufficient evidence to put the plaintiff on trial. While the preliminary inquiry judge ultimately came to the conclusion that there was insufficient evidence to proceed to trial, the Crown’s office was clearly of a different opinion. Ms. Nelles brought an action against Her Majesty the Queen in right of Ontario (the Crown) and the Attorney-General for Ontario (the Attorney-General) after the charges against her were dropped, but the Ontario court found, and the Court of Appeal agreed, that there were no legal grounds for her action: See \textit{Nelles v Ontario} (1985), 21 DLR (4th) 103.
It is common for those entrenched in the legal tradition – legal scholars, judges and lawyers – to view criticisms of the system as exaggerated and misguided. The Anglo-American legal system, based as it is on precedent, is by its nature conservative and draws strength from a long historical foundation. With this in mind, the scope of the critique and the reforms ultimately suggested in the last chapter are relatively moderate. There is much that works well with the existing system and no desire to throw the baby out with the bathwater.

At the same time, the future is in statistical evidence and the law must adapt to embrace it. Intuition and expertise are evolving to interact with data-based decision making. New technology has made huge pools of data easily available for almost every kind of decision. Google’s spell-checker and translators discussed above are excellent examples. Business has been affected by the ready availability of data. Some rental car companies refuse to serve people with poor credit scores because data shows a high correlation between low credit scores and a high likelihood of an accident. Practices such as these are sometimes referred to as tera mining because they “mine” terabytes of information. Unlike the memory capacity of your personal computer, business and government datasets are measured not in mega or gigabytes but in terabytes. Even dating is becoming increasingly a matter of data matching. The highly popular eHarmony dating website is a system of data-driven matching that generates potential romantic matches by entering the answers to 436 personal questions into an algorithm designed to generate compatible matches. And very notably, the medical profession is increasingly embracing tera mining in basing treatment choice on statistical analyses rather than just intuitions or clinical expertise.\textsuperscript{40} This relatively recent shift in the medical profession will be discussed in some detail in chapter five and used as a possible source of inspiration for change in the law.

The world is changing and readily available statistical analyses are the wave of the future. Meaningful incremental change in the legal system is salutary and it does occur. My goal is to illuminate an area of the law that is in need of modernization. My hope is that once illuminated, the process of modernization can begin. Small but significant changes can be made by educating judges and lawyers and amending jury charges, as well as subtly changing the way in which

\textsuperscript{40} Ian Ayres, \textit{Super Crunchers} (Toronto: Bantam Books, 2007).
experts present their evidence. Ultimately, it is law schools that could effect the greatest change by including in the legal curriculum some basic education about how to understand statistical information and avoid common misconceptions.
Chapter 2
What is Legal Evidence? Constraints on Assertion and Justification

In applying a particular technique in a practical problem, it is vital to understand the philosophical and conceptual attitudes from which it derives if we are to be able to interpret (and appreciate the limitations of) any conclusions we draw.⁴¹

2 What is Legal Evidence?

In what follows, I will posit that a Bayesian updating model of justification proves useful in tackling pressing legal problems. This claim can only be made against some background understanding of what constitutes acceptable justification in the legal context. The purpose of this chapter is to supply this background. The chapter proceeds as follows:

- Assertion and justification: to each context, a different standard of caution
- Legal evidence: constraints on assertion and justification in the legal context
- Is there a unifying theory of legal evidence?

In later chapters, constraints on justification in the legal context will be contrasted with the different parameters relevant to medicine. Armed with this understanding of the rules of engagement, the next chapters will show how Bayesian updating, intuitively understood, can serve as a helpful model.

2.1 Assertion and Justification: To Each Context, a Different Standard of Caution

To assert that “P” is to say anything from a strong “Yes, P is true” to a weak, “I have a hunch that P”, or “I’d guess that P”. You have asserted “P” to be the case if, when P turns out to be false, it is appropriate to say that you were wrong. Other remarks that tend to indicate that P is so without the commitments of assertion include, “it is probable that P” or “there is good

---

evidence for P”. If it is discovered that P does not obtain, such statements are not undermined in the same way as an assertion of P would be. That is, it may be the case that P is not so, without it being false that P is probable. Sometimes the improbable occurs. It is in this way that most people are mistaken in chiding the weatherman on any given day for a false assertion. The weatherman may have predicted sunshine for the day, causing you to leave your umbrella at home. If his prediction carried with it a probability, he can only legitimately attract rebuke if, in the long run, it rains more than 20% of the time that he predicts sunshine with an 80% probability.

According to Robert Brandom⁴², one of the earliest uses of the word “assert” is as a commitment to justification. Cognitive discourse, which consists of knowledge claims, takes the form of assertion. And knowledge claims purport to express appropriately justified true belief. To assert something, or put it forward as true, is to assume responsibility for justifying it. The young Frege treated asserting P as putting it forward as true and so supporting inferences from P.

In a social context, to assert P is to commit to the truth of P and to endorse it. That is, it is socially customary only to assert things with the intention of saying something that is true. Fiction is an exception to this social custom, since fictional assertions are not taken as commitments to their truth.⁴³

An assertion cannot form the basis for further inference unless it is defensible or justifiable in some way. Assertions are backed up by both personal and content-based authority: the former means that one can reference the assertor him or herself in defending the claim while the latter arises by way of reference to the content of other assertions. Bare assertions are an exception to this rule in that they rely only on personal authority and often by their nature cannot be bolstered by way of reference to other claims.

Considering the distinction between justified and true claims reveals important differences between the medical and legal context. An expert’s assertions in a legal trial may be justified by the methods she has used to arrive at them, but what ultimately matters is the truth as seen from

---


the judge or jury’s perspective. In medicine, each physician is not in the same way entitled to maintain a personal truth that is distinct from the justified assertions of her colleagues. Medical guidelines, established in a justifiable way, supersede the individual physician’s personal judgment on many matters: Dr. M may assert that all inflamed gallbladders should be removed, but if the guideline states that they are only to be removed under specifically defined circumstances, Dr. M. risks sanctions by her professional governing body if she removes an inflamed gallbladder in the absence of those specifically defined criteria.

In “Justified Assertion and the Relativity of Knowledge”\(^{44}\), Robert Hambourger explains that the strength of evidence one needs to be justified in making any particular assertion varies with the assertion itself, but also with the situation. The constraints on the evidence allowed into legal and medical discourse can be understood as situation specific in this way. One needs stronger evidence to say “P is so” than to say “I have a hunch that P”. But whether one is asserting P at a picnic or in a court of law is also relevant. Hambourger refers to these differences in terms of “standards of caution”. That is, a picnic requires a lower standard of caution than a court of law. There is a rich spectrum of standards of caution even in a single context and the standard can shift even in the course of a single conversation. The example Hambourger gives is as follows: One person asks another if she knows the time. The response: “Yes, it’s 4:45 p.m.”. The interlocutor then says, “Look do you know the time? I want to make a long-distance call and don’t want to be charged the full rate”. The response now shifts: “Well, no I’m not sure, but my watch is usually accurate and so it’s probably 4:45”.

An assertion can be considered epistemically justified if, under the circumstances, one has sufficient evidence or other grounds, to support it. Some statements are epistemically unjustifiable. An example of this would be Moore’s paradox “I believe that P, but not-\(P\)”. Contradictions in general will always be epistemically unjustifiable.

In any given situation, standards of caution are defined by the competing interests in speaking freely on the one hand, and the dangers of mistake on the other. Standards of caution may also be artificially raised where the dangers associated with mistake are not particularly substantial,

but there is a shared expectation of an elevated standard. Academic debate or posing questions such as “are you sure?” raise the standard of caution without necessarily introducing increased danger associated with error. Another way to think of the elevated standard of caution in such cases, however, is in terms of reputational danger. That is, in academic discourse or when an interlocutor has probed the speaker on the issue of certainty, the reputational damage risked by erroneous assertion rises.

Hambourger takes the idea of standards of caution further to suggest that knowledge is always relative to a particular standard of caution. Not only is there a loose sense of the word “know” in which we “know” many things, and a strict sense in which we know very little, Hambourger’s view is that there are perhaps infinitely many gradations of knowledge relative to a correspondingly varied spectrum of standards of caution. Whether we are speaking in casual conversation, in an academic debate, or considering skeptical problems, the applicable standards of caution will vary accordingly. In the first category of conversation, we may claim to “know” a great deal, while in the last we may concede to knowing nothing for certain (other than perhaps that we know nothing for certain). It is the goal of this chapter to shed light on the constraints that apply to the unique context of legal discourse. The rules of evidence that limit what is admissible into the realm of legal decision making will be discussed and placed in historical context. The question of whether there is a unifying theory of legal evidence will also be touched upon. Some understanding of these rules of legal play will be essential to considering the central question of how best to integrate explicitly statistical evidence into the conversation of law.

2.2 Verdict Accuracy vs. Verdict Acceptability: Constraints on Assertion and Justification in the Legal Context

One helpful way to think of constraints on legal discourse is in the context of standards of caution. The law limits the types of assertions that can count as evidence in a way that is consistent with the relevant standard of caution. Here the standard of caution is not to be confused with the standard of proof, which is also context dependent: in the criminal context the standard of proof required to secure a conviction is “beyond a reasonable doubt” while in the civil realm, a successful plaintiff need only make out her case “on a balance of probabilities”.
Before arriving at the question of whether the standard of proof has been met, there is a preliminary inquiry into whether an assertion meets a threshold standard. It is this threshold standard that I suggest is fruitfully contemplated in the context of standards of caution. It is this standard that is defined by law’s rules of evidence.

In the legal domain, “evidence” encompasses everything that decision makers consider in coming to their conclusions about the “facts” of the case. To decide what the facts are, jurors are told to consider only the evidence that they hear and see in the courtroom. Jurors are instructed that evidence is the testimony of witnesses, what each witness said in answering questions, and things produced as exhibits. It may also consist of admissions, namely the facts that the Crown and defence agree on.\(^{45}\)

Ultimately, legal conclusions are based on applying the governing legal principles to the facts as they have been found. In the context of a jury trial, it is the trial judge’s role to instruct the jury on the legal test while the jury’s job is to make “findings of fact”. In judge-alone trials, the judge makes legal determinations and findings of fact. When a decision is appealed, the legal principles applied are held to a rigorous standard of “correctness” while findings of fact must betray a “palpable and overriding error” to be overturned. That is, findings of fact made by the decision maker at first instance are entitled to considerable deference.

The legal definition of evidence is more restrictive than the general use of the term which, as defined in Achinstein’s Nature of Explanation, refers to “any matter of fact, the effect, tendency, or design of which, when presented to the mind, is to produce a persuasion concerning the existence of some other matter of fact: a persuasion either affirmative or disaffirmative of its existence.”\(^{46}\)

The rules of legal evidence control the information that enters into the realm of consideration. Given constraints on time and resources as well as other considerations as to what information should properly be excluded from consideration, the “evidence” that is actually considered by


decision makers is restricted. That is, as Thayer conceived of it, the law of evidence is a series of disparate exceptions to a single principle of free proof.  

The law of evidence controls fact-finding in three basic ways: (i) by guiding the decision maker in the deliberation process; (ii) by regulating the deliberation process; and (iii) by excluding certain information from consideration.

The law of evidence provides general guidance to fact finders by defining the *kinds* of things that count as legal evidence. For example, jurors are given general directions at the outset of the trial about what does and does not constitute evidence. That is, they are told that the testimony of witnesses and witnesses’ answers to questions as well as exhibits and admissions are the kinds of things that count as evidence. In contrast, the charge in the indictment that is read out at the outset of the case – which states that the accused has been charged with crimes x, y, z – is not evidence. What the lawyers and the judge say when they address the jury during the trial, and what people outside the courtroom say, also do not qualify as evidence. In this way, radio, television, newspaper and internet reports about a case are not evidence and jurors are instructed to ignore them completely and avoid all media coverage of the case they are deciding.

Regulation directs the jury about what it may or may not do in a way that is more than suggestive. That is, a judge may inform a jury that it is prohibited from making certain inferences or following particular lines of reasoning. For example, a jury instruction may caution jurors that they are prohibited from using the accused’s failure to testify as grounds for a finding of guilt. Or, if the accused does testify, the fact of his prior conviction for fraud may be permitted into evidence in order to undermine his credibility as a witness. A jury would not, however, be allowed to use this information as the basis for an inference that the accused was more likely to have committed the crime with which he is presently charged. The trial judge would therefore instruct the jury on the permissible and impermissible uses of such information: it would be permissible for a juror to use the fact of the accused’s prior conviction to undermine her degree of belief in his account of what he was doing on the night of the crime; it would be

---


48 *Supra* note 45.
impermissible for her to use the information to directly infer that the accused had committed the
particular crime in question.

Finally, there are exclusionary rules, which are the primary focus of theoretical inquiry into the
law of evidence. For example, the so-called opinion rule requires that non-expert witnesses
restrict their evidence to facts within their personal knowledge and not stray into the realm of
opinions or beliefs about those facts. In this way, a police officer providing evidence at a trial
would be permitted to testify that he saw a red mark on the finger of the accused, but would not
be permitted to opine on the genesis of such a mark. As a police officer is no expert on the types
of marks produced by different means – scratches, bites, knife cuts – his opinion on the mark’s
origin would be excluded evidence according to the rule against opinion evidence.\(^49\) There is
also a general exclusionary rule against the admission of what is called “hearsay” into evidence.
That is, the law generally eschews witness evidence based on the reports of others as opposed to
the witness’ own personal knowledge. The main reason for this exclusion is that unlike first-
hand testimony, hearsay cannot be tested by the rigors of cross-examination. The law of
evidence is also comprised of multiple “traditional” exceptions to the rule against hearsay that
developed in the common law. For example, excited utterances, statements expressing the
declarant’s impression of an existing condition, declarations of present state of mind, business
records, and prior inconsistent statements are considered exceptions to the general exclusionary
rule. There are a number of additional exceptions to the rule against hearsay that apply when the
declarant is unavailable as a witness. For example, dying declarations, admissions against
interest, prior testimony and admissions of guilt are all admissible if the declarant is unavailable
(as he always will be in the case of “dying declarations”). Also, if the party who is
disadvantaged by the hearsay statement brought about the declarant’s unavailability, he forfeits
his right to rely on the rule against hearsay. This means that if you “get rid” of a witness who
said he saw you at the scene of the crime, the hearsay report of another witness who heard the
statement of the first becomes admissible (although finding another witness to come forward
under such circumstances may prove tricky).\(^50\)

---


The three categories of controls on fact-finding are not mutually exclusive and do overlap. For example, many exclusionary rules are based on a presumption that the admission of certain evidence would most likely lead to prohibited lines of reasoning (e.g. once a criminal, always a criminal) or that decision makers would likely put too much weight on it (e.g. by relying on evidence that the accused was “not a nice guy” to undermine the presumption of innocence).51

The rules of evidence come from a number of different sources. Evidential law is made up of common law principles, that is, principles that have been developed by judges in the context of deciding particular cases over the years, statutory provisions that are enacted by elected legislators, and constitutional principles. The purpose of the rules of evidence is to allow relevant, but only relevant, information before the decision maker and also to honour certain fundamental policies of the law. Some exclusionary rules speak to a belief that some kinds of evidence are more likely to mislead than to guide judges and juries to the truth. Other exclusionary rules reflect the position that certain policy considerations trump the truth when it comes to legal determinations:

The rules of evidence control the presentation of facts before the court. Their purpose is to facilitate the introduction of all logically relevant facts without sacrificing any fundamental policy of the law which may be of more importance than the ascertainment of the truth. What are the logically relevant facts in any particular cases, whether civil or criminal, is decided by the substantive law governing the cause of action or offence set out in the pleadings or the charge, as the case may be. These matters can tangentially affect the evidentiary principles in any given case, but they do not make up a part of the law of evidence. Rules governing court practice and procedure can govern the conduct of litigation in a manner similar to evidentiary rules, but again, these are matters ancillary to evidence law...Evidentiary principles, on the other

---

hand, regulate (1) what matters are or are not admissible before the
court; and (2) the method by which admissible facts are placed
before it. [Emphasis added.]\textsuperscript{52}

Some of the fundamental policy objectives of the law that are distinct from the search for truth
are as follows: (i) The principle of party autonomy – namely that the parties and their lawyers
hold control over what is at issue, who are the witnesses and so on. This distinguishes the
Anglo-American adversarial process from, for example, the French inquisitorial proceedings in
which the judge acts as an investigator; (ii) The court as umpire – namely that the court is
inactive, passive, remote, neutral and independent; (iii) The principle of specialization of
functions – namely that questions of law, questions of fact and questions of disposition are kept
distinct and often decided by different people; (iv) The principle of orality – namely, that
arguments and witness evidence are presented orally in open court; (v) The principle of publicity
– namely that as a rule all proceedings are public to expose the process to scrutiny; (vi) The
principle that decisions are based on the issues and the evidence and not on a judgment about any
person as a whole; (vii) The principle of procedural fairness – namely, that the process for
arriving at decisions is one that is both efficient and perceived as in accordance with “due
process”; (viii) The principle of protecting the accused against mistaken conviction embodied in
the presumption of innocence; and (ix) The principle of the protection of suspects from improper
treatment.\textsuperscript{53} The supremacy of both the accused’s right to remain silent, and of solicitor client
privilege, are specific and powerful examples of policy objectives that are often at odds with
truth seeking. Perhaps above all, the law is concerned that justice not only be done but also be
seen to be done. In espousing this view, the law clearly is committed to values in addition to,
and at times over and above, the discovery of the truth in any given case. This given, it is
important to distinguish between verdict accuracy and verdict acceptability. A verdict may be
accurate and still unacceptable if it offends some legal policy objectives other than the search for
truth. At the same time, a verdict may be inaccurate but at the same time acceptable (until

\textsuperscript{52} John Sopinka, Sidney Lederman & Alan Bryant, \textit{The Law of Evidence in Canada}, 3d ed (Toronto: LexisNexis
Canada Inc., 2009) at 3.

\textsuperscript{53} \textit{Supra} note 47.
demonstrated inaccurate) if it was arrived at in a manner that is ordinarily truth-conducive and respects the other policy objectives of the law.

A potentially compelling example of a policy interest that may supercede truth detection involves the use of brain scans to detect lies. It is quite possible that at some point in the near future, functional magnetic resonance imaging (fMRI) will be proven highly reliable in the detection of mendacity by “seeing” inside the brain.\textsuperscript{54} An fMRI machine tracks blood flow to activated brain areas. The use of fMRIs in lie detection is premised on the idea that certain areas of the brain become engaged in the telling of a lie and that the regions that do more work get more blood. During lie studies, certain areas of the brain light up when lies are told that remain unengaged (and so unlit on scans) in telling the truth.\textsuperscript{55} While the technology and research is still underdeveloped, it is quite plausible that even if fMRIs were proven to be highly reliable in detecting lies, the legal system would not adopt them as an essential component of the legal process. It is arguable whether a refusal to adopt such new technology would be wise, but any proposal for its adoption would certainly be controversial. That is, it is clear that there are interests at stake here other than the goal of truth discovery. There is a high level of unease associated with the idea of relinquishing the privacy of our minds: from the husband who claims not to have noticed the attractive woman he just undressed in his mind, to the employee who tells his boss he’s happy to work over the weekend, white lies are as common as they are embarrassing if revealed. Accordingly, it is easy to imagine circumstances under which truth is not always put above all else.

To promote both truth-finding and other legal goals, specific rules of what should and should not be permitted into evidence have developed. Historically, the rules of evidence existed as an assortment of rules and exceptions, most of which were developed by the courts over time. The rule against hearsay, as well as the multiple exceptions to this rule, is a prime example of this pattern of development. The rule provides that, “written or oral statements, or communicative

\textsuperscript{54}The science of lie detection may already be sufficiently reliable for use in some criminal and civil trial contexts. See Frederick Schauer, “Can Bad Science Be Good Evidence? Neuroscience, Lie Detection, and Beyond” (2010) 95 Cornell L Rev 1191.

conduct made by persons otherwise than in testimony at the proceeding in which it is offered, are inadmissible, if such statements or conduct are tendered either as proof of their truth or as proof of assertions implicit therein.\(^{56}\) In this way, a judge or jury would be prohibited from considering a witness, John’s, statement that Bob said he saw the accused running away from the scene of the crime. The rule is motivated by the absence of truth safeguards for hearsay evidence that are thought to exist in the case of direct statements gained from first-person experience. That is, if John states that he saw the accused running from the scene of the crime, the following truth safeguards apply: the right of cross-examination, which is believed to ferret out inconsistencies characteristic of untruth, the oath or affirmation that John took before making this statement, and the threat of punishment for perjury. In the context of direct evidence, the decision maker also has the opportunity to observe John give his testimony and assess his credibility on this basis. Does John look shifty? Does he deliver his statement calmly and confidently, or is he uncomfortable and hurried in a manner suggesting an attempt to deceive?\(^{57}\)

Unfortunately, there is considerable research to suggest that people are in fact woefully unable to assess by observation whether or not a witness is telling the truth. For example, studies suggest that physiognomic, personality, cultural and behavioural characteristics of an individual tend to dictate an observer’s assessment of his or her truthfulness. That is, baby-faced, non-weird, extroverted and non-idiosyncratic people are more likely to be judged truthful than their harsh-featured, weird, introverted and idiosyncratic counterparts. In general, there is also a strong correlation between the attribution of positive traits to a person and the assessment that he or she is telling the truth. The inverse also applies; people are likely to assess as untruthful those to whom they attribute other negative characteristics such as ugliness, unfriendliness, or strangeness. Observers also tend to use mistaken cues for identifying truthfulness and deception. For example, it is widely believed that the failure to make eye contact is correlated with deception. However, since most liars know about this basic rule, they actually increase eye

\(^{56}\) Supra note 52 at 229.

\(^{57}\) In addition to the absence of what are thought of as truth safeguards, the rule against hearsay may also relate to certain process preferences in the law. That is, hearsay does not comport with the legal preference for direct assertion and reply between the involved parties. For more on this, see Bruce Chapman, “Defeasible Rules and Interpersonal Accountability” in J Ferrer & GB Ratti, eds, Essays in Legal Defeasibility (Oxford: Oxford University Press, forthcoming 2011).
contact when telling a lie.  

Other beliefs about the circumstances and detection of truth inform numerous exceptions to the general rule permitting exceptional admission of hearsay evidence. For example, where it is impossible or difficult to secure other evidence, where the author of the hearsay statement was not an interested party in the sense that the statement was not in his favour, where the statement was made before the dispute in question arose, or where the author of the statement had a peculiar means of knowledge not possessed in ordinary cases, hearsay may be exceptionally admitted.  

It is in this way that the rules of evidence and multiple exceptions to them developed over time.

The theory underlying these assorted exceptions is that all the circumstances behind them were thought to make the statements sufficiently reliable to be heard by a trier of fact. For example, excited utterances were thought to be reliable because the state of excitement was considered inconsistent with cunning design. A “dying declaration” was similarly considered to be reliable based on the belief that anyone facing death would have no earthly motivation for deceit and every reason to tell the truth in the face of imminent judgment at St. Peter’s gate.

While these rationales may strike us as quaint at best, it was not until very recently that the Supreme Court of Canada adopted a “modern” and “principled” approach to hearsay. In R. v. Starr, [2000] 2 S.C.R. 144, the Supreme Court indicated that the long-recognized exceptions to the general rule against hearsay may now be reassessed in light of the overarching requirements of necessity and reliability. At p. 243, Justice Iacobucci, writing for a majority of the court stated:


60 Supra note 52 at 245.
Up to the present, this Court’s application of the principled approach to hearsay admissibility in practice has involved only expanding the scope of hearsay admissibility beyond the traditional exceptions. The focus of the Court’s analysis and commentary has been upon the need to increase the flexibility of the existing exceptions, and not specifically upon the need to re-examine the exceptions themselves. However, this case requires that we examine an exception to the hearsay rule and determine its coexistence with the principled approach. As I will discuss further, to the extent that the various exceptions may conflict with the requirements of a principled analysis, it is the principled analysis that should prevail.

Justice Iacobucci indicated that the two main reasons for reconsidering previously accepted exceptions to the hearsay rule were trial fairness and the integrity of the justice system, as well as the intellectual coherence of the law of hearsay. At the same time, the court indicated (at p. 247) that the traditional exceptions would “continue to play an important role under the principled approach.”

As an example, in the 2008 case of R. v. Blackman, [2008] 2 S.C.R. 298, the Supreme Court of Canada considered the question of the admissibility of the deceased’s statements to his mother in establishing the motive of the accused. The accused was charged with first degree murder after the shooting death of E. The Crown’s theory was that the accused shot E in retaliation for E’s previous violent attack against him. Central to the Crown’s theory was E’s mother’s testimony that he had told her he had stabbed a man over a pool debt and was subsequently shot at by three men. While such testimony would ordinarily be excluded under the general prohibition against hearsay evidence, the court endorsed the application of the principled approach and a consideration of the factors of necessity and reliability to agree with the trial judge that the mother’s evidence was admissible. E’s death was sufficient to meet the criterion of necessity, since this evidence could not be entered by way of direct testimony. As for reliability, the trial judge found that the circumstances under which the statement had been made were not “inherently unreliable” and the Supreme Court endorsed the trial judge’s approach.
What the above illustrates is that what counts as “evidence” for the purposes of legal decision making is subject to a number of restrictions. These rules and restrictions, as well as the principles that guide them, are what constitute the law of evidence. Taken together, they define the standard of caution that applies to legal discourse and that is distinct from other conversational contexts.

The question remains, however, whether these rules and restrictions can be explained by a single overarching theory. That is, is there a principle that explains at a more fundamental level the controls placed on legal evidence by the various exclusionary rules?

2.3 Is There a Unifying Theory of Legal Evidence?

While many areas of the law are the topic of extensive theoretical inquiry, historically, the rules of evidence have received relatively sparse theoretical treatment. There are those who suggest that attempts at theorizing in the realm of evidence are unlikely to bear fruit; that is, that the rules of legal evidence have developed as a historical patchwork created by the courts, responding to situations as they arose, and that no unified theory either gave rise to these rules nor can the rules be explained retrospectively in a truly coherent way.

The common law of evidence provides a good example of what may be called Anglo-Saxon pragmatic evolution. According to William Twining in his Theories of Evidence, much of Anglo-American writing about evidence theory has been in this vein:

Despite these strains and disagreements there is a truly remarkable homogeneity about the basic assumptions of almost all specialist writings on evidence from Gilbert through Bentham, Thayer and Wigmore to Cross and McCormick. Almost without exception, Anglo-American writers about evidence show very similar assumptions, either explicitly or implicitly, about the nature and ends of adjudication, about knowledge or belief about past events

---

61 Supra note 47.
and about what is involved in reasoning about disputed questions of fact in forensic contexts.\footnote{William Twining, \textit{Theories of Evidence, Bentham and Wigmore} (Stanford: Stanford University Press, 1985) at 13.}

The basic assumptions adopted by this tradition arise from the fact that the Anglo-American system accepted a “rational” mode for determining facts in contrast to the medieval “irrational” modes. The “rational” mode for determining facts involved questions of evidence and argument. This was in contrast to “irrational” modes of proof that assumed that proof of fact was in divine hands. During the Middle Ages, irrational proofs included trial by ordeal and trial by battle. In a trial by ordeal, a person’s innocence was tested by subjecting him to danger. If the person survived, this was considered “proof” that God was protecting him and thus of his innocence. Ordeals included swallowing poison, pulling an object out of burning oil or walking over burning coals. If a person’s burn blisters became infected, he or she was judged to be guilty. Similarly, civil disputes were decided in a fight to the death. The survivor would be “judged” innocent since God would only protect the innocent.\footnote{Austin Sarat, Lawrence Douglas & Martha Merrill Umphrey, \textit{How Law Knows} (Palo Alto: Stanford University Press, 2007).}

In addition to valuing evidence and argument over divine revelation, the Anglo-American system assumed the classical empirical view of rationality found in the writings of Bacon, Locke and John Stuart Mill. Differences relate most commonly to the scope of and need for rules of evidence in general, and to the details of particular rules.\footnote{\textit{Supra} note 62.}

Common epistemological assumptions typical of the classic Anglo-American approach include the following: (i) events and states of affairs occur and have an existence independent of human observation; (ii) true statements are statements which correspond to the facts; (iii) present knowledge about past events is possible; (iv) in this context, “knowledge” means warranted beliefs that satisfy specified standards of proof relating to the truth of statements about facts in the real world; and (v) present knowledge about past events is typically based upon incomplete evidence. It follows from the first five assumptions that establishing the truth about alleged past
events is typically a matter of probabilities or likelihoods falling short of complete certainty, and that judgments about the probable truth of allegations about past events must, generally speaking, be based on the available “stock of knowledge” about the common course of events in the external world. In any given society at a particular time, the “stock of knowledge” includes, in a descending scale of probability, generalizations accepted by the scientific community as established, the opinions of experts, and “common-sense” generalizations based on the experience of members of society. 65

An early important treatise by Geoffrey Gilbert attempted to treat all of evidence law under the single principle of the “best evidence rule”. The idea here was that it was the main task of the law to base its decisions on the best evidence available under the circumstances. From this idea, Gilbert endeavoured to establish a formal hierarchy of evidence with public records crowning that hierarchy. Gilbert also attempted to ground his theory of evidence in the notion of probability. His treatise on the law of evidence, first published in 1754 and then reissued four times by 1801, opens with this passage from Locke:

In the first Place, it has been observed by a very learned Man, that there are several Degrees from perfect Certainty and Demonstration, quite down to Improbability and Unlikeliness, even to the Confines of Impossibility; and there are several Acts of the Mind proportion’d to these Degrees of Evidence, Which may be called the Degrees of Assent, from full assurance and Confidence, quite down to Conjecture, Doubt, Distrust and Disbelief. 66

Bentham’s first work in evidence in the early 1800s focused on finding the flaws in Gilbert’s approach. In particular, he objected to Gilbert’s emphasis on formal, rigid rules to govern the


weighing of evidence. Also, by referring exclusively to the rules of probability, according to Bentham, Gilbert had conflated questions of evidentiary admissibility with those of evidentiary weight. Bentham placed high value on the discovery of true facts in the decision-making process, but also emphasized utility and efficiency. Much of Bentham’s work spoke to eliminating many formal rules of evidence. Due to his efforts, rules excluding parties and others as competent witnesses as well as formal rules governing questions of weight or credibility of evidence no longer exist. The one area where evidentiary constraints still exist is in exclusionary rules, such as the rule against hearsay, referred to above.67

It was in the early 1900s that John Henry Wigmore entered the legal scholarship scene and created what is to this day the most complete and exhaustive treatise on the Anglo-American law of evidence. The aim of his work is set out in the preface to the first edition of his Treatise:

First, to expound the Anglo-American Law of Evidence as a system of reasoned principles and rules; secondly, to deal with the apparently warring mass of judicial precedents as the consistent product of these principles and rules; and thirdly, to furnish all the materials for ascertaining the present state of the law in the half a hundred independent American jurisdictions.68

For Wigmore, the objective of academic analyses was to evaluate the evidence and determine whether a case was rightly, or at least justifiably, decided with reference to some articulated standard. (Of course there can be a significant difference between the former and the latter, but this does not appear to be a distinction upon which Wigmore was particularly focused.) For Wigmore, the evaluative standard might be articulated as follows: given the substantive law and

the burden of proof found applicable by the court, did the evidence support the factual conclusion necessary to sustain the result?\textsuperscript{69}

To give a sense of the Wigmorean approach, consider this “Wigmorean protocol” for analyzing problems in the use and admissibility of evidence:

1. What is the ultimate probandum [thing that needs to be proven] for the case in which the proffered evidence is being offered? What are the penultimate probands?
2. To what fact or consequence is the proffered evidence relevant? Specify the necessary inferential steps.
3. Is there a rule of evidence that the opponent could invoke to challenge the admissibility or use of the proffered evidence? If so, specify the rule, and make the argument, initially in syllogistic form.
4. What, if any, are the arguments that the proponent of the evidence may make to avoid the application or minimize the effect of the rule or rules? Begin by challenging the opponent’s rule-based argument.
5. Unless it is clear that the evidence must be excluded under the rule or rules invoked, what is the legitimate probative value of the proffered evidence? What are its improper prejudicial effects?
6. What is the strongest argument that can be made to persuade the court that the improper prejudicial effects substantially outweigh the legitimate probative value?
7. Should the proffered evidence be admitted? If admitted should its use be limited? Why?\textsuperscript{70}

In the 1940s to 1960s a number of States produced evidentiary codes and consolidations, for example, the California Evidence Code. In the 1970s a lively debate began in the United States about the uses and abuses of mathematics in litigation. This debate was partly stimulated by a


\textsuperscript{70} \textit{Ibid.} at 306.
high-profile case, *People v. Collins*, which was described in the previous chapter and which prominently featured mistaken reasoning with probabilities relating to base rate neglect.\(^\text{71}\)

It was partially in response to this that in 1977 the British philosopher L. Jonathan Cohen put forward the position that reasoning about probability in law wasn’t mathematical (Pascalian) but could be approached using non-mathematical (Baconian) criteria. Cohen uses the term “Pascalian”, which for our purposes is interchangeable with “Bayesian”. Since my project assumes a mathematical approach to reasoning with probability in the law, I will take some time to discuss what I consider to be the limitations of Cohen’s proposed approach.

### 2.4 Cohen’s Non-Mathematical Approach: A Critique

In *The Probable and the Provable*, Cohen takes an innovative approach to evidence theory and attempts to demonstrate that what is loosely called “probable” reasoning in the courts as well as in everyday explanations, does not involve a Bayesian probability satisfying the usual axioms of a probability calculus. Instead he suggests an alternative inductive probability which he argues resolves many of the problems with the Bayesian treatment.\(^\text{72}\) Cohen proposes a method whereby rival hypotheses may be graded ordinally in accordance with the fact-finder’s conclusions as to how well each hypothesis is accounted for by the evidence, how well each withstands attempts to eliminate it based on the evidence, and the extent to which any given hypothesis shows itself to be evidentially superior to the others in terms of completeness and quality of the evidence.\(^\text{73}\)

For example, if the hypothesis (H) is that the defendant (D) stole a diamond worth $1,000 from the victim (V), and the evidence (E) is that D took the diamond from V’s house and placed $1,000 in cash on the table, any fresh evidence that the diamond was in fact worth $1,000 or less would be a relevant variable tending to undermine H. On the other hand, further evidence that the diamond was worth $10,000 or more would tend to support H. Analyzing each new piece of

\(^{71}\) *Supra* note 67.


evidence in relation to the range of competing rival hypotheses allows one to rank the hypotheses in order of strength based on the evidence. Cohen describes this process as “Baconian” to distinguish it from the Bayesian process of arriving at a single arithmetical probability assessment for a hypothesis.\textsuperscript{74}

In a way, Cohen’s contrast between what he calls the “Baconian” and “Bayesian” assessment slides by the central area of divergence between the two approaches. If his point is that precise probabilistic calculations are often impossible in the context of legal decision making, he need have no quarrel with the Bayesians. But his claim goes further than that to deny that a Bayesian model is normatively compelling and to argue that the “Baconian” model is preferable. Let’s revisit the diamond example above. A Bayesian might approach it in this way: in instances where theft is alleged and the defendant and victim had no prior communication about the diamond, a legitimate business transaction is an unlikely explanation. However, fresh evidence that the diamond was worth $1,000 or less would likely cause the Bayesian to update his belief in the plausibility of this explanation. Since thieves do not exchange money of equal value for the goods they take, the fresh evidence provides a very strong reason to update one’s belief to favour misunderstanding as the most likely explanation. In this way, the Bayesian still “ranks” competing hypotheses; the only difference is that the Bayesian’s ranking is informed by fundamental probabilistic principles, such as the relevance of base rates, while the Baconian ranking is not.

Cohen aims to support his approach on the basis that forensic proof cannot be analysed in terms of the mathematics of chance given how serious and numerous the paradoxes are to which this analytical framework gives rise, and that a non-Bayesian concept of probability based on degree of provability is a superior option. I will discuss – and hopefully dispel – the six paradoxes that he discusses.

I suggest that Cohen’s so-called paradoxes generated by a mathematical approach to legal analysis can be avoided in one of three ways: either (i) by denying that his descriptions of legal

\textsuperscript{74} Ibid.
The first paradox concerns the difficulty about conjunction. That is, the rule for civil suits requires a plaintiff to prove each element of his case on a balance of probabilities where each element is construed as independent. According to Cohen, if this probability is construed as mathematical probability, the conjunction principle for such probabilities would impose some curious constraints on the structure of the proof. Namely, each element would have to be proven to an extremely high degree of probability in order for the joint outcome to be greater than 0.5.

In civil suits such as a claim in negligence, the law requires the plaintiff to prove the following elements: (i) the existence of a duty of care; (ii) the breach of this duty; (iii) that the breach of the duty was a proximate cause of the harm suffered by the defendant; and (iv) that the plaintiff actually suffered harm. The law’s requirement is that each of these elements be proven on a balance of probabilities: “the plaintiff in a tort action has the burden of proving each of the elements of the claim on the balance of probabilities. This includes proving that the defendant’s impugned conduct actually caused the loss complained of.”

If a plaintiff slips and falls on a neighbour’s ice-covered sidewalk, then in order to succeed in a claim for negligence, she must prove that the neighbour had a duty to take reasonable precautions to keep her sidewalk clear, that she failed to take such reasonable precautions, that it was the failure to take those precautions that was the proximate cause of the defendant’s fall, and lastly that the plaintiff did indeed suffer some harm or loss as a result of the fall.

The existence of a duty of care is a purely legal, rather than a factual (or mixed factual and legal) question, and so is not properly characterized as probabilistic in nature. The answer to the question of whether or not there was a duty of care to take reasonable precautions to keep her sidewalk clear will either be a “yes” or a “no”. In Toronto, for example, there is a bylaw that requires non-senior, non-disabled residents to clear snow within 12 hours of a snowfall. While the remaining elements will be proven to some degree of probability, they are neither entirely independent nor are they unlikely to be proven to a high degree of probability. Two events are

75 Stewart v Pettie, [1995] 1 SCR 131 at para 60.
“independent” in the technical probabilistic sense if the outcome of one has no effect on the outcome of the second. It is only the probability of simultaneous occurrences of independent events that is the product of the probabilities of these events. The questions of whether the defendant failed to clear her sidewalk, whether this failure caused the plaintiff’s fall and whether the plaintiff was hurt as a result, are not independent of one another. Once you have an icy sidewalk, you have an increased chance of someone falling. And once someone falls, the chances of injury increase substantially. So, while the law requires the plaintiff to make out each element on a balance of probabilities, Cohen misleads by characterizing these legally distinct inquiries as factually independent. Once the probabilities are construed as dependent, it is no longer correct, as a matter of probability, to multiply them to generate the likelihood of their combined occurrence. When events are dependent, that is, when two outcomes are related to one another, the combined probability is the product of the probability of obtaining one event multiplied by the conditional probability of obtaining the other event, given that the first event has occurred:

\[ p(A \& B) = p(A) \times p(B|A) \]

It is not then, as Cohen suggests, that a probabilistic analysis leads to a paradox, but rather that an incorrect application of probabilistic principles does.

This is not to suggest that the elements in a tort claim are always dependent. That is, there may well be cases in which some or all of the elements to be proven are truly independent. In these cases, however, it would simply be correct to say that where two elements are established with a 70% probability, looked at together, the likelihood of both elements combined is only 49%. This is no paradox though. You should be less convinced that two uncertain, independent, events occurred than that one of them did. If there’s a 70% likelihood of rain in the morning and a 70% likelihood that your favourite coffee shop will be closed at 9 a.m., the chance that you will arrive at work both wet and caffeine-deprived is less than the chance that only one of these unfortunate events will occur. In a legal negligence case, if it is uncertain both that there was a breach of the duty – for example, the defendant maintains that he cleared his sidewalk of snow, but there is also evidence to the contrary – and it is uncertain whether the plaintiff actually suffered any harm – the plaintiff complains of continuing back problems from the fall but there are also indications to the contrary – then the chance that the defendant is liable in negligence is less than if there was uncertainty with regard to only one element.
In this case, however, I would suggest that the result generated by the application of Bayesian probability is not paradoxical but rather is unproblematic. In this instance, multiple sources of uncertainty should and do raise the overall level of uncertainty.

The second so-called “paradox” raised by Cohen relates to inference upon inference: Cohen claims that whereas in law there is a restriction that all inferences save the last must be established either on a balance of probabilities or beyond a reasonable doubt depending on the overall standard of proof, Bayesian probability does not impose any such limitation.

The obvious response to this criticism of Bayesian probability is that mischaracterizes the law. Inferences are never the focus of the standard of proof. The only question is whether taken as a whole the ultimate legal question has met the applicable standard of proof.76

The third difficulty raised by Cohen relates to the non-Baconian approach to negation: because of the principle that \( p(S) = 1 - p(\sim S) \), Cohen states that the mathematical analysis implies that in civil cases, the system is officially prepared to tolerate a quite substantial mathematical probability that the losing defendant deserved to succeed. That is, the probability that the trial outcome is wrong might be as high as, say 0.499.

While Cohen casts this as a difficulty about a probabilistic characterization of the legal system, it is really just an unfortunate reality about the law itself. There is nothing unique about this. All decisions made on the basis of “majority rules” are susceptible to the same criticism. That is, we accept in such cases that 49.9% of people may be opposed to the decision made on this basis. A standard of proof of a “balance of probabilities” means precisely that the probability that a trial outcome is wrong might be as high as 0.49. So long as matters aren’t decided on the basis of an error probability greater than 0.5, the system is working as well as it can. If the civil system is viewed as dedicated to re-apportioning losses, then it is appropriate to decide a case in favour of the party in whose favour the scales tip even slightly. In the standard civil case, someone has suffered a loss. There is no way to undo that fact. What can be done, however, is to re-apporion that loss where appropriate. If the loss was caused by the party who was injured or purely by

chance, then the loss rests where it falls. If, on the other hand, there is another individual who acted wrongly and is responsible for the loss, it is just that she should at least share, or in some cases bear the full equivalent of, the burden of the loss suffered. Viewed in this way – that is that losses happen and must either be borne by one or several people – it is not troubling that there could be a 0.499 probability that a trial outcome is wrong.

In fact, there is good reason to believe that an “all or nothing rule” that fully indemnifies the plaintiff where liability has been established to a degree greater than 0.5, and entirely denies damages where the probability is less than 0.5, optimally minimizes error. Error is defined here as any amount of money that either party is ordered to pay when it was another who was responsible for the loss. That is, if you compare two possible decision-making rules, namely an “expected value” rule and a “maximum likelihood rule”, where the former apportions liability in proportion to the degree of evidence (e.g. giving 49% of the total damages incurred where the probability of negligence has been established with a probability of 0.49) and the latter abides by the current all-or-nothing rule, the existing maximum likelihood rule actually minimizes overall error. In general terms, this result is explained by the fact that the expected value rule “gets it a little bit wrong” in every case; that is, in each individual case, the defendant is either fully or not at all responsible such that apportioning damages according to degree of proof never gets it exactly right. The maximum likelihood rule, on the other hand, gets it totally wrong, but only in a few cases.77

Let’s look at an example to illustrate the point. Imagine that you have 10 cases, each involving a $100 loss. As a matter of fact, the defendant is fully responsible for the losses in 6 of those cases. Now let us assume that the plaintiff in each case is able to marshal evidence demonstrating the defendant’s liability with a 60% probability (this is consistent with a base rate probability of defendant responsibility equal to 6 out of 10 cases). According to the expected value rule, the plaintiff in each case would be awarded $60, consistent with the 60% probability of liability supported by the evidence. On the maximum likelihood rule, however, the plaintiff would be awarded the full $100 of damages in every case since in every case, the plaintiff will

have “proved” her case on a balance of probabilities, namely to a probability of 0.6. Now the question is which approach minimizes error? On the expected value rule, the defendant pays $60 in each case. At the same time, the defendant actually owes $100 in 6 cases and $0 in 4 cases. This means that an amount of $40 is misallocated in 6 cases (the difference between the $100 actually owed and the $60 awarded) and $60 is misallocated in the other 4 (the difference between the $60 owed and the $0 awarded). A simple calculation reveals that this totals $480 in misallocated money: (6 cases x $40) + (4 cases x $60). On the all-or-nothing rule, on the other hand, the defendant pays $100 in every case. Since the defendant actually owes $100 in 6 cases and $0 in 4 this amounts to an error in allocation of just 4 cases x $100 = $400: namely, $80 less than on the expected value rule.

All this is to say that Cohen’s observation that the probability that the trial outcome is wrong might be as high as 0.499 is actually not so troubling. Someone has to bear the loss and there does not appear to be an alternative that better allocates losses.

A fourth “paradoxical” point put forward by Cohen is this: we are more inclined to doubt a conclusion because there is a particular, specifiable reason for doubting it, than to doubt the conclusion because it falls short of certainty. The significance of the statistical probability has to be assessed in light of other evidence by a mode of assessment, which need not itself have anything to do with mathematical probability.

The point that Cohen is making here is taken up by others such as Alex Stein, Ronald Allen and Michael Pardo. The academic legal literature is replete with attempts to formulate a justification for the rejection of statistical evidence in favour of “direct” evidence. Direct evidence, and first-hand witness testimony in particular, is in fact favoured by common law courts. But as Amit Pundik argues in “What is wrong with statistical evidence: The attempts to establish an epistemic deficiency”, all attempts to meaningfully distinguish explicitly statistical evidence from direct evidence fall short. Many of the traditionally accepted types of evidence are highly

78 Supra note 72.

fallible. Our inability to accurately assess credibility from demeanor is just one example.\(^\text{80}\) If the law is to be consistent, then it should not hold other types of evidence up to a higher standard. In fact, there is good reason to think that the law inappropriately holds scientific evidence, which is usually presented in statistical form, up to a higher standard of reliability than it does other types of evidence. As Frederick Schauer suggests in “Can Bad Science be Good Evidence? Neuroscience, Lie Detection and Beyond”, scientific evidence that is not yet sufficiently reliable or valid for scientific purposes may still be good enough for a variety of legal ends. There is good reason to suppose that the question of threshold reliability for scientific evidence should relate to the evidence’s purpose rather than to scientific standards.\(^\text{81}\) By relying on scientific standards to govern the admissibility of many kinds of statistical evidence, the law stacks the deck against such evidence without justification.

Whether it is Cohen’s concept of weight, or Stein’s notion of case-specificity and the problem of creating new knowledge, there is no \textit{mathematically} principled reason to ignore evidence that 99% of concert goers failed to pay for their tickets in considering witness testimonial that defendant X slid in under the fence. Witness evidence should always be subject to scrutiny on the basis of non-case-specific information, such as the information that eyewitness evidence is only 80% reliable, or that brothers are more inclined than strangers to lie in each other’s favour. Statistical evidence is inescapable and the onus would be on those who eschew non-specific evidence to establish a principled reason to reject some varieties and not others.\(^\text{82}\)

Unfortunately for defenders of the law’s reluctance to embrace statistical evidence, the bald fact that we are more inclined to think in a certain way does not make it right. That is, attempts to epistemically justify certain legal practices tend to conflate the descriptive and the normative.


\(^\text{82}\) Alex Stein argues that a notion of “causal probability” is better suited to the law than mathematical probability. Causal probability would dictate that a man diagnosed by a doctor as having a disease should act on a belief in this disease and ignore base-rate evidence since it is not causally related to him in any way. I, of course, think that this is just wrong. See Alex Stein, “The Flawed Probabilistic Foundation of Law & Economics” (2011) 105 Nw U L Rev 199.
There are plenty of examples of ways in which we are inclined to think that are just plain wrong. For example, it is a well-known cognitive bias that we prefer vivid anecdotal evidence to impersonal data collected from a large sample. The fact that your sister suffered major complications from laser eye surgery will act as a far stronger deterrent than the presentation of data that serious complications occur in 0.1% of the population. In the medical context, our inclination to feel better when given a pill with no active ingredients is not interpreted as proof that sugar pills are health enhancers, but rather as proof that we cannot always trust our inclinations.

In addition, Cohen tells us that no familiar criterion of mathematical probability is applicable to the evaluation of juridical proofs. Statistical criteria are inapplicable. This is the fifth “paradox” that he raises. To suppose that jurors should evaluate proofs in terms of a coherent betting policy is to ignore the fact that rational men do not bet on issues where the outcome is not discoverable otherwise than from the data on which the odds themselves have to be based. That is, the model of a betting quotient is not applicable to the legal context: a man only wagers on discoverable outcomes.

But surely it is untrue that we only bet on discoverable outcomes. Every day we are faced with choices – big and small – about our lives. Many, perhaps even most of them, have non-discoverable outcomes in the sense that we never know what would have been had we taken a different path. Should I take a job as a researcher or as a technician? Hardly any of the decisions we make have fully discoverable outcomes. I may assess that I have an 80% chance of happiness in my life as a researcher, but I will never know if this is actually the case. I will either end up happy, in which case I may think I was right in my assessment (and in any case I won’t care since I’ll be happy!). Or, I may end up miserable. On either outcome, my 80% assessment may have been spot-on or wildly misguided. It is only if I predict an outcome with 100% certainty that the outcome has the potential for discoverability. And only a fool makes predictions of certainty.

Even more to the point, betting represents only one of many models for understanding probability. That is, it is in no way necessary to represent probability-based decision making as betting on outcomes. We make decisions all the time. In making these decisions, we can use probabilistic information without modeling our decision making as wagering.
The sixth and last “paradox” put forward by Cohen pertains to Bayesian probability’s consideration of prior probabilities in the face of the law’s insistence on the presumption of innocence.

There are a number of possible responses to this concern. Perhaps the simplest for present purposes is that the law is perfectly free to stipulate a low prior probability as a starting point in the process. This is just what is meant by a presumption of innocence; that is, importing a high prior probability of guilt or responsibility is expressly prohibited. A stipulation of this kind places a constraint on the Bayesian probability analysis but in no way undermines its usefulness.

In sum, the so-called “difficulties” with a probabilistic understanding of legal evidence as described by Cohen can be avoided in one of three ways: either by denying that his descriptions of legal evidence are apt, by denying that probability dictates the conclusions he describes, or by denying that the conclusions he describes are problematic. In any case, all I hope to have demonstrated with the above analysis is that Cohen has failed to show that a probabilistic account is necessarily problematic. The ultimate question will be whether it is the probabilistic account that is most compelling. That assessment can only be made in light of what is to come.

### 2.5 Looking Forward

What is perhaps remarkable about the history of the Anglo-American approach to the law of evidence is how little the nature of the debates has changed. Debates about presumptions, hearsay or the best evidence rule, about the pros and cons of the jury or the adversarial system or judge-made law versus codification have changed little since the 1700s. As Twinning observes, “[t]here is a remarkable degree of continuity about some of these controversies: for example, a reader of the *Edinburgh Review* in the eighteen-twenties and thirties would have found a great deal that was familiar in recent debates in England about the reform of criminal evidence and procedure.”

---

83 *Supra* note 67 at 75.
This is not only remarkable but unfortunate. In particular, I will argue that the law’s reluctance to embrace mathematics and scientific results expressed in mathematical terms comes at an unacceptable cost. Twinning puts it well when he says that,

…theories of probability are central to evaluation of evidence and to elucidating basic evidentiary concepts such as ‘standards of proof’, ‘probative value’ and ‘prejudicial effect’; and statistical evidence of increasing practical importance in many different types of proceeding ranging far beyond the familiar examples of paternity, discrimination and fingerprinting. The lawyer of today needs to be a master of elementary statistics.84

The focus of the next chapters will be how basic probabilistic evidence should best be presented to and used by legal decision makers. The avoidance of rudimentary errors will be of particular importance. Statistical evidence is regularly before the courts and should be before them even more often. While statistical evidence may be misused, so may the traditional types of evidence. The sooner the law develops a useful approach to integrating statistical evidence into the decision-making process, the better off we will be.

84 Supra note 47 at 224.
Chapter 3
Bayesianism: A Legitimate Epistemological Approach to Statistical Evidence

3 Bayesianism

3.1 Introduction to the Theorem

“‘BEWARE of geeks bearing formulas.’ So saith Warren Buffett.” So begins a recent article in the New York Times by Richard Dooling. Richard Dooling’s musings concern the role of technology in the current episode of financial unrest, but a similar fear of formulas can also be found in the context of statistical evidence and legal decision making.

The purpose of this chapter is to discuss what is meant by a Bayesian updating approach and so to elucidate the discussion that will follow. The basics of Bayesianism are set out, as well as the role of Bayesianism in an attitude towards knowledge. The role of priors is discussed and some challenges to Bayesianism are considered. The hope is that this chapter will be persuasive in suggesting that Bayesianism is a good candidate for a normative model against which the law can distinguish right from wrong when it comes to using and interpreting statistical evidence.

The chapter covers the following topics:

- The derivation of Bayes’ theorem
- Interpreting probabilities
- The role of prior probabilities
- Encouraging good Bayesian reasoning
- Some challenges to Bayesianism as a normative model
- An application to medicine
- An application to the law

The “Bayesian” label is used to identify many different kinds of claims and to mark many different distinctions. For example, claims and/or distinctions that are often thought to define “Bayesianism” include the following: that belief is a matter of degree and degrees of belief may be measured numerically; that degrees of belief should obey the axioms of probability; that statistical reasoning requires a prior probability distribution to proceed; that people have prior
distributions in their tacit degrees of belief; that probabilities should be updated by conditionalization; or that decisions should be made in accordance with the expected-utility maximization rule.

For our purposes, it will suffice to think of Bayesianism as an approach to learning under circumstances of uncertainty. It is the business of statistics to quantify uncertainty. Any statistical approach is defined by how it deals with this uncertainty. The “Bayesian” approach is a model of learning whereby as new information becomes available, you update what you know. Put in very basic terms, the Bayesian is persuaded by the following reasoning:

(i) It is plain silly to ignore what we know; (ii) It is natural and useful to cast what we know in the language of probabilities; and (iii) If our subjective probabilities are erroneous, their impact will get washed out in due time, as the number of observations increases.\(^{85}\)

Add to this that one should update one’s belief in light of new evidence and you have the basics of the Bayesian approach.

3.2 The Derivation of Bayes’ Theorem

Before proceeding any further, a few words must be said about definitions and the basic mathematical derivation of Bayes’ theorem. Bayes’ theorem reveals a fundamental relationship between the conditional and non-conditional (also called marginal) probabilities of two events, A and B. More specifically, the theorem states that the probability of A given that B, on the one hand, is equal to the product of the probability of B given that A and the probability of A, divided by the probability of B:

\[
p(A|B) = \frac{p(B|A) p(A)}{p(B)}
\]

Conditional probabilities are those that relate the probability of one event to another. For example, the probability “p” of B given that A is conditional since the probability of B depends upon the occurrence of A. The convention is to write conditional probabilities using a vertical bar. “p(B|A)” is read as “the probability of B, given that A”; that is, the probability of B taking A as certain. In contrast, a marginal probability, such as the probability of B, p(B), can be obtained by adding up all of the conditional probabilities of that same event B, each multiplied by the marginal probability of the condition. That is, the marginal probability of B is equal to the probability of B given that A, multiplied by the probability of A, plus the probability of B given that not-A (~A), multiplied by the probability of ~A:

\[ p(B) = p(B|A) p(A) + p(B|\sim A) p(\sim A) \]

Probability values range from 0 to 1 with “0” representing impossibility and “1” representing certainty.

Bayes’ theorem is derived directly from Kolmogorov’s probability axioms and the definition of conditional probability. It is therefore only the interpretation and application of the theorem – not the theorem itself – that is the subject of some controversy.

Bayes’ theorem can be derived by substituting the following identities into:

\[ p(A|B) = \frac{p(A \& B)}{p(B)} \]

This probability equation is true by definition, as long as p(B) ≠ 0. The identities (which are fairly apparent on their face) are as follows:

(i) \[ p(B) = p(A \& B) + p(\sim A \& B) \]
(ii) \[ p(A \& B) = p(B | A) p(A) \]
(iii) \[ p(\sim A \& B) = p(B | \sim A) p(\sim A) \]

---

Using identity (ii), $p(A|B) = \frac{p(A \& B)}{p(B)}$ can be transformed into this statement of Bayes’ theorem:

$$p(A|B) = \frac{p(B|A) p(A)}{p(B)}$$

By relating conditional and marginal probabilities in this way, Bayes’ theorem allows you to calculate the probability of some fact of interest “A” given that you know something about some other related fact, “B”.

### 3.3 Interpreting Probabilities

But what do we mean by probability anyway? There is considerable controversy about how best to define probability. Many interpretations have been proposed and important criticisms raised. For example, it is argued that the frequency interpretation does not make sense in the single-case, that subjectivist accounts lead to arbitrariness in probability assignments, and that logical interpretations, while applicable to the gambling context, are of little use to science. Given the difficulty in devising a satisfactory definition or interpretation of probability, it is quite possible that we will ultimately arrive at some kind of deflationary theory of probability. Just as a deflationary theory of truth sees nothing useful to add to the concept of truth than that “Snow is white” is true if and only if snow is white, so too may we ultimately conclude that there is nothing useful to add to a definition of probability beyond what we ordinarily understand it to mean. Our efforts are likely best directed towards learning how to manipulate probabilities rather than define them.

Probability interpretation nonetheless enjoys a rich history. The first serious attempts to calculate probabilities are attributed to Blaise Pascal (1623-1662) and to Pierre-Simon Laplace (1749-1827) who focused on calculating probabilities associated with games of chance. Around the same time, probabilities began to be used for other purposes as well: by merchants contemplating the safe arrival of shipments, by historians considering past significant events, by

---

87 This is a view held, for example, by Timothy Williamson, *Knowledge and its Limits* (Oxford: Oxford University Press, 2000).
theologians reflecting on miracles, and also by legal scholars interested in the credibility of multiple witnesses to an event.  

The frequency interpretation of probability defines probability as the number of times an outcome occurs relative to the total number of events. For example, the frequency of a coin flip resulting in heads is the actual number of times you get heads relative to the actual number of coin tosses you performed. This interpretation is obviously limited in that it can only apply where you can actually count the total number of events. While all events are one-of-a-kind, the frequency theory assumes that one-of-a-kind events can be fruitfully classified as part of more general class such that a particular coin flip is also part of the class of coin flips more generally. Interpretations with wider applicability are the finite frequency theory of probability or the long-run relative frequency interpretation, which define probability as the number of times the outcome occurs relative to the number of times that it could have occurred, or the limiting frequency with which the outcome appears in a long series of similar events. This approach was developed by Venn in *The Logic of Chance*, 1866 and Reichenbach in *The Theory of Probability*, 1935.

Propensity theorists think of probability as a physical propensity, disposition, or tendency of a physical situation to give rise to an outcome of a certain kind. In this way, propensities are conceived of as the causes of observed probability patterns. Each interpretation of probability has its own advantages and is better-suited to certain contexts than to others. The frequency and propensity theories of probability are objective theories of probability since they present probability as an expression of the tendency of the world to operate in a certain way. In contrast, subjective or epistemic probability theories view probability not as a feature of the external world but rather as an expression of our uncertainty.

---


about the likelihood of some event. Formally, probability remains unchanged by the
interpretation that is imposed upon it.\textsuperscript{90}

When the probabilities generated by the application of Bayes’ rule are construed as a measure of
an agent’s uncertainty, Bayesianism becomes an epistemic theory of probability.\textsuperscript{91} That is, so-
called subjectivist Bayesians interpret probabilities as reflections of an agent’s uncertainty about
states of affairs instead of as a feature of the world. Viewed in this way, Bayes’ theorem can be
construed as a model for learning: taking one’s subjective probability of “A” and then revising it
in light of new information, “B” and the relationship of “B” to “A”.

3.4 It is Plain Silly to Ignore What We Know: The Role of Prior Probabilities

As mentioned above, the Bayesian accepts certain core propositions. The first is that you should
always consider what you already know about something when contemplating new information
about the topic.

Bayes’ theorem crucially revolves around the notion of base rates and prior probabilities. The
terms “base rates” and “prior probabilities” are often used interchangeably although there is a
clear distinction between them in some versions of Bayesianism: “base rates” refer to objective
probabilities whereas “prior probabilities” may represent a subjective degree of belief.
Epistemic Bayesianism equates prior probabilities with a subject’s existing belief about a state of
affairs. If, as the result of experience we come to believe proposition P to a degree of probability
x, then when new evidence arises, a good Bayesian must adjust her probability estimate
accordingly. The new probability function is obtained from the old by updating the probability
in light of the new evidence.


\textsuperscript{91} Bayes’ theory was first published by his friend, Richard Price, in 1763. Beginning in the 1920’s, an alternative,
“classical” approach to probability, advocated by RA Fisher, Karl Popper, and others, achieved dominance. It was
not until the late 1980s and early 1990s that the Bayesian approach to science began to gain favour.
Diagnostic medicine provides a very intuitive and relatively uncontroversial example of the importance of prior probabilities and the intuitive process of updating our knowledge about the world in light of both prior probability and new evidence.

Nervous Nelly is over 40 and tests positive for breast cancer. You know this about the test: it correctly identifies the disease 80% of the time. That is in 80% of cases where there is cancer, the test correctly identifies it. Conversely, the test misses 20% of patients who have cancer (false negative). The test also has a 15% likelihood of yielding a false positive. That is, in 15% of cases where there is no cancer, a positive result will arise. What is the likelihood that Nervous Nelly actually has the disease? Should she quit her job, withdraw every last penny from her bank account and “live it up” assuming she has nothing to lose? Or, is she in fact more likely than not to be cancer free?

The Bayesian updating approach requires that you take into account your existing knowledge of the prevalence of the disease in the population in arriving at your answer to these questions. That is, if you already know that 1% of women over age 40 have breast cancer, then your answer to the question, “how likely is it that Nervous Nelly in fact has the disease” will be very different than if you know that 50% of women in that population have the disease.

The application of Bayes’ formula yields the result that if 1% of women over 40 have breast cancer, given the parameters of the test accuracy described above, only 5.1% of women who get a positive test result will in fact have cancer. That means that even with a positive test result, Nelly is almost 95% likely to be cancer free!

If, on the other hand, 50% of women in the population have the disease, then given the same test fidelity, there would be an 84.2% chance that Nelly has cancer. That is, the odds of cancer versus good health are essentially reversed.
Table 1: Definition of Symbols for Breast Cancer Test Example

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>p</td>
<td>probability</td>
</tr>
<tr>
<td>C</td>
<td>the existence of cancer</td>
</tr>
<tr>
<td>M⁺</td>
<td>a positive mammography result</td>
</tr>
<tr>
<td>↓C</td>
<td>no cancer</td>
</tr>
<tr>
<td>p(C)</td>
<td>the probability/prevalence of cancer in the population</td>
</tr>
<tr>
<td>p(M⁺</td>
<td>C)</td>
</tr>
<tr>
<td>p(M⁺</td>
<td>↓C)</td>
</tr>
<tr>
<td>p(C</td>
<td>M⁺)</td>
</tr>
</tbody>
</table>

Here’s how the calculation works. Table 1 above defines the symbols used. Nervous Nelly has received a positive mammography. So what we are interested in knowing is the probability that there is actually cancer given a positive mammography result: p(C|M⁺). What we are looking for is the ratio of true positives to all positive results given the information we have. We find this by first taking the probability of a positive mammography given that there is cancer and multiplying it by the prevalence of cancer in the population: p(M⁺|C) p(C). This gives us the probability of a true positive result in this population. On the bottom of the ratio we add together the probability of true positives, p(M⁺|C) p(C), and of false positives, p(M⁺|↓C) p(↓C). Notice that the probability of a false positive result takes into account the known actual likelihood of the disease in the population such that the chance that Nelly received a false positive result is higher.
where fewer people have the disease. This result is highly counterintuitive at first, but this is because of a misunderstanding that will be illuminated below:

\[
p(C \mid M^+) = \frac{p(M^+ \mid C) \cdot p(C)}{p(M^+ \mid C) \cdot p(C) + p(M^+ \mid \neg C) \cdot p(\neg C)}
\]

\[
= \frac{80\% \times 1\%}{(80\% \times 1\%) + (15\% \times 99\%)}
\]

\[
= \frac{0.008}{0.008 + 0.149}
\]

\[
= \frac{0.008}{0.157}
\]

\[
= 5.1\%
\]

Where, on the other hand, the prevalence of the disease is 50%, we get the following:

\[
p(B \mid M^+) = \frac{p(M^+ \mid B) \cdot p(B)}{p(M^+ \mid B) \cdot p(B) + p(M^+ \mid \neg B) \cdot p(\neg B)}
\]

\[
= \frac{80\% \times 50\%}{(80\% \times 50\%) + (15\% \times 50\%)}
\]

\[
= \frac{0.4}{0.4 + 0.075}
\]

\[
= 84.2\%
\]

At first glance it can appear quite counterintuitive that a test that correctly identifies the disease 80% of the time and only yields 15% false positives can ever have such a high likelihood of getting it “wrong”. But of course, the likelihood of a false positive is only very high where the prior probability of the disease is very low. And it is simply essential to take the base rate into account.

Presented in terms of natural frequencies, the answer becomes more intuitively clear: 10 out of 1000 women over age 40 who participate in routine screening have cancer. Given a test with 80% true positives, you can expect about 8 out of those 10 women to get a positive.
mammography (an accurate result). But 80% true positives tells you very little unless you also know how many false positives the test yields and how big the relative pools of cancer ridden and cancer free patients are to which the true positive and false positive percentages apply. That is, we know too that 15% or approximately 149 out of the remaining 990 women without breast cancer will also get a positive mammography. Since 80% of a small number is a small number and 15% of a big number is a big number, looking only at 80% versus 15% is highly misleading.

Seen in this way, it becomes quite clear that only approximately 8 out of 157 (8 + 149) women with a positive test actually have cancer, which immediately appears to be a small percentage (and can quickly calculated as ~5.1%). This makes a good deal of intuitive sense when you consider that the test has more opportunities to make mistakes in testing the larger pool of cancer free women.

Alternatively, in a population where 50% of the women have cancer, we can think of it in this way: Let us say that 500 out of 1000 women over age 40 who participate in routine screening have cancer. Given a test with 80% accuracy, approximately 400 of those 500 women will get a positive mammography. At the same time 15%, or approximately 75 of the remaining 500 women without breast cancer will also get a positive mammography result. In this way, approximately 400 out of the 475 (400 + 75) women with a positive mammography result will actually have cancer. This is ~84.2%.

Now it must be conceded that presenting probabilities in terms of natural frequencies is not strictly correct. That a test has 80% accuracy does not justify one in precisely specifying that 80 out of 100 women will get a positive mammography. 80% accuracy only means that in the long run positive results will approach 80%. Since 100 is not the long run, we have no guarantees that 80 is the right number here. Nonetheless, the natural frequency approach provides a very useful tool for working with probabilities and generating approximations sufficiently accurate for most legal purposes. The relative ease with which information presented in this way can be grasped is also likely to yield useful results for those without statistical expertise.

92 Supra note 90 at 23-24.
More will be said in the chapters that follow about the benefits of presenting evidence in terms of natural frequency. The object of the examples in this chapter, however, is to illustrate what is meant by base rates and why their consideration is so important. Thoroughly misguided results arise from a failure to take base rates (in the previous example, the overall prevalence of the disease) into account.

Of course, sometimes one receives information relating to a topic about which one has no prior knowledge. Under such circumstances, one is left to rely only on, for example, the false positive and true negative rate (namely the sensitivity) and the true positive and false negative rate (namely the specificity) of the test. But here too, it is important to recognize how little you know when all you know is test sensitivity. Knowledge of whether the disease is common or rare in the population is critical to making sense of such results. To know test sensitivity alone is, in the case of Nervous Nelly described above, to know that she could have anywhere from a more than 96% chance that she is cancer free to an 80% chance that she is an unfortunate victim of the disease: two radically different scenarios.

In the legal context, prior probabilities are often supplied by the frequency of certain events or individuals within a population such as the likelihood of infant deaths in hospital or false DNA tests.

### 3.5 Encouraging Good Bayesian Reasoning

Even if it is accepted that Bayesianism makes for a good normative model, it must be conceded that people do make plenty of “errors” according to this normative standard. There is, however, considerable psychological research to suggest various modes of presentation that encourage people to reason in accordance with the dictates of Bayesianism.

In Bayesian Rationality, Mike Oaksford and Nick Chater offer ways to improve human reasoning performance as judged against a Bayesian model. Take, for example, the Wason selection task. The task involves cards with numbers on one side and colours on the other. You are shown cards with the following sides facing up: 3, 8, red and brown. The question posed is

---

as follows: which cards should you turn over to test the rule that if a card has an even number on one side, it has a primary colour on the other. People famously fail at this selection task when it is framed in this way; that is, they assume that both the card with the 8 and the red displayed must be turned over since the rule pertains to even numbers and primary colours. In fact, the card bearing the 8 and the brown card must be turned over since only a card with an even number on one side and a non-primary colour on the other could disprove the rule. That is, the red card need not be turned over since there is no way in which it can offend the rule: either the number on the reverse is even in which case it complies with the rule, or it is odd, in which case it is irrelevant to the rule. What is interesting is that when the task is reframed in terms of testing the rule, “if you’re drinking beer, you’re over 19”, subjects far more accurately identify the information they would need to test the conditional. That is, subjects quickly realize that they would only have to investigate those who are drinking to test the rule. Framed in this way, they do not make the mistake of thinking that all those over 19 must be looked at (which is analogous to thinking what is on the reverse of the red card matters). Oaksford and Chater suggest that reframing the task in this way makes it more straightforward because of the connection between human rationality in familiar versus abstract contexts. No drinking for those under 19 is a rule to which people are accustomed and attuned; the connection between numbers and colours on cards is not.

Familiarity is only one possible explanation for this result. Cosmides and Tooby have argued that it is not familiarity but rather other content effects that explain the superior results when the selection task is framed in terms of underage drinking. Specifically, having looked at a number of equally “familiar” scenarios to frame the task, they found that the drinking scenario was significantly more effective at eliciting the right result. Their preferred explanation is that “social contract” problems elicit superior performance: that is, we have evolved to readily identify cheaters in situations of social exchange. On this explanation, subjects would be expected to fare less well at an equally common, but non-cheating-related, task. An example of this would be testing the rule that “If you’re in Boston, you take the subway”. As it turns out, when faced with four cards with cities on one side and means of transportation on the other

---

(Boston, Arlington, subway, cab) subjects fare far worse in assessing that they only need turn over the “Boston” and “cab” cards to test the rule. In fact, a number of studies show that subjects are not very good at Wason selection tasks even when the rules deal with familiar content drawn from everyday life.\textsuperscript{95}

The idea that human rationality fares better in familiar as opposed to abstract contexts could have some obvious implications for legal decision making. That is, these findings suggest that where possible, information should be framed in a way that is most familiar to legal decision makers. For example, if statistics on whether man-made greenhouse gases contribute to global warming are relevant to an environmental case, as in the case of \textit{EPA v. Massachusetts} mentioned earlier, then they might be presented to a jury in terms of familiar and visible substances.

Alternatively, if Cosmides and Tooby’s explanation is preferred, then deontic formulations, that is, formulations relating to rules and cheating, would be most successful in eliciting accurate results. On this view, environmental statistics about greenhouse gases should not only be presented in terms of more familiar and visible substances but also in terms of rules about such substances. It would seem to follow from this view that human reasoning is at its best in the legal context: the law is all about rules and rule breaking.

A better descriptive understanding of the ways in which people actually go about making decisions can inform choices about how to present evidence so that it is used in compliance with Bayesian standards. In “Simple Heuristics and Legal Evidence”\textsuperscript{96}, Alvin Goldman suggests that while Bayesianism is attractive as a normative model, a plausible descriptive account of human rationality involves the use of heuristics that are “fast and frugal”. As an example, he suggests that we use a recognition heuristic to make us relatively good at recognizing which of two cities is bigger based on whether or not we recognize the city’s name.


While some simple heuristics are relatively successful in yielding results in line with good Bayesian reasoning, others seem less so. Eyewitness testimony, criminal confessions, police testimony and expert testimony are just some of the examples that Goldman provides. This suggests that heuristics show more promise as a descriptive, rather than necessarily optimal, model of human reasoning.

Kahneman and Tversky suggest that people rely on a limited number of heuristic principles that reduce the complex tasks of assessing probabilities and predicting values. In general, this approach is quite helpful, though it can sometimes lead to severe and systematic errors.

For example, when asked if X is an engineer or a lawyer and given descriptions, subjects ignore whether they are told that X was drawn from a pool of 70 engineers and 30 lawyers, or vice versa. Subjects do however use prior probabilities correctly when given no other information. Kahneman and Tversky note that prior probabilities are effectively ignored when a description of X is introduced, even when this description is totally uninformative (such as a description of X as having a wife named Bunny or wearing black pants). When no specific evidence is given, prior probabilities are properly utilized; when worthless evidence is given, prior probabilities are ignored.

Kahneman and Tversky also report that subjects are insensitive to sample size in answering the question of which of two hospitals would be more likely to record that 60 percent or more of the babies born in a given year were boys, assuming that on average, about 50 percent of all babies born are boys. That is, the smaller the sample size, the greater the chance that proportions will differ significantly from the mean. If only 5 babies are born in a small rural town in one year, then the birth of 3 boys and 2 girls results in a 60% boy birth rate in that hospital. In fact, given a total of 5 births, it would be impossible for this small hospital’s birth rate to get any closer than this to the overall average. If, on the other hand, 4000 babies are born in a large city’s hospital, the chances are far greater that the number of boys born would more closely approximate 50%. That is, given the 50% overall average, it is far more likely that the number of boy babies born

would not be 24000 (which is 60% of 4000), but would rather be within a couple hundred of 2000.

In terms of misconceiving of chance, subjects are apt to view chance as a self-correcting process in which deviation in one direction induces a deviation in the opposite direction to restore the equilibrium. This gives rise to the gambler’s fallacy that a string of losses is bound to be followed by a win.

Where one group is asked to estimate the product of $8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1$ and the other asked to estimate the product of $1 \times 2 \times 3 \times 4 \times 5 \times 6 \times 7 \times 8$, the median estimate for the ascending sequence is 512, while for the descending sequence it is 2,250. The correct answer is 40,320. These types of distortions are referred to as “anchoring and adjustment” errors.

Heuristics of representativeness, availability of instances, and adjustment from an anchor are highly economical and usually effective heuristics, but they lead to systematic and predictable errors. Kahneman and Tversky suggest that a greater awareness of them could improve judgments and decisions in situations of uncertainty.

For example, subjects are apt to prefer causal explanations over diagnostic considerations where a given scenario supports both interpretations. This is illustrated by the following problem and response. The problem is described in this way: let $A$ be the event that before the end of next year, Peter will have installed a burglar alarm system in his home. Let $B$ denote the event that Peter’s home will be burglarized before the end of next year. Let $\neg A$ and $\neg B$ denote the negations of $A$ and $B$. Subjects are then asked the following questions: “Which of the two conditional probabilities $p(A|B)$ or $p(A|\neg B)$, is higher?” and “which of the two conditional probabilities, $p(B|A)$ or $p(B|\neg A)$, is higher?”

The first question asks whether the probability that an alarm system is installed in a certain year is higher if there is a burglary that year or if there is not. The second asks whether the probability of a burglary in a certain year is higher given that an alarm system either is installed or is not installed in that same year. The large majority of subjects stated that $p(A|B) > p(A|\neg B)$ and that $p(B|A) < p(B|\neg A)$. That is, the large majority state that the probability of alarm installation is higher if there is also a burglary and the probability of burglary is higher if there is no alarm installed that same year. These results suggest that subjects misread conditional
probabilities so as to place the condition temporally prior to the target event. That is, they read $p(A|B)$ as the probability of alarm installation given that there has been a burglary when in fact, using the definitions provided, $p(A|B)$ also covers the situation in which the alarm is installed before the burglary occurs. It is only if you read the conditional probability as placing the burglary first in time that you would have reason to suppose that $p(A|B) > p(A|\sim B)$. That is, you may think that the burglary would cause alarm installation in response and so make alarm installation more likely than in its absence. But a conditional probability need not tell a causal story. That is, the probability of alarm installation in a given year may also be greater given that there is no burglary that year. The absence of the burglary could be the effect rather than the cause of the alarm installation in that year. Kahneman and Tversky explain subjects’ results in terms of a preference for causal explanations (the burglary caused the alarm installation) over diagnostic ones (the absence of a burglary was an indication that the alarm had been installed).

In fact, there is nothing in the information provided to suggest that one probability is higher than the other. Instead, a correct probabilistic analysis of this problem would require subjects to bring their prior probabilities to bear: it is more likely that a home will not be burglarized in any given year than that it will be.

When it comes to the evidential impact of base rates, certain results suggest that subjects sometimes rely on representativeness, ignoring base rates in the process. In a test in which subjects are presented with a description of a professional, and are asked to predict her profession, posterior probability judgments are determined primarily by the degree to which the description is representative of the relevant stereotype. That is, even in a librarian-dense population, if the individual is described as aggressive and competitive, subjects say she is most likely a lawyer.

Kahneman and Tversky interpret this as base rate frequency neglect. While it may well constitute such neglect, this is not necessarily the case. That is, if, for example, there is no such thing as an aggressive and competitive librarian (for the purpose of this example, we’ll assume this proposition with a probability of 1), it would not matter that the population is librarian-dense; the individual described would still have to be a lawyer.
Another interesting set of results relates to information acquisition and overconfidence. That is, in circumstances in which predictive accuracy reaches a ceiling at some early point in the information-gathering process, confidence in one’s decisions continues to climb steadily as more information is obtained. In this way, as one reaches the end of the information-gathering process, there is a high degree of overconfidence in subjects’ judgment assessments.

Kahneman and Tversky use the character of Joseph Kidd from the book *Life in Progress* as an example. Subjects, unfamiliar with Kidd’s character in the book, are given the following description: “Joseph Kidd (a pseudonym) is a 29 year old man. He is white, unmarried, and a veteran of World War II. He is a college graduate, and works as a business assistant in a floral decorating studio.” At the second stage of the study, subjects are provided with additional information about Kidd’s childhood. At stage three, they learn about Kidd’s high school and college years and stage four covers his army service and later activities up to the age of 29.

Subjects, in this case comprised of clinical psychologists, undergraduates and psychology graduate students, proceed to answer 25 questions about Kidd’s character and are also asked how confident they are in the answers. As subjects acquire more information about Kidd, their confidence ratings soar, but their accuracy, as measured against the character as described in White’s book, does not.

Results relating to overconfidence in decisions with increased information are of course highly relevant to the legal decision-making context. Judges and juries at the end of month-long trials are likely to feel far more secure in their decisions than they would after hearing only two-days-worth of evidence. It is in large part because of this that we hold lengthy trials. But it is questionable whether decision accuracy actually increases as a result.

Results such as those reported above are often used as demonstrations of rampant human irrationality. I would suggest that the better interpretation is that they actually *presuppose* Bayesian rationality as a model of optimal performance against which human judgments can be compared. That is, while the heuristics seem error-prone in many contexts, it makes most sense to interpret them as such only against a background of largely successful human rationality.\(^98\)

\(^98\) Successful rationality is here judged against a normative standard of Bayesian rationality.
Studies of human decision making and irrationality can, however, provide insight into common biases. This kind of knowledge can be used to inform the presentation of evidence in ways that encourage the alignment of ordinary decision making and correct Bayesian results.

3.6 A Challenge to Bayesianism as a Normative Model

As a normative model, Bayesianism does face challenges in various forms. There is no need to attempt to catalogue them here. I will, however, say a few words about the so-called problem of priors, since this is a non-technical problem that is readily identified upon first acquaintance with Bayesianism. Briefly stated, the problem of priors is that while Bayesian calculus provides a rigorous and objective model for generating posterior probabilities given existing prior probabilities and new evidence, the prior probabilities themselves are relatively unconstrained. So-called subjective and objective Bayesians differ on the extent to which prior probabilities may be constrained – subjectivists are defined by their emphasis on the relative lack of constraints on prior probabilities while objective Bayesians introduce varying degrees of rational constraint on priors. For objective Bayesians, rationally admissible prior probabilities are often constrained in various ways, although with all manner of constraints proposed, none uniquely determine the prior in all cases. In this way, where the available evidence is taken as a given, disagreements between good Bayesians must come down to differences in prior beliefs, since inferences from the evidence are fully constrained. But if this is the case, then there is no way to mediate between disparate views since they differ based only on completely subjective prior probabilities. This can be viewed as a significant weakness for a normative model of rationality, one of the goals of which may be to mediate between disparate views. That is, norms in reasoning allow us to debate those with opposing views in the hopes of uncovering a flaw in our opponent’s chain of reasoning. If, however, it were always possible to end a debate with the discovery of disparate priors, the exercise would be quite unsatisfactory (“Oh, you started out convinced that the moon was made of green cheese. Well, in that case, I have nothing more to say”).

It has also been demonstrated that if a community of Bayesians contains even two agents who modestly differ on a hypothesis, where “modestly differing” is defined as having a prior that is less than 99.9% of the other agent’s prior, then convergence to agreement can occur only if the whole community comes to agree that the new evidence is strong enough to confirm the hypothesis in question and refute its competitors.100

The Bayesian is not, however, without a reply to this; while initial prior probabilities may differ considerably (although less so for the objective than the subjective Bayesian), with each new experience and new piece of evidence, the difference in starting points becomes less and less salient. There is also a sense in which one good Bayesian encountering another with a different belief should adjust her belief accordingly on the assumption that her fellow Bayesian’s beliefs differ based on exposure to different evidence: evidence that had she encountered it herself, would have led her to similarly adjust her beliefs. The corollary to this is that those with the same priors cannot agree to disagree. That is, if two people have the same priors and their posteriors for a given event A are common knowledge,101 their posteriors must be the same. This is so even if their posterior probability assessments are based on quite different pieces of information.102

In addition, it can be argued that the subjectivity of Bayesianism adds to its plausibility as a realistically fallible and incomplete model of our understanding of the world.103 As Colin Howson suggests in “A Logic of Induction”,104 where he considers the so-called problem of objective priors, the subjective element may be a reflection of its connection to the realities of the world:

100 For the full proof see James Hawthorne, “On the Nature of Bayesian Convergence” (1994) 1 Philosophy of Science Association 241.
101 There is “common knowledge” of p in a group of agents G when all the agents in G know p, they all know that they know p, they all know that they all know that they know p, and so on ad infinitum.
103 Supra note 90.
…to assume that the theory is in deficit on the objectively grounded true methodological judgments that can be made begs the question that it is possible to make objectively correct judgments where that theory fails to do so. This is a claim that is unsubstantiated and which there seems reason to doubt. We have seen that the claim of superiority on this score made by a rival account, Neyman-Pearson theory, is untenable: that theory’s apparently greater power generates unsound fallacious inferences.  

Howson proceeds to explain that to assume that there is actually an objective yes or no answer to every question is to assume a completeness that Gödel, Church, Tarski and others have shown does not exist. Instead, Howson posits that the Bayesian theory may represent a necessary limit on theoretical completeness beyond which the theory becomes unsound (i.e. contains untrue premises). Admittedly, this approach does little to allay concerns about vast differences in subjective priors. But at the same time, with a few notable exceptions, there are compelling reasons to think that our perspectives on the world are largely shared. 

It should be clear by now that Bayesianism enjoys at least reasonable plausibility as a normative model for rationality. And as De Finetti himself said, “…people noticing difficulties in applying Bayes’ theorem remarked, ‘We see that it is not secure to build on sand. Take away the sand, we shall build on the void.’”

I am inclined to think that the “problem of priors” is consistent with viewing Bayesianism epistemology as akin to democracy: the worst there is except all the others that have been tried. Of course, if your prior probability on this differs significantly from mine, you may not be convinced. It is easy enough to identify deficiencies with a model of rational knowledge that relies on prior beliefs; on the other hand, we all must start somewhere. So long as there is strong

105 Ibid. at 283.

evidence with which to update one’s priors, convergence on similar beliefs will occur. And in the absence of strong new evidence, it is appropriate that we are unable to convince each other to adopt one belief over another; in these instances, all our beliefs are relatively thin and unsubstantiated. These are the circumstances under which we would be advised to say that “I choose to believe that X” rather than “I believe (or know) that X”.

Ultimately, the question is not whether Bayesianism has its flaws as a normative model and may be susceptible to error, but whether it fares well relative to the other approaches on offer.

3.7 An Application to Medicine: Bayesian Analysis of Clinical Data

Bayesian methods can also be adopted in the context of clinical medical approaches. They can be used to supplement rather than to replace traditional analyses in assessing health care interventions. In general, Bayesian analyses make wider use of probability distributions to convey prior opinion about proportions, event rates, and other unknown quantities. ¹⁰⁷

A Bayesian analysis can take into account the many sources of evidence already available that may be inadequately addressed by the traditional single “alternative hypothesis” invoked in standard analyses. Whereas standard statistical methods are designed to summarize evidence from single studies or pool evidence from similar studies, Bayesian analyses are better-equipped to reflect multiple sources of evidence. ¹⁰⁸

For example, suppose you’re interested in whether, and by how much, a particular treatment, such as chemotherapy, increases the survival rate of patients. This is called the “treatment effect”.

A standard statistical analysis of chemotherapy trial data would yield the following results. It would yield a p-value for the null hypothesis that the treatment effect is zero. The p-value is the probability of obtaining a result at least as extreme as the one actually observed, assuming that

---

¹⁰⁷ D J Spiegelhalter, K R Abrams & J P Myles, Bayesian Approaches to Clinical Trials and Health-Care Evaluation (Mississauga: John Wiley & Sons Ltd, 2004).

¹⁰⁸ Ibid.
the null hypothesis is true. That is, assuming the result observed was that the mean patient survival rate increased by 7 years, the p-value would represent the probability that this result would be observed were the chemotherapy treatment to have no effect. The statistical significance of the result varies inversely with the p-value: that is, the lower the p-value, the less likely one is to observe the result on the assumption of the null hypothesis. P-values are often 0.05 or 0.01, corresponding to a 5% chance or 1% probability of an outcome that extreme, given the null hypothesis. The results of a test are either statistically significant or they are not.

A standard statistical analysis of the data would also yield a point estimate, which is a single value used to estimate the population parameter. It involves the use of sample data to calculate a “best guess” value for an unknown population parameter. That is, in the case of chemotherapy treatment the point estimate would tell you how many of the people involved in the study improved as a result of the treatment, but you would not know from this alone whether that estimate was representative of the population as a whole. You may have just randomly selected a group with an on average higher improvement rate than that of the general cancer patient population.

Standard statistical analyses also typically yield a confidence interval, which is a particular kind of interval estimate of a population parameter. The confidence interval is the point interval plus or minus a margin of error. The confidence interval serves to indicate the reliability of an estimate, namely how likely the interval is to contain the parameter. One speaks, for example of an 80% confidence level in a confidence interval defined by confidence limits. The higher the confidence level, the wider the confidence interval will be. All other things being equal, a smaller confidence interval is more desirable than a larger one because a smaller interval means the population parameter can be estimated more accurately.

The information generated by a typical statistical analysis gives us some sense about what the chemotherapy trial data reveals in relation to the treatment efficacy in terms of the defined goal of increasing survival rate.

The Bayesian analysis adds to the information generated by the standard statistical analysis and summarized by the p-value, point estimate and confidence interval. That is, the Bayesian analysis tells us something about how the trial should change our opinion of the treatment effect based on three additional considerations: (i) the reasonable prior opinion about the plausibility
of different values for the treatment effect, referred to as the prior distribution (which excludes the evidence from the focal trial); (ii) the support for different values based solely on the focal trial data, referred to as the likelihood; and (iii) a final opinion about the treatment effect, referred to as the posterior distribution.

Some advantages of Bayesian analysis in health care that add to the standard statistical analysis include the following: (i) it allows all evidence considered thus far (i.e. up to the point of the particular study in question) to be taken into account; (ii) it requires explicitness about the external evidence and judgment that is brought to bear on the results of a particular study; and (iii) it allows for the pooling of evidence among studies. 109

The Bayesian approach also provides a clear view about the kinds of conclusions that scientists in general, medical scientists in particular, may draw:

Scientific methodologies have to take a view about the kinds of conclusion scientists draw, and should provide appropriate mechanisms for those inferences. The bayesian method meets these conditions by characterizing a scientific conclusion about a hypothesis as a statement of its probability, and by providing Bayes’s theorem as the mechanism for calculating that probability. 110

In this way, a Bayesian model can lend additional clarity to medical conclusions. In particular, it helps to make explicit the limitations of the conclusions drawn in any particular case by expressing conclusions probabilistically. While the details of Bayesian statistical analyses in medicine are complex, the potential value of Bayesianism to medical analyses is apparent: given our necessary lack of certainty, all accurate statements about the world should involve some statement of probability. A Bayesian approach makes this uncertainty explicit, thereby reducing the chance that uncertainty will be overlooked.

109 Ibid.

3.8 Two Applications to the Law: Identification Evidence and Base Rates

In the legal context, prior probabilities are supplied by the frequency of certain events or individuals within a population. These frequencies are referred to as “background” or “base” rates.

Whether it’s DNA evidence or environmental evidence, explicitly statistical evidence arises regularly in the law. The law, however, lacks a systematic and workable approach to integrating this evidence in decision makers’ reasoning process. For this reason, the law often shies away from such evidence, losing out on all it has to offer.

Identification evidence provides an example. “Identification evidence” is any evidence that speaks to the factual question of the identity of the culprit in a legal case. Identity is not the central question in many cases. For example, where the charge is sexual assault, more often the defendant’s position is not that he “didn’t do it”, but that all the activity that took place was consensual. In other criminal cases, however, the main focus of the trial is to answer the question of whether the defendant is the “right guy”. That is, there is no dispute that the crime took place – for example, it is agreed by all that someone shot the victim and that a crime was committed – the question is by whom?

Part of the courts’ hesitance to employ explicit statistical evidence stems from its frequent misuse in the legal deliberation context. One of my objectives is to provide a workable approach to statistical evidence that would reduce confusion and misuse.

One common confusion about the role of statistics in law is the belief that their use often inappropriately constraints decision makers. In fact, inferences of identity based on statistics will generally be weaker than expert judgments expressed in the usual way; few, if any, evidentiary traces can be shown by statistics to be unique. In the face of evidence that, for example, fingerprints with particular characteristics occur approximately once in every thousand people, this suggests only that a defendant with this print is a thousand times more likely than someone chosen at random to have left it. But the question of the defendant’s likelihood of guilt remains entirely open. In a population of 30 million, 30,000 people would possess such “one-in-a-thousand” characteristics.
In the legal context, evidence must be looked at as a whole and guilt determined by accumulating probabilities. Even when statistics are not involved, this cumulative perspective controls the significance of any piece of evidence. The same is true where explicitly statistical evidence is invoked. Where there is evidence of a match between a print at the scene of the crime and the accused’s finger, in addition to the consideration of non-uniqueness just mentioned, there are also other factors to be considered. The accused may have been present at the scene of the crime without having committed it. Or, the accused may have been framed, the real culprit having planted an object with the accused’s fingerprints at the scene of the crime.

In “A Bayesian Approach to Identification Evidence”¹¹¹, the authors, Michael Finkelstein and Fairley, point out that statistical evidence can be used modestly to improve verdict accuracy. No piece of statistical evidence can ever map onto the probability of guilt; it is the jury’s role to consider whether they are satisfied beyond a reasonable doubt that it was the accused, and not someone else, who committed the crime. That the chances are very good that the accused committed the crime will not satisfy the requisite standard of proof.

But courts regularly shy away from statistical evidence and sometimes even explicitly reject it. In one case mentioned by Finkelstein and Fairley, the court rather shockingly rejected statistical evidence on the basis that probability can only be used with respect to future events and not past ones. While there may be legitimate reasons for rejecting statistical evidence, this is not among them: there is no relevant difference between the future and the past when it comes to mathematical probability. The principles of probability apply so long as uncertainty is involved. Finkelstein and Fairley ultimately conclude that a Bayesian approach that begins with a subjective prior for the non-average case would probably improve the performance of judges and juries. Mathematics correctly used should lead to a fairer evaluation of identification evidence.

And while there are those who may worry that statistical evidence would lead decision makers to convict with false certainty, Finkelstein and Fairley suggest that statistical evidence is actually more likely to lead to the opposite result since it is never conclusive in the way qualitative expert

opinions sometimes are. That is, statistical evidence always wears the risk of error on its sleeve while qualitative evaluations often do not.

The story of Dow Corning, and the proliferation of breast implant lawsuits in the U.S., provides a useful lesson about the dangers inherent in ignoring some basic statistical principles. In order to be successful in a product-liability action, a plaintiff must demonstrate that the injury-causing product was defective, that the defect existed at the time the product left the control of the defendant, and that the defect was the proximate cause of the plaintiff's injury. The stream of lawsuits against the breast implant manufacturers, of which Dow Corning was the largest, began after David Kessler, then head of the U.S. Food and Drug Administration (“FDA”), banned silicone-gel filled breast implants from the market in 1992. In 1991, a deadline had expired for implant manufacturers to prove the safety of their product to the FDA. No manufacturer had offered any convincing proof on the matter.

In 1993 the parties tentatively agreed to settle the class-action products liability lawsuit for $4.75 billion. The prospective settlement fell through however after 440,000 women registered for the settlement. Dow Corning filed for bankruptcy.

It was alleged that breast implants were responsible for a number of disorders known as connective tissue disease. But since connective tissue disease can develop among women with or without implants, to evaluate the claim that the implants were responsible for the disorders suffered by the women would require knowledge of the general rate of occurrence of connective tissue disease in the female population (the base rate). It is only in light of the known base rate for the disorder that it could be demonstrated that the risk of the disease was significantly higher for women with implants than without. To discover the base rate incidence of connective tissue disease in the U.S. female population at large would have required epidemiological studies, that is, scientific surveys of the incidence of the disease among different groups. No such base-rate evidence was ever collected. After it was determined at a preliminary hearing that causation was an issue of fact that would need to proceed to a full trial, Dow Corning decided to settle. They likely suspected that juries would sympathize with ill plaintiffs rather than a big pharmaceutical company. Had they thought that base rates would have been properly considered at trial, the story may have unfolded differently.
A few years later in 1999, an independent panel of 13 scientists convened by the Institute of Medicine at the request of Congress concluded that silicone breast implants do not cause any major diseases. The report, which was more than 400 pages long identified that there was a safety issue with implants relating to their tendency to rupture or deflate and to lead to infections or hardening or scarring of the breast tissue. The report was equally clear however that there was no basis for belief that the implants caused rheumatoid arthritis, lupus, or any of the other systemic diseases that formed the basis of the plaintiffs’ claim. 112

Without confidence that a trial would properly consider base rates – in this case the prevalence of connective tissue disease among women in the general population – it would appear that both the plaintiffs and the defendant viewed the co-existence of a breast implant and connective tissue disease as persuasive to settlement.

The acceptance of Bayesianism as a normative model for considering factual legal claims along with the provision of basic education to avoid rudimentary conceptual errors would assist in preventing such injustices. That is, the true hazard lies not with the geeks bearing formulas but with the failure to understand the message they bring.

If it is accepted that mathematics correctly used can improve the quality of decision making in a number of legal contexts, then the focus turns to the best ways in which to present such evidence for optimal use. That is the topic of the discussion to follow.

Chapter 4
The Probability “Paradoxes”: Need Legal Intuition and “Good” Bayesian Results Clash?

4 The Probability “Paradoxes”

In this chapter, well-known paradoxes designed to illustrate the irrelevance of base rates in the legal context are discussed. Critics of Bayesianism generally commit one of three errors: they claim the irrelevance of statistical evidence at the same time that they presuppose it, they fail to separate the legitimate invocation of non-truth-related trial process values from errors in factual determination, or they misunderstand the correct application of statistical principles. The chapter proceeds as follows:

- Description of the “paradoxes”
- Common responses to the “paradoxes”
- Koehler and Shaviro’s resolution to the impasse
- An alternative resolution based on corroborating evidence
- Simple heuristics as an alternative response
- A possible place for base rates

The purpose of this chapter is to dispel the notion that a Bayesian approach, which requires the consideration of base rates in forming updated beliefs in light of new evidence, is somehow misplaced in the context of legal decision making. The chapter treats some well-known “paradoxes” from the literature on probability and the law, and aims to demonstrate the failure of these examples to show Bayesianism ill-suited to legal evidence.

The logic of the paradoxes is as follows: since the consideration of base rates in the legal context gives rise to dramatically counterintuitive results, Bayesianism must be inapplicable to that domain. This argument is vulnerable on two fronts. One may either deny the premise that counterintuitive results arise, or the inference that because the results are counterintuitive, they’re wrong. Both options are discussed. I suggest that solutions based on rejecting the basic premise that probability and intuition regularly conflict possess certain benefits particular to the realm of law. It is a fundamental tenet of our legal system that justice must not only be done, but also be
seen to be done. If Bayesian results and intuition can be reconciled on most occasions, then the Bayesian result can both be, and appear to be, the right result. In this way, it becomes possible for probability theory to play a normative role in correcting intuition every now and then.

I offer a unique solution to one of the supposed paradoxes that allows intuition and Bayesian probability to align. I also consider an alternative approach to improving the accuracy of legal decision making, one that is based upon looking for expedient approaches to particularly error-prone intuitions. This approach is, however, susceptible to multiple practical problems and also fails to provide any normative framework. This adds further support for the Bayesian model on offer. I proceed to consider the presentation of base rates in a form that encourages the alignment of intuition and probability. To demonstrate that disagreement between probability and ordinary legal intuition involves only penumbral cases is to allow the two to exist in a beneficial dialectic. If probabilistic principles can be more easily understood, then they can shape intuitions without fear that probability will lead us astray.

It is only when the lessons learned from probability theory can be absorbed intuitively that base-rate evidence may successfully be integrated by decision makers with other non-numeric evidence to arrive at confidence-inspiring conclusions.

### 4.1 Description of the “Paradoxes”

There are a number of purported paradoxes that repeatedly show up in one guise or another in the literature on base rates and legal evidence. The subject matter ranges from prisoners to bunny rabbits, but the ultimate purpose remains clear: namely, to demonstrate that the use of base rates in the context of legal decision making yields unacceptably counterintuitive results. In this section, three such stock hypotheticals and the paradoxes they are said to generate are described:

- The gatecrasher “paradox”
- the green cab or blue cab “paradox”; and

---

• the Todhunter/lottery “paradox”

It is worth noting at the outset that a failure to keep verdict accuracy and verdict acceptability distinct, that is, a failure to appreciate that truth is not the law’s only policy objective, is at least partly responsible for the generation of these supposed paradoxes.

4.1.1 Gatecrasher “Paradox”

The first hypothetical is most commonly known as the “gatecrasher paradox”. In this scenario, the issue is not whether base-rate evidence should be contemplated as part of an assessment of other direct evidence, but rather whether base-rate evidence alone is sufficient to find that the defendant committed the offence in question. The scenario is as follows: 100 football fans attend a game. 90 people fail to pay. The proprietor of the stadium sues a randomly chosen individual, Mr. V.G. Fan, for the price of admission. Should we conclude that Mr. V.G. Fan did not pay on the basis that there is a 90% probability that any given attendee, including Mr. V.G. Fan, was a gatecrasher, and given that the burden of proof in civil matters is a balance of probabilities or a probability of more than 50%?114

You would be hard pressed to find any legal case in which someone successfully sued on the basis of base-rate evidence alone. To begin with, there is only what the law calls “circumstantial evidence” in this case; there is no direct evidence relating to the individual defendant. The prior probability of a particular individual committing a crime may not be unknown – if for example, the person is a prior offender – but the law chooses not to take prior offences into account. For those familiar with the Anglo-American legal tradition, and quite likely even for those who are not, there is a strong intuition that it would be inappropriate to conclude on this evidence alone that, on a balance of probabilities, Mr. V.G. Fan did not pay for his ticket. That is, if the only evidence that the stadium proprietor has is evidence that there is a 90% likelihood that any given individual failed to pay, he hasn’t got anything on Mr. V.G. Fan. Certainly, when we think of ourselves in Mr. Fan’s situation, we would feel unjustly treated were we to be convicted of wrongdoing on the basis of this statistical evidence alone. That is, a single piece of statistical evidence relating only to a group of which I’m a member – e.g. the group of women, Ph.D.

candidates, or those with hazel eyes, and so on – can’t alone be used to implicate me because to do so would be to fail to treat me as an individual. It is true that direct evidence, such as an eyewitness’s account that she “saw me” at the scene of the crime may also invoke group membership; for example, such evidence may be presented against a backdrop that the witness in question is more accurate in her identification of women than of men. However, it is not the invocation of a group in and of itself, but rather the use of group membership by itself that offends our sense of individuality. It is in this way that the gatecrasher hypothetical is intended to draw out the intuition that the failure to provide any evidence pertaining to the individual in particular is a critical failing from the perspective of legal evidence.

4.1.2 Green Cab or Blue Cab “Paradox”

The next hypothetical scenario is described by Laurence H. Tribe in “Trial by Mathematics”\textsuperscript{115} in one guise, and was later recast by Kahneman and Tversky for use in an experiment on decision making under uncertainty. Both versions of the hypothetical are said to have been inspired by an actual 1945 case by the name of Smith v. Rapid Transit. In that case, a plaintiff was injured in an accident caused by a negligent bus driver. It is worth noting that in this case the plaintiff sued Rapid Transit on the grounds that because Rapid Transit chartered most of the buses on the street where her accident had occurred, it was probably Rapid Transit that was responsible. The court did not accept that this base-rate evidence was sufficient to find that it was the defendant’s bus that was involved:

The most that can be said of the evidence in the instant case is that the mathematical chances somewhat favour the proposition that a bus of the defendant caused the accident. This was not enough.\textsuperscript{116}

Note that the court did not venture an opinion as to whether the mathematical chances were relevant to the ultimate determination, but simply found that they were “not enough” to find that the defendant company’s bus was involved. In this sense, while the case is cited as the


\textsuperscript{116} Smith v Rapid Transit Inc (1945) 58 N E 2d 754 at 754.
inspiration for the blue bus/red bus and blue cab/green cab hypotheticals used by Tribe and by Kahneman and Tversky, it is actually more closely aligned to the gatecrasher hypothetical from the perspective of the role that base-rate evidence plays within it.

I note though that there are cases in which the law has considered evidence of “market share” to find in favour of the plaintiff(s). For example, in *Sindell v. Abbott Laboratories*, 26 Cal. 3d 588 (1980), the plaintiff was a young woman who had developed cancer as a result of her mother’s use of the drug diethylstilbestrol (DES) during pregnancy. There were a number of manufacturers of DES. While the plaintiff had succeeded in proving that there was a breach of the duty of care by the manufacturers of the drug and that she had been harmed by this breach, she could not prove which company had actually manufactured the particular batch of pills ingested by her mother.

In a 4-3 decision, the California Supreme Court decided to impose liability in spite of the plaintiff’s inability to prove causation in relation to any particular defendant on a balance of probabilities. Instead, the court imposed liability based on market share, holding that such liability could be imposed under conditions where,

1. all defendants named in the suit are *potential* tortfeasors (that is, they did produce the harmful product at issue);
2. the product involved is fungible;
3. the plaintiff cannot identify *which defendant* produced the fungible product which harmed *her* in particular, through no fault of her own;
4. a *substantial share* of the manufacturers who produced the product during the relevant time period are named as defendants in the action; and
5. the plaintiff can prove actual damages.

The existence of these conditions were held to give rise to a presumption in favour of the plaintiff open to rebuttal by the defendant (for example, by presenting evidence that the company did not manufacture the product at all during the time period at issue).\textsuperscript{117}

\textsuperscript{117} *Sindell v Abbott Laboratories* (1980) 26 Cal 3d 588.
The essential difference between *Smith v. Rapid Transit* and *Sindell* is that all of the defendants in Sindell were shown to have fallen below the requisite standard of care. So while manufacturers that had not actually caused the harm to Ms. Sindell were found liable, no defendant was held liable who had not engaged in blameworthy conduct.

In contrast to the *Smith v. Rapid Transit* case and to *Sindell*, which feature only statistical evidence on the factual issue of causation, Kahneman and Tversky’s version of the cab problem involves the contemplation of direct evidence *against the backdrop* of a base rate. That is, the issue becomes to what use, if any, the base-rate evidence should be put, in assessing the direct evidence that is also presented:

A cab was involved in a hit-and-run accident at night. Two cab companies, the Green and Blue, operate in the city. You are given the following data: (i) 85% of the cabs in the city are Green and 15% are Blue. (ii) A witness identified the cab as a Blue Cab. The Court tested his ability to identify cabs under the appropriate visibility conditions. When presented with a sample of cabs (half of which were Blue and half of which were Green) the witness made correct identifications in 80% of the cases and erred in 20% of the cases. Question: what is the probability that the cab involved in the accident was Blue rather than Green?*

Applying Bayes’ theorem, and using the relative frequency of blue and green cabs as prior probabilities to be accounted for, the 80% accurate witness would be expected to say “blue” (when it was indeed blue) in approximately 12 out of 100, or 80% of 15, cases, and would also identify the cab as blue (when it was in fact green) in 17 out of 100, or 20% of 85. So the probability that the cab was in fact blue given that the witness proclaimed this to be the case is only approximately 12/29, that is, approximately 41%. As explained in a previous chapter, these numbers are only approximate since 80% accuracy means 80% accuracy *in the long run* such that we cannot know precisely what percentage of 100 cabs will be correctly or incorrectly

\[\text{Supra note 97.}\]
identified by the witness. For our purposes, however, treating “100” as “the long run” is the best that we can do. This lack of strict accuracy simply reflects the lack of complete certainty that is appropriate to any contemplation of real world events.

The purported paradox to which this scenario gives rise is as follows: intuition tells us that an 80% reliable witness who says that the bus was blue provides good grounds to believe that her report that the blue bus “did it” is accurate; probability theory tells us otherwise. That is, taking the base rates into account, we should actually be only about 41% confident in the eyewitness’s report. Given a civil burden of proof of the preponderance of evidence (i.e. more than 50% in the plaintiff’s favour), this difference would reverse the result. That is, if you believed that the witness’s report was 80% reliable, in the absence of other evidence, you would find in the plaintiff’s favour; if you believed that the chance that a blue bus was responsible given the witness’s report was only 41%, you would not. Whether or not this result is properly construed as “paradoxical” at all will be discussed below.

4.1.3 Todhunter/Lottery “Paradox”

The third and related hypothetical is called the Todhunter problem, named after the man who devised it. This hypothetical shows the putative dilemma in its purest form. The scenario is described as follows: “suppose that a witness who is correct 99.9% of the time claims that the winning number in a lottery of 10,000 tickets was 297. Intuitively, it would seem that we can be 99.9% certain that 297 really was the winning number. The number of tickets in the lottery appears at first blush to be irrelevant.”119 (The contours of this problem may appear familiar to you as they mirror those described in the medical context in the example of nervous Nellie).

The paradox generated by this problem is this: taking into account the prior probability of 1 in 10,000 that any one ticket bears the winning number, even in the face of the testimony of a highly reliable witness, Bayes’ theorem generates a posterior probability of only 9% that the ticket identified by this unusually reliable witness is in fact the winning ticket.

Where we designate the fact that the winning ticket is actually ticket number 297 as 297, and the witness’ statement that the winning ticket is number 297 as “297”, according to L.J. Cohen and others, the application of Bayes’ theorem yields a very small chance that the winning lottery ticket was 297 given the witness’ say so:\(^{120}\):

\[
p(297 \mid "297") = \frac{p("297" \mid 297) p(297)}{p("297" \mid 297) p(297) + p("297" \mid \sim297) p(\sim297)} = \frac{0.999 \times 0.0001}{0.999 \times 0.0001 + 0.001 \times 0.9999} = 0.09
\]

The purported paradox is generated by the contrast between the intuition that a 99.9% reliable witness gives rise to high confidence in the conclusion that 297 was the winning ticket, and the solution that it is claimed is generated by Bayesian probability. In taking the prior probability into account, according to some, probability theory generates a posterior probability of merely 9%.

What all the above hypotheticals have in common is the use of a base rate (or prior probability) acting either alone or in concert with another piece of evidence to generate what appears to be a counterintuitive result. The gatecrasher problem is distinctive in the sense that it is intended to draw out the intuition that base-rate evidence alone is insufficient to find that the defendant committed the offence in question. By contrast, the blue bus/green cab hypothetical and the Todhunter paradox both feature eyewitness testimony delivered against the backdrop of a prior probability.

### 4.2 Common Responses to the “Paradoxes”

All of the scenarios described above exist in a context of private law. That is, they concern the legal response to conflicts between two or more individuals (as opposed to between the state and an individual, as is the case with criminal law). Inherent in the structure of private law is the

---

\(^{120}\) Ibid.
Aristotelian idea that private law, otherwise known as corrective or restorative justice, has three essential features: (i) it is independent of social rank or moral character (“whether a worthy person takes something from an unworthy person makes no difference”); (ii) it regards the transacting parties as equals; and (iii) it focuses on the immediate relationship of doer to sufferer. This framework is important in so far as it establishes the fundamental values underlying the private law system. In all the cases that give rise to the paradoxes then, it is important to appreciate that the goal is to address wrongs between specific individuals, to restore matters back to their previous state, and not, by way of contrast, to address other kinds of injustice. For example, it is irrelevant to the aims of private law whether the plaintiff is otherwise hard-done-by, whether she’s down on her luck or has been dealt a harsh hand in life. These factors are the concern of distributive justice, which focuses on fairness more broadly, but not of corrective justice. The defendant need only answer to harms that he directly inflicted upon the plaintiff. In the absence of such harm, the plaintiff leaves empty-handed.

There are those who assume a radical response to the so-called paradoxes described above: that is, they contend that the typical (and intuitive) legal approach to all the scenarios above is simply wrong. On this view, even in the case of the gatecrasher scenario where the only evidence presented is a base-rate statistic, the defendant should be found to have failed to pay for his ticket. This contention is based on the simple fact that the base-rate evidence creates a balance of probabilities in favour of the plaintiff and against the defendant. After all, since either the plaintiff or the defendant must bear the cost, given the absence of any further information, the odds are still in the plaintiff’s favour.

On this view, the case of Smith v. Rapid Transit and many others like it would also be evaluated as wrongly decided. While this view is not entirely indefensible (precious few are), it is fundamentally inconsistent with the common law tradition. Finding against the defendant in the gatecrasher scenario would fail to treat him or her as an individual in a way that is seriously at odds with a fundamental guiding principle of the existing legal tradition.

Since the law is fundamentally conservative and changes are almost always incremental rather than radical, the onus is on radical reformers to establish why their approach should be preferred.

Instead, an approach that is largely consistent with existing law but still provides a framework for more subtle directional shifts enjoys an advantage. Proceeding on this assumption lends credence to the view that base-rate evidence alone does not provide a sufficient basis for a finding in favour of the plaintiff. Part of the explanation for this can be found in the distinction that exists between probability and weight of evidence. A 50/50 prior probability is assigned under circumstances of complete ignorance, but this does not mean that one has much confidence in a 50/50 objective chance under such circumstances. For a court to make a finding in the absence of any individual evidence may be to make a finding where there is a dearth of evidentiary weight. The need for sufficient evidentiary weight coupled with the law’s insistence on treating people as unique individuals rather than as members of a group, are arguably considerations of verdict acceptability that extend beyond the sphere of probability theory’s guidance.

It is important to distinguish between two questions: (i) the question of whether, based on the available evidence, the audience member paid for his ticket; and (ii) the question of what to do about it. In relation to the first question, on the available evidence, the answer is “probably not”. So if you are only interested in what’s most likely true on the available evidence, then the base-rate statistic will provide the best available answer. But this is distinct from the question of whether the law should take action based on evidence of base rate alone. The law is clear that the mere fact that someone probably did something is often not enough to empower another person to take action against him or her and is almost never enough for the State to.

The law’s procedures and presumptions include the thought that everyone is to be judged as an individual. That is, our justice system is shaped both by the desire that legal outcomes track truth but also by certain non-epistemic policy values that do not concern error reduction. Included in this class of non-epistemic policy values are considerations that sometimes bar the use of evidence illegally obtained without warrant (despite its relevance and truth-tracking

---

potential), the rule against re-trying the same matter twice (referred to as “res judicata” in Canada and otherwise known as “double jeopardy” in the U.S.) and the rule of spousal incompetency, which protects spouses from having to testify against one another. All of these rules are informed by policy considerations entirely distinct from truth seeking. The first is intended to protect citizens from inappropriate action by the police, the second to allow people to move on with their lives at the conclusion of a trial and the courts to move on to another case, and the third to protect the privacy and intimacy of the marital union. To infer, however, from these rules that the law rejects a standard probability analysis is a total non sequitur. An illegal wiretap may provide overwhelming evidence that the accused committed the crime. If this is the only available evidence, however, the accused may well escape conviction. It is in this way that the law, choosing to emphasize individuality, will not find someone guilty based on base-rate evidence alone; this does not amount to a rejection of the probability that the person did it, but rather to a non-epistemic choice to promote other goals in addition to truth seeking.

Any in-depth defence of the importance of these considerations to the law would take us beyond the scope of this thesis. Suffice it to say that in many different realms, we consider treating people as individuals to be essential to human dignity. Part of what we find so appalling about racism is not just the possession of negative views about a particular group of people, but rather the attribution of characteristics to an individual based on his or her membership within a particular group. Any proposal that runs contrary to this approach faces an uphill battle. It is not my aim to promote a particular view of how competing interests in the law should be balanced. I concede that there are circumstances where other considerations rightly trump verdict accuracy. My only concern here is to identify common loci for mistakes in arriving at verdict accuracy where verdict accuracy is what is at issue.

Unlike the gatecrasher scenario in which there is arguably an issue of insufficient evidence relevant to the individual, individual-specific evidence against relevant background evidence is at stake in the blue/green cab and Todhunter examples. While the academic literature tends to treat all of these problems as of a kind, there is a distinction to be made between considerations

---

124 In Truth, Error, and Criminal Law (Cambridge: Cambridge University Press, 2006), Larry Laudan argues that germane inculpatory evidence, even if illegally obtained, should be placed before the jury.
of weight – which are at issue in the gatecrasher paradox – and those of accuracy. Where uncertain direct evidence exists against the backdrop of a known base rate, to ignore the base rate is clearly to compromise accuracy and indeed runs contrary to sound probabilistic principles.

With respect to the blue/green cab problem, the debate between defenders of the Bayesian approach and those against it reaches an impasse. Defenders assert that the base rate should be considered as a countervailing consideration to the eyewitness testimony. The claim is that base-rate evidence just is no less relevant than case-specific evidence. It is argued that evidence that the applicable base rate is 80% should be treated as on a par with the evidence of an 80% reliable eyewitness. If an eyewitness were to report that she saw an alien, we would factor in our prior belief in the likelihood of alien existence in assessing the truth of the statement. It is contended that it is simply unjustifiable to ignore the base rate. Some base rates are more informative than others, but there is no minimum standard for which a base rate is relevant.125

On the other hand, those who see no legitimate place for probability in the law assert the irrelevance of base rates by reference to our intuition that eyewitness testimony is a reliable mechanism for generating true beliefs. There is also a sense in which the legal system may, as a matter of public policy, be loath to attribute features of the aggregate to the individual. Taking base-rate evidence, e.g. market share, into account may cause successful companies to attract more liability than those that fare less well in the market. But by this logic, we should be equally troubled if a cab company with a monopoly in a certain region were to be assumed responsible where it was established that a cab driver negligently hit the plaintiff. I don’t think the law would have trouble attributing responsibility in that case. In addition, it is possible to respond to the notion that eyewitness testimony is reliable by demonstrating that the mechanism is not as good as one might think and that our intuitions here are simply misguided. After all, there is reason to think that the law has traditionally had a strong empirical bias, relying too heavily on direct eyewitness testimony to the detriment of accuracy. Studies show that this evidence is in fact quite weak,126 and on this basis it might be argued that the law is simply due for reform. That is, while eyewitness testimony that the accused was at the scene of the crime properly

125 Supra note 119.
increases the probability that he did it, it may do so only slightly. In fact, rather than treating eyewitness testimony as highly instructive, the onus may properly be on the party relying on the eyewitness evidence to provide additional reasons for believing this particular witness to be especially reliable. That is, instead of presuming the accuracy of eyewitness testimony, the existence of strong evidence of eyewitness error should shift the burden of witness-specific evidence from the one challenging the witness to the one relying on her. In this way, it may be properly sufficient upon cross-examination of an eye witness to highlight the absence of any evidence indicating that she is more reliable than most; this would be in contrast to the current prevalent assumption that there must be some evidence of particularly poor eyesight or memory in order to place any significant limit on the witness’s reliability. And while non-Bayesians may very well accept that we should be careful to give evidence only the weight it properly deserves, the critical point here is the acceptance of the statistical base-rate evidence about witness reliability in general (or of this particular type if such base-rate evidence is available) as a starting point.

It is also the case that in other intellectual domains, we have come to accept that our intuitions can be seriously mistaken and that they should be revised to bring them into line with sound mathematical thinking. So-called “cognitive biases” are the focus of a large portion of psychological research. By “cognitive bias” I refer to distortions in the human mind that are difficult to eliminate, and that lead to perceptual distortion, inaccurate judgment, or illogical interpretation. Anchoring, the bandwagon effect, framing effects, and status quo bias are all common examples of cognitive biases. Anchoring is the common tendency to rely too heavily on a single fact when making a decision to the exclusion of other relevant information; the bandwagon effect describes the tendency to rely on the beliefs or actions of others as an indication of truth; framing effects are said to occur when different conclusions are drawn from the same pieces of information depending on the manner in which it is presented; and status quo bias requires no further explanation. For example, we often overemphasize the relevance of vivid anecdotes in the face of accurate statistical evidence. This could be viewed as a kind of anchoring effect. Consider this by way of illustration. Mary and Jane, both strongly myopic, are

127 Daniel Kahneman & Amos Tversky, "Subjective probability: A judgment of representativeness" (1972) 3 Cognitive Psychology 430.
asked whether they would consider undergoing laser eye surgery to correct their vision. They are also both provided accurate statistics on the risk of adverse complications. Let’s say for the purposes of this example that they are both told that there is a 95% success rate (where success is defined as an outcome that is the same or better than the one prior to the procedure) and a 5% chance of procedure failure (where failure is defined as an outcome that leaves the individual more visually impaired than before, due for instance to post-operative scarring). Mary has two friends who underwent successful laser eye surgery and are thrilled with the results. Mary may well be considerably more likely than Jane to opt for the surgery herself. If Mary were to elucidate the reasons for her choice and refer to her friends’ successes with the procedure, we would consider this to be an irrational basis for the decision; since Mary already knows that there is a 95% success rate with the procedure, she does not have any more relevant information than does Jane. In this example, we do not claim that the statistical evidence is somehow less relevant than the anecdotal; instead we consider the reliance on anecdotal evidence misplaced and pejoratively refer to the phenomenon as a “cognitive bias”. Once identified as such, the focus becomes how best to mitigate the bias and promote rational decision making.

4.3 Koehler and Shaviro’s Resolution to the Impasse

Let’s return now to the impasse between those who defend the law’s tendency to minimize or ignore base-rate evidence in the face of case-specific testimony and those who insist that base-rate evidence must be taken into account. A promising way forward may lie in showing that intuition and probability are not in contest in most cases and that where they are, probability theory can reveal cognitive biases. If this is so, then the proper application of probabilistic principles can perform both a descriptive and normative role for the law. In the same way that demonstrations of human “irrationality” only make sense against a backdrop of usually rational behaviour, an application of probabilistic principles to legal problems is most viable if it largely accords with existing legal practice. In this way, the application of probabilistic principles can be more easily accepted and used to correct only the penumbral cases.

To resolve the Todhunter problem, Koehler and Shaviro embrace the idea that intuition and probability actually align. I will argue that while their particular proposal suffers from a number of difficulties, the idea that it is possible to reconcile probability and intuition presents a promising way forward.
The paradox generated by the Todhunter problem is difficult to dismiss. While Bayesian defenders can plausibly posit that in the blue/green cab scenario, the witness who is 80% reliable simply provides insufficient evidence when presented against a high competing base rate, the Todhunter problem puts the contrast with intuition into stark relief. A 99.9% reliable eyewitness provides what would ordinarily be thought of as strong direct evidence, and yet does nothing to offset a very low prior probability. More disturbing still is the realization that such low prior probabilities are plausibly the norm in real trial situations. It is trivially true that the chance that any given individual committed a particularized offence is small. If the Todhunter problem cannot be resolved, either the common law approach to legal evidence is truly threatened, or probability theory must be conceded as irrelevant.

There is a certain similarity between the argument implied by the Todhunter problem and the Scottish philosopher David Hume’s discussion of miracles. That is, where Hume believes that he has found an argument against the existence of miracles based on the definition of a “miracle” and the strength of evidence required to offset an extremely high prior probability of impossibility, the Todhunter problem is designed to show that the low prior probability that any given individual committed a prior offence could in a similar way render the proof of guilt extremely onerous. Hume defines a miracle as a “violation of the laws of nature” such that the event in question has, by definition, an extremely low prior probability. The chance that any given individual would commit a particularized offence, while small, has by comparison a much larger prior probability. Hume argues that, “the proof against a miracle, from the very nature of the fact, is as entire as any argument from experience can possibly be imagined”. That is, by definition, a miracle has the lowest imaginable prior probability. If it did not, it would not be a miracle. As Hume says, “it is no miracle that a man, seemingly in good health, should die on a sudden: because such a kind of death, though more unusual than any other, has yet been frequently observed to happen.” The only way to “prove” a miracle would be by evidence equal in strength to all the evidence existing against it, and this, by definition, would never be

---

128 David Hume, *An Enquiry Concerning Human Understanding* (New York: Collier & Son, 1910) at section IX.
129 Ibid.
130 Ibid.
found. If we were to start observing men walking on water on a regular basis, we would at some point begin to look for an explanation for a change in the state of nature and no longer call this a miracle. Hume’s famous discussion of miracles nicely illustrates the power of prior probabilities and the need to take them into account when updating. His reflections also demonstrate why it is necessary to address the Todhunter problem in order to reconcile the common law approach and probability theory. We would not want the application of fundamental principles of probability to render findings of guilt “miraculous”.

Koehler and Shaviro, defenders of the relevance of base rates to verdict accuracy, offer a resolution to this dilemma on the basis that probability theory does not actually support the result upon which the paradox is based. Instead they say that a crucial mistake is made: the proposed probability analysis confuses the probability that the witness made some mistake (i.e. 0.1 %) with the probability that the witness made this particular mistake. On the latter interpretation, we arrive at a posterior probability of 99.9% that the number was in fact 297. Since this result is entirely consonant with intuition, no untoward consequences need follow for either the existing approach to legal evidence or a probabilistic understanding of it.131

\[
p(\text{"297"} \mid \sim \text{297}) = p(\text{mistake}) \times p(\text{"297"} \mid \text{mistake})
\]

\[
= \frac{1}{1000} \times \frac{1}{9999}
\]

\[
= 1.0001 \times 10^{-7}
\]

If all mistakes were equally likely, then the probability that the witness would say “297” given that he had made a mistake would be only 1/9999 since there are 9999 possible mistakes in a lottery of 10,000 tickets.

According to Koehler and Shaviro, the correct Bayesian solution to the problem is as follows:

\[
p(297 \mid \text{"297"}) = \frac{0.999 \times 0.0001}{(0.999 \times 0.0001) + \left(0.001 \times \frac{1}{9999} \times 0.9999\right)}
\]

\[
= 0.999
\]

131 Supra note 119.
4.4 An Alternative Resolution Based on Corroborating Evidence

Koehler and Shaviro’s suggestion that intuition and probability theory need not diverge will be embraced below. The particulars of the solution that they provide, however, do not suffice to allay the concerns generated by the paradox.

To begin with, there is not enough information provided by the example to generate a Bayesian solution. All we have is the witness’ accuracy, which tells us the probability of true positives and true negatives (namely 99.9%) and conversely generates the frequency of all false reports (namely 0.1%). But that is insufficient to generate the four probabilities necessary to solve the problem. In both the original formulation and Koehler and Shaviro’s version, $p("297" \mid \sim 297)$ is equated with 0.1%. But $p("297" \mid \sim 297)$ represents the number of false positives generated by the witness, which is not 0.1%. Instead, it is some unknown fraction of the total number of errors – both false positives and false negatives. In short, knowledge about the witness’ accuracy, which is all that we have in this case, is insufficient to generate the probability that the number is 297 when the witness reports that it is. See Table 2 below.

<table>
<thead>
<tr>
<th>Test Outcome</th>
<th>Actual Condition</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>True</td>
<td>False</td>
</tr>
<tr>
<td>Positive</td>
<td>True positive</td>
<td>False positive</td>
</tr>
<tr>
<td></td>
<td>←Positive predictive value</td>
<td></td>
</tr>
<tr>
<td>Negative</td>
<td>False negative</td>
<td>True negative</td>
</tr>
<tr>
<td></td>
<td>←Negative predictive value</td>
<td></td>
</tr>
<tr>
<td></td>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td>Sensitivity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Specificity</td>
<td></td>
<td>Accuracy</td>
</tr>
</tbody>
</table>

So, the simple solution to the “paradox” is to say that any attempt to apply a Bayesian analysis will not generate a paradox: it will tell you that you don’t have enough information, which is a quintessentially non-paradoxical response.
There is, however, another way to think about the problem. As has already been noted, what the “paradoxes” discussed above have in common is a paucity of evidentiary support. They rely on either base-rate evidence alone or base-rate evidence interacting with a single additional piece of evidence. In this, they create artificial scenarios that are not well suited for consonance with our intuitions. In the context of real trials, it is not usually a single piece of direct evidence, but rather multiple pieces of evidence that are at stake. In the absence of more than a single piece of evidence, the law legitimately draws a negative inference from the lack of any further evidence.

Thinking along these lines leads us to an understanding of why it might be appropriate to regard the relevant probability in the Todhunter paradox as the probability of the particular mistake. That is, we would indeed be interested in the chance that the eyewitness made the mistake of identifying 297 as the winning ticket on the added assumption that the number 297 had already been independently identified. If 297 were already suspected as the winning lottery number on the basis of any other evidence, another piece of eyewitness testimony, an entry made in the lottery books, the recollection of the convenience store clerk, etc., then the question would properly concern the chances that the eyewitness would make this particular mistake. Certainly the chances of some generic mistake, given the low prior probability of any given ticket winning, would still be high, but the chance that the witness would be mistaken in identifying the number 297, the very same number that was already (and independently) suspected, would be very small.

This solution differs from the one offered by Koehler and Shaviro in that instead of relying, as they do, upon a single witness’ testimony to say that the issue is the probability of saying “297” in error, 297 has already been identified by some other piece of evidence. I suggest that it is only in this way that it is plausible to suggest that we are interested in the probability of making this particular mistake. Absent corroborating evidence, we can only be concerned with the probability of the witness making any mistake. And that probability is high. The number 297 is in no way special to us until it is identified by some other evidence. So it is only by adding corroborating evidence to the picture that Koehler and Shaviro’s idea of focusing on the probability of making this particular mistake is salvaged.

This way of thinking about the problem suggests the power of independently corroborating evidence. Even where two independent pieces of evidence each have a significant probability of
error, the corroborative effect is extremely powerful, and can easily overcome even a very low prior probability.

Adding the assumption of two independent sources of identification, the scenario becomes this. Two witnesses independently testify that they saw the number 297 on the winning lottery ticket. We know that the number of values that the tickets may take on is 10,000. What is the probability that the winning number was actually 297? Even assuming that the witnesses are only 80% reliable, the result is as follows:

\[ X = \text{the winning number is 297} \]
\[ A = \text{witness 1 reports that } X \]
\[ B = \text{witness 2 reports that } X \]

What we are interested in is the probability of \( X \) given \( A \) and \( B \): \( p(X \mid A&B) \).

Given that,

\[ r = \text{the probability that } X \text{ is reported given } X, \text{ i.e., } p("X" \mid X) \]
\[ n = \text{all possible values of } X \]

According to Bayes’ theorem,

(i) \[ p(X \mid A&B) = \frac{p(X) p(A&B|X)}{p(A&B|X) p(X) + p(A&B|\sim X) p(\sim X)} \]

(ii) \[ p(X) = \frac{1}{n} \]

(iii) \[ p(\sim X) = 1 - \frac{1}{n} \]

(iv) \[ p(A&B \mid X) = p(A \mid X) p(B \mid A&X) \]

Since we have assumed that \( A \) and \( B \) are independent, we can treat \( p(B \mid A&X) \) as equivalent to \( p(B \mid X) \).
We have also assumed that both reports are equally reliable/unreliable, which brings us to this:

\[(v)\quad p(A\&B \mid X) = p(A \mid X) \cdot p(B \mid X)\]

Next, the chance of an incorrect report is 1-\(r\), and there are \(n-1\) possible values that could be randomly reported, such that

\[(vi)\quad p(A \mid \sim X) = (1 - r) \times \frac{1}{n - 1}\]

Given (v) and (vi), this becomes

\[(vii)\quad p(A\&B \mid \sim X) = \left((1 - r) \times \frac{1}{n - 1}\right)^2\]

Substituting the above into (i), the following can be derived:

\[p(X \mid A\&B) = \frac{\frac{1}{n} \times r^2}{r^2 \times \frac{1}{n} + \left((1 - r) \times \frac{1}{n - 1}\right)^2 + \left[1 - \frac{1}{n}\right]}\]

\[= \frac{nr^2 - r^2}{nr^2 - 2r + 1}\]

Substituting our assumed values of \(n = 10,000\), and \(r = 0.8\) yields a value for \(p(X\mid A\&B)\) of 99.9%.\(^{132}\)

What the above derivation is intended to demonstrate is the power of corroborative testimony. Even when starting with a low base rate, or dealing with multiple pieces of evidence that have a not-insignificant possibility of error, two independent pieces of corroborating evidence can lend great support. This is both borne out by the probability calculus and by intuition: if \(X\) were not the case, it would be highly unlikely for two independent pieces of evidence to point to it.

Intuitively, the greater the array of available possibilities, the less likely it is that the same particular individual or event would be independently accidentally identified.

All this is to say that while the gatecrasher paradox does make the point that base-rate evidence is admittedly insufficient to ground a legal determination, none of the paradoxes leads to the conclusion that the consideration of base rates along with other error-prone evidence (as is contemplated in the red bus/blue bus example where both market share and witness evidence are available) is not appropriate to the pursuit of verdict accuracy. That is, it is possible to make a compelling case that the Bayesian result is the right one and that a single-piece of error-prone evidence against a low prior probability should not lead to the conclusion that the hypothesis has sufficient support. In the context of real legal decision making, it would be extremely odd to be faced with anything close to the low prior probability described in the Todhunter case. This alone should allay concerns that Bayesianism yields counterintuitive results. In addition, I have demonstrated that even were decision makers to face very low initial prior probabilities, this issue would in most cases be quickly addressed by even a single piece of corroborating evidence. That is, even a single piece of corroborating evidence would allow the direct evidence adequately to countervail even very low prior probabilities. Accordingly, to embrace a Bayesian approach need not open the door to drastic revisionist consequences for legal evidence. On the contrary, Bayesianism and intuition align in most cases.

Admittedly, we are still left with the potentially troubling prospect that a single piece of eyewitness testimony introduced against the backdrop of a low prior probability event should not lead to the conclusion that the witness’s claim is true. The path of intellectual honesty may lead to results that we otherwise eschew. What I have tried to demonstrate though is that upon reflection this conclusion is not actually unpalatable. It is not just fancy statistical equations that would lead us to seek further corroboration for an eyewitness claim made against a background of low prior probability; clear and considered thinking will yield the same result.

Of course, the multiple probabilistic approaches to the Todhunter problem discussed above do yield another non-trivial concern: if it is possible to demonstrate a posterior possibility of both 9% and 99.9% depending on two not utterly implausible interpretations, then is not the normative power of the probabilistic approach undermined? Specifically, was Tribe right to
worry that, “the costs of attempting to integrate mathematics into the fact-finding process of a legal trial outweigh the benefits”\textsuperscript{133} and are apt to confuse rather than clarify?

In the last section of this dissertation it is argued that the answer to this question is wrapped up in the question of whether probability can be made truly intuitively accessible. Any blind application of the probability calculus to factual determinations in the legal decision-making process is indeed unhelpful and counterproductive. This is so both because of the complexity of the operations involved and the possibility for grossly erroneous results based on faulty assumptions that are too easily made. Instead, it is argued that we can extract from the probabilistic approach certain basic principles that can be intuitively absorbed and applied to improve verdict accuracy.

The section above has contributed to dispelling the misconception promoted by the so-called paradoxes that the consideration of base rates necessarily yields counterintuitive results. If intuition and basic principles of probability can be brought into line, a beneficial symbiosis between the two is rendered possible.

4.5 Simple Heuristics as a Possible Alternative

Before embarking on a discussion of whether base rates can and should play a useful role in the decision-making process, it makes sense to consider the alternative. If we do not allow probabilistic principles to inform factual determinations in judicial decision making, upon what will we base our decisions instead? It may be easy to find fault with the Bayesian approach, but if no alternative normative framework exists, then the position advocated in this work is strengthened.

Left to their own devices, most people rely on so-called “simple heuristics” to resolve complex problems involving uncertainty. This descriptive fact about human reasoning is not in dispute. Research into human reasoning, such as the studies most famously conducted by Kahneman and

\textsuperscript{133} \textit{Supra} note 115 at 1377.
Tversky, reveal that people rely on a limited number of “fast and frugal” heuristics to solve complex problems quickly.\textsuperscript{134}

That humans use the “fast and frugal” heuristic approach is also tightly tied up with the notion of bounded rationality. As Andy Clark puts it,

\begin{quote}
Human minds were not designed as instruments of unhurried, fully informed reason. They were not designed to adhere to rigid or even consistent preference orderings or to act so as to maximize reward on all occasions. Human minds, it is now widely agreed, are better understood as loci of only bounded rationality: reason restricted by a variety of evolutionary and pragmatic factors.\textsuperscript{135}
\end{quote}

The question of present interest is whether we should be satisfied from the perspective of accuracy with the results generated by the “fast and frugal” heuristic approach. That is, assuming that people do, as a matter of fact, reason using simple heuristic strategies, do these strategies yield good results from the perspective of verdict accuracy? If so, then perhaps probability theory does not have a useful role to play.

The suggestion that simple heuristics might be not only efficient but also effective short-cuts that “make us smart” has been made by Gerd Gigerenzer among others and is discussed at some length by Alvin Goldman in “Simple Heuristics and Legal Evidence”.\textsuperscript{136}

One candidate for a simple heuristic that might “make us smart” is the “take the best” heuristic (“TTB”). According to this rule-of-thumb, once a particular cue or indicator has been judged to have high cue validity, then this cue alone is used to make determinations, ignoring other potentially relevant information. The idea is to “take the best, and ignore the rest”. For example, if presented with pairs of cities and asked to judge which city is larger, the “take the best”

\begin{footnotes}
\textsuperscript{134} Supra note 118.
\textsuperscript{136} Supra note 96.
\end{footnotes}
strategy would indicate that once a good cue such as “possession of a soccer team” has been identified, it is used to make all determinations, ignoring other potentially relevant information. After considering an array of potentially relevant cues – whether the city is the capital, whether the Olympics were ever held there, and so on – the best cue among the bunch is selected and used exclusively to make the city-size determination. Gigerenzer et al. claim that the accuracy potential for the TTB strategy is high and that this simple heuristic as well as others fare well when compared to more complex strategies. ¹³⁷

For our purposes, the potential merits inherent in embracing the heuristic approach are apparent. To begin with, decisions arising out of heuristic intuitions are likely to conform to the intuitions of those who cursorily evaluate these decisions, thus bolstering the fundamental legal tenet that justice should not only be done but also appear to be done. That is, if a result is intuitive, then it will also appear correct to casual observers. If complex formulas and strategies requiring greater time and effort are necessary to arrive at accurate conclusions, those who superficially review these conclusions are less likely to read the rectitude of the decision on its face.

Another plausible advantage to this approach is that it seems to reduce the potential for the kinds of glaring errors that are often made when statistical evidence is introduced into the decision-making process but is not fully understood by decision makers.

In light of this concern, the appeal of embracing intuition is apparent. Perhaps in the most ideal of worlds, intuition would be accurate. The question remains, however, whether the fast and frugal approach does in fact generate accurate results. As Goldman points out, the accuracy of the TTB approach is critically dependent upon the accuracy of the cue. That is, if possession of a soccer team does in fact correlate highly with city size, then the TTB approach in conjunction with knowledge of a highly correlated cue will yield highly accurate results. Of course, if the “best” cue has been selected in error, then the accuracy of the strategy is also lost. ¹³⁸

¹³⁷ Gerd Gigerenzer & Peter Todd, Simple Heuristics That Make Us Smart (Toronto: Oxford University Press, 1999).
¹³⁸ Supra note 96.
Goldman looks at a few selected situations in the legal context where cue validity may be thought to be poor: eyewitness identification testimony, criminal confessions, and expert testimony. The assumption is that these cues are mistakenly considered highly reliable. Goldman canvases a number of possible remedies to address this potential problem. He considers cue exclusion, scientific testimony about a cue’s validity, juror empanelment and influencing cue validity. While the first two are self-explanatory, the second two require some further elucidation. The juror empanelment “solution” involves selecting for the decision-making job only those jurors with good cue validities. That is, presumably questions could be put to potential jurors in order to assess whether or not they would be likely to overestimate the cue validity of eyewitness identification testimony, criminal confessions, or expert testimony. For example, hypothetical scenarios could be put to potential jurors that involve considering a few different types of evidence. If a potential juror overemphasizes one type of evidence versus another, this could potentially disqualify him or her. Such questioning would be not unlike the existing “challenge for cause” procedure whereby potential jurors are asked the following question: “Would your ability to judge the evidence in this case without bias, prejudice or partiality be affected by the fact that the person charged is Black”\textsuperscript{139} The fourth remedy, namely influencing cue validity, involves changing objective cue validities to bring them in line with subjective cue validity. The idea here would be to take steps to actually render eyewitness identification testimony, criminal confessions, police testimony, and expert testimony more accurate.

The suggestion to exclude the cues in question because of their potentially distorting effect is highly problematic. Goldman himself mentions the troubling fact that we have no way of knowing whether the second-best cue that would replace the offending cue would be any more reliable. Exclude eyewitness testimony and perhaps the decision maker will now rely too heavily on the circumstantial evidence instead. There is also a sense in which eyewitness testimony, confessions, and expert evidence aren’t “cues” in the way that soccer teams are cues of city size. Large cities happen to have soccer teams, but the evidence to which Goldman

\textsuperscript{139} Whether racial screening questions are effective at reducing biased results is discussed in RA Schuller, V Kazoleas & K Kawakami, “The Impact of Prejudice Screening Procedures on Racial Bias in the Courtroom” (2009) 33 Law Hum Behav 320.
refers, while possibly prone to error, is not merely *incidental* to the question at issue. In addition, Goldman points out that for excluding evidence with poor cue validity to work, judges would be required to accurately assess both juror’s subjective cue validity and objective cue validity so as to make an ultimate assessment as to whether exclusion would or would not be appropriate. This places an unrealistic burden of knowledge on judges. However, Goldman sees this solution as potentially redemptive at least in the criminal context where the standard is “beyond a reasonable doubt”. At least here, if the mistakenly inflated cue is removed, then verdict accuracy will be improved: the individual will not be convicted given the absence of any sufficiently reliable evidence.

The more obvious problem with this proposed solution is that it would exclude far too much. Remove eyewitness identification testimony, criminal confessions, police testimony, and expert testimony, and precious little is left. One would have to have an extremely high level of confidence in the reliability of the remaining cues for it to make even the slightest sense to entertain such a possibility.

The second option, namely the introduction of scientific testimony about a cue’s validity, is equally impractical. Goldman suggests that there is reason to think that this would ameliorate the problem. But on what grounds? Since expert testimony has already been identified to have poor cue validity – that is, decision makers put a disproportionate amount of weight on the opinions of experts – introducing additional expert evidence will surely compound rather than correct the problem. Under such circumstances, it is more than likely that decision makers would judge the expert testifying about the other expert evidence to be the “ultimate” expert and accordingly entirely defer to her testimony. As a result, decision makers may simply take to heart the message that the other experts are not to be trusted. In this way, the usefulness of the first-order expert testimony is negated and we are returned to the problems inherent in the cue exclusion solution.

The third remedy seems most impractical of all: surely it is difficult enough to attempt to eliminate jurors with extreme prejudices that could dramatically bias verdicts. The suggestion that jurors undergo additional testing for cue validity choices that could potentially arise in the course of the trial is absurd. To begin with, at the outset of a trial, how is one to know which cue validities will be invoked as the evidence unfolds? Even if one were to compose a list of likely
suspects (excuse the pun), there would be the same concern expressed above with respect to the exclusion of cues. We may exclude jurors on the basis of overconfidence in expert testimony, but there is no assurance that the alternate cues relied on by the remaining jurors will yield better results. In addition, the process of juror elimination based on poor cue validities might ultimately yield a jury of social scientists which would have to be defended as desirable.

Even Goldman dismisses rather readily the fourth remedy: namely, actually changing objective cue validities to bring them in line with subjective cue validities. There are many circumstances under which this goal may be laudable. For example where coerced confessions are primarily responsible for the fallibility of confessions, eradicating this source of error would indeed be desirable. However, given the loftiness of this goal, it is hardly a plausible candidate to address the problem of inaccurate cue validity. This suggestion is a little like proposing to deal with the school bully by achieving world peace.

Despite acknowledging certain difficulties inherent in the practical implementation of his suggestions to improve upon the heuristic approach, Goldman ultimately thinks that proceeding along one or several of the lines suggested may be ultimately ameliorative.

There is, however, an aspect of epistemic circularity or bootstrapping that attaches to any suggestion to redress deficiencies in cue validities: the same problems inherent in the original cue validities apply to the processes introduced to redress them. The concern as to whether or not the approach taken in processing the information is reliable remains, irrespective of how many layers of additional information are added. Those who embrace the intuitive approach need to answer to the charge that there is no explicit standard against which these heuristics can be judged to yield accurate results. If probability calculus is the standard against which these heuristics are judged, then the onus is on proponents of the heuristic approach to show that standard probabilistic principles such as the one dictating the relevance of base rates do not apply. If probability theory is not acknowledged as the appropriate evaluative standard, then what is?

Goldman’s account also fails to make explicit the important role played by priors in identifying what the “best” cue is. That is, we assess whether a cue is working well or not based on our existing knowledge about the right answers to the questions the cue is helping us to answer. Without any knowledge about which cities are larger than others, we would have no hope of
discovering that the existence of a soccer team provides a good cue. We judge cues based on a comparison of the cue indication to a known set of information: if I know that Toronto is bigger than Sudbury and Montreal is bigger than Lefebvre and I know that Toronto and Montreal have soccer teams and Sudbury and Lefebvre do not, then I can begin to consider the possession of a soccer team a reliable cue for city size. Without prior knowledge, there is no way to even know where to begin.

One of the major weaknesses of Goldman’s approach is that if we simply accept our intuitions, not only do we beg the question of whether Bayesian probability is actually counterintuitive, but we are also left without any normative framework. Surely there is room for improvement in arriving at accurate conclusions, but a heuristic approach fails to provide any direction in this regard. It also fails to lead the way in terms of how decision makers can effectively make use of statistical evidence. From DNA evidence to actuarial tables, explicit statistics can provide extremely useful information to decision makers; an account of legal decision making that does not show how such information may best be approached is clearly lacking.

4.6 A Possible Place for Base Rates

Two major points were argued for above: (i) that there is an alternative and preferable explanation from probability that resolves the Todhunter paradox and also accords with intuition; and (ii) that the heuristic approach alone as an alternative to Bayesianism has unacceptable deficiencies. Even conceding that the potential for multiple probabilistic interpretations of the same scenario raises certain significant concerns, these two points taken together do suggest that there may be a place for base rates in our approach to legal evidence.

In particular, the approach based on corroborating evidence that was set out above suggests that it is possible to reconcile the constraints imposed by the probability calculus with the bulk of common legal intuitions. What Bayesianism in general, and the consideration of base rates in particular, have to offer is a normative framework for decision making under uncertainty that has proven fruitful in other disciplines. Since the shortcomings of the use of explicitly statistical

---

140 Supra note 93.
evidence focus on concerns that the results generated are counterintuitive and that such evidence is apt to mislead, an alignment of probability and intuition is poised to solve both problems.

Tribe’s statement of the problem in his classic paper on the topic nicely encapsulates the concern:

The very mystery that surrounds mathematical arguments - the relative obscurity that makes them at once impenetrable by the layman and impressive to him - creates a continuing risk that he will give such arguments a credence they may not deserve and a weight they cannot logically claim.\footnote{\textit{Supra} \mbox{note 115.}}

To put the concern in this way, however, is also to point the way to a resolution: if the mathematical arguments can be made penetrable and their fallibility exposed, then the risk will be dispelled along with the mystery. If probabilistic principles can be absorbed intuitively upon some more reflection, then explicitly probabilistic evidence, grasped in this way, also loses its ability to mislead. By understanding the relevant probabilistic principle fully, rather than perceiving it as confusing or difficult, decision makers will not be tempted to abdicate their decision-making responsibility.

While the studies of Kahneman and Tversky suggest that the use of heuristics leads to gross errors from a Bayesian perspective, other studies have indicated that when the same problems are presented differently, the rate of error declines sharply.\footnote{Leda Cosmides \& John Tooby, “Are Humans Good Intuitive Statisticians After All? Rethinking Some Conclusions from the Literature on Judgment Under Uncertainty” (1996) 58:1 Cognition 26.} For example, a study by Gigerenzer and Hoffrage\footnote{Gerd Gigerenzer \& U Hoffrage, “How to improve Bayesian reasoning without instruction: Frequency formats” (1995) 102 Psychological Review 684.} showed that some ways of phrasing story problems are much more likely to evoke responses that accord with Bayesian probability. What has been called “natural frequencies” tend to draw out results most in accord with those generated by the probability calculus.
As was mentioned in an earlier chapter, presenting probabilities in terms of natural frequencies is in a sense not strictly correct. That a test has 80% accuracy does not justify one in specifying that 80 out of 100 results will reflect reality. 80% accuracy only means that in the long run accurate results will approach 80%. Since 100 is not the long run, we have no guarantees that 80, as opposed to, say 79 or 81, is the right number here. Nonetheless, the natural frequency approach provides a very useful tool for working with probabilities and yields more accurate results for those without statistical expertise.

To understand what it is meant by a “natural frequency”, let us look at an example. Suppose that a barrel contains many small plastic eggs. Some eggs are painted red and some are painted blue. 40% of the eggs in the bin contain pearls, and 60% contain nothing. 30% of eggs containing pearls are painted blue, and 10% of eggs containing nothing are painted blue. What is the probability that a blue egg contains a pearl? Under a “natural frequency” presentation, the information about the prior probability is included as part of the conditional probability information. So, to present the egg example in terms of natural frequencies would be to say that 40 out of 100 eggs contain pearls, 12 out of 40 eggs containing pearls are painted blue, and 6 out of 60 eggs containing nothing are painted blue. On this presentation, it becomes far clearer that the probability that a blue egg contains a pearl is 2:1.

But what does this mean in the legal decision-making context? Returning to the Todhunter problem, could we find a natural frequency presentation that would coax a response in line with the probabilistic analyses offered above?

Presented in terms of frequencies, the answer that accords with Bayesian probability can be rather readily understood. Out of 10,000 tickets, only 1 is the winning ticket. 9999 tickets are losers. Given the witness’ accuracy rate of 99.9%, if she is shown the winning ticket, the chance is 99.9%, or 0.999, that she will identify it as the winner. Out of the remaining 9999 tickets that witness could be shown, she would falsely identify 0.1% of these as winners. 0.1% of 9999 is 9.999 tickets. Therefore, the total number of tickets that the eyewitness would identify as winners in 10,000 tickets would be 0.999 (correctly) + 9.999 (incorrectly) = 10.998.

---

144 Supra note 90 at 23-24.
0.999/10.998 = 0.09, that is 9%. And as the number of tickets in the pool increases, so does the absolute number of losing tickets and accordingly the proportion of losers incorrectly identified as winning tickets. That is, the bigger the pool of tickets, the less we should rely on the 99.9% accurate witness.

The alternate solution to the Todhunter paradox based on corroborating evidence presented above also suggests what further conclusion might plausibly be drawn from such a realization: namely that some additional independent corroborating piece of evidence would be required to arrive at a sufficiently supported conclusion. What has been argued in this chapter is that this result is both in accord with considered intuition and sound probabilistic reasoning.

The complex calculation of the power of corroboration demonstrated earlier in this chapter comes down to the following simple concept understood in terms of frequencies: witness 2 also has an accuracy rate of 99.9%. The chance that she will identify the winning ticket correctly is 99.9%. We have already established for witness 1 that the chance of falsely identifying a losing ticket as a winner is 0.1% of 9999 or 9.99. But the chance that witness 2 would identify any of the tickets already falsely identified by witness 1 as a winner would be 9.999/9999, i.e. only 0.1%. Therefore the chance that the ticket that both witnesses identify as the winning ticket would in fact be the winning ticket would be 1-0.1% = 99.9%.

The implications of any demonstration that intuition and probability often align could be considered a victory for either those who are satisfied with the heuristic approach or those who believe that Bayesianism provides a good normative framework for decision making. If it is possible however to bring probability and intuition largely in line, concerns about mystifying mathematics may be left behind while retaining the benefits of the probability calculus as a potential constraint on certain ill-considered intuitions. Empirical data supporting the conclusion that individuals do in fact take base rate information into account when it is presented in a more accessible form would lend credence to this view. To accept that probability may have this role to play in legal decision making could also pave the way for a less fearful approach to other varieties of statistical evidence.
Chapter 5
Medical Evidence: Constraints on Assertion and Justification

Algemon: The Doctors decided that Bunbury could not live, so Bunbury died.

Lady Bracknell: I’m glad to see that Mr. Bunbury had such confidence in the opinion of his physicians and acted under proper medical advice.

~Oscar Wilde, The Importance of Being Earnest

5 Medical Evidence

The purpose of this chapter is to juxtapose the modern medical profession with the law and to encourage comparison: that is, while medical practitioners fall prey to the same kinds of conceptual failures we see in legal discourse, these disparities are not generally used to criticize Bayesianism as a relevant normative model but rather to inform physicians’ education. The law can accordingly stand to learn from modern medicine’s approach to statistical evidence.

While not invulnerable to Wilde’s trenchant sarcasm, a physician’s opinion has more to recommend it today than it did in past centuries. This improvement is due in part to more rigorous methodology, in part to an expanded knowledge base acquired through the application of this methodology, and in part to the more widespread accessibility of up-to-date information. So-called modern medicine is distinguished by its reliance on large scientific studies to inform best practices. On the basis of such studies, medical colleges produce treatment guidelines that, at least to a certain extent, systemize and regularize physicians’ approaches.

In theory, Bayesian approaches figure significantly in both the development of medical guidelines and in medical practice. Prior probabilities must be accounted for and updated in light of new information. Ignoring prior probabilities may yield more glaring errors in medicine than in law; a sick patient is a clearer sign than a faulty verdict. Despite this, when reasoning under uncertainty, medical practitioners are susceptible to error in much the same way as are legal decision makers. At the same time, while there is theoretical controversy in the medical domain, less deference appears to be paid to the existing practices of medical practitioners. That is, there appears to be a stronger normative aspect to theorizing in medicine than in the law. This chapter
looks at the medical approach to knowledge acquisition with a view to applying lessons learned there to legal decision making.

The first section provides some illustrations of the importance of considering prior probabilities in the design and interpretation of medical research. Next, the integration of basic probabilistic principles at the level of medical study design is explored. This is followed by a discussion of principles underlying medical guidance and pedagogy:

- The role of priors in medical research: some examples
- Medical evidence: constraints on assertion and justification in the medical context
  - Medical study design: randomization and the importance of priors
  - Theories about the practice of medicine
    - EBM
    - Criticisms of EBM
  - Actual medical practice
    - A model of clinical medical reasoning
    - Teaching doctors to be better Bayesians

Techniques for teaching doctors to be better Bayesians may have the most direct application to the legal domain once the normative role of basic principles of probability in the law is acknowledged.

5.1 The Role of Priors in Medical Research: Some Examples

As mentioned above, prior probabilities are significant in both law and medicine. In medicine, however, they are more widely recognized as such, and overlooking prior probabilities leads to more glaring errors.

An example of the explicit consideration of prior probabilities in medical study design illustrates the point. A recent article in the New York Times suggested that expensive shoes are no more effective for injury prevention than are cheap running shoes. The article refers to a number of

---

studies in support of the proposition that injuries occur with equal frequency over time, whether the runner wears cheap, moderately priced or expensive running shoes. In one British study, cheap shoes were defined as those that cost £40-45. Medium-priced shoes were defined as those that cost £60-65, and expensive shoes could be obtained for £70-75. One of the studies referred to in the article was reported in the British Journal of Sports Medicine in 2008. In this randomized controlled trial (RCT), all subjects had size 8 or 10 feet, no gait abnormality or previous or current history of lower limb pathology, no foot or leg length discrepancy, and no other apparent disability that might in some way affect their gait, such as a visual impairment or a walking aid. The subjects were then randomly assigned cheap, medium or expensive shoes (as defined above) and plantar, i.e. foot, pressure was measured. Plantar pressure measurements were recorded from under the heel, across the forefoot and under the great toe. Differences in plantar pressure were recorded between shoe models and between shoe brands in relation to cost. The results of this study showed that low- and medium-cost running shoes in each of the tested brands provided the same (if not better) cushioning of plantar pressure as did the high-cost running shoes. And since other studies had confirmed that runners who exhibit relatively high and rapid impact forces, that is, low cushioning effect, are at increased risk of injury, the further conclusion could be drawn that high-cost running shoes are no more likely to prevent injury than their low or medium cost alternatives.

Particular prior probabilities must be taken into account for the results of such a study to be meaningful. For example, if a subject had a gait abnormality, previous or current history of lower limb pathology, a foot or leg length discrepancy, or some other apparent disability that may affect her gait, this would affect the prior probability of injury. That is, we would already expect someone with known foot pain to be more likely to suffer future pain, all other things being equal. If a clinical study were to show that those who wear orthotics have more problems with their feet than those who don’t, we should look to see whether prior probabilities have been accounted for; those with troublesome foot and leg structures are the ones who get orthotics in the first place. If they also are more prone to injury, it may be because their prior probability of injury is already higher than the norm.

---

By selecting subjects with, “no gait abnormality or previous or current history of lower limb pathology, no foot or leg length discrepancy, and no other apparent disability that might in some way affect their gait”, the study controlled for known factors that would affect the prior probability of the circumstance under investigation. The further random assignment of subjects to cheap, medium or expensive shoes is intended to control for unknown variables. More will be said below on the merits of randomization. For present purposes, however, it is sufficient to note that prior probabilities play a relatively apparent role in the design of medical studies.

In general, medicine tends to adapt to high-quality research results and to change practices accordingly more quickly than does the law. But, change in practice does take time even in the medical context. For example, in September 2008, the Globe and Mail ran an article on a study reported in the New England Journal of Medicine that confirmed the results of a 2002 study demonstrating that arthroscopic surgery for osteoarthritis of the knee is ineffective. The results of the 2002 study were questioned on the basis of that study’s methodology and the widespread use of arthroscopic surgery for osteoarthritis of the knee continued in the medical community.

In 2008, a large randomized controlled trial was conducted to compare physical and medical therapy on one hand with arthroscopic treatment in addition to those therapies. The study was conducted over eight years and involved nearly 200 participants and seven different surgeons. This 2008 study confirmed the results of the 2002 study in showing that there was no benefit from arthroscopic treatment when performed in addition to physical and medical therapy. Patients who received the surgery did have improved results in the first three months post-surgery, but this positive effect was both short-lived and anticipated since other studies have shown that “sham” surgery is associated with a fairly large but short-term placebo effect.

The example of arthroscopic knee surgery demonstrates that time and the repetition of results are often necessary to change prevalent medical practices. This fact itself may actually be used as an


example of good Bayesian practice. That is, if the medical community has a sufficiently high prior belief that a certain surgical intervention is effective, it may properly take more than one study to derail this belief. All studies have methodological limitations such that the effectiveness of any one study to change medical practice properly depends both on the prior probability of the originally-held belief and the strength of the new evidence. The possibility of follow-up studies that address some of the methodological concerns and either confirm or call into question previous results leads to a rigorous approach within the discipline. Since any discipline starts with its own priors – it is, after all, necessary to start somewhere – strong evidence is required to modify these beliefs. The more entrenched a prior probability, that is the more evidence there is already in support of that prior, the more new evidence to the contrary is required to dislodge it.

5.2 Medical Evidence: Constraints on Assertion and Justification in the Medical Context

Bayesian probability principles enter into medicine at various levels: at the level of clinical study design, the development of clinical guidelines, and the practices of individual physicians. The law can learn from the debate that has occurred in medicine at all these levels of discourse. Current educational approaches to teaching physicians how better to understand proper probabilistic principles are of particular interest.

Medical evidence is provisional, defeasible, emergent, incomplete, constrained, and collective.\textsuperscript{149} In this, medical evidence is both similar to and different from legal evidence. Legal evidence is provisional and defeasible in that until a final decision is made – although only until then – additional evidence can be brought to refute it. Legal evidence is also subject to such checks as cross examination. Ultimately, legal evidence can either meet the standard of proof – whether it is a balance of probabilities in a civil case or beyond a reasonable doubt in the criminal context – or fall short of it. If the evidence fails to meet the applicable standard of proof, then there is automatically a victory for the defendant or the accused.

A good example of the provisional and defeasible nature of medical evidence is the development of the medical understanding of peptic ulcers. In the 1970s, peptic ulcers were thought to be

\textsuperscript{149} Ross Upshur, “Seven characteristics of medical evidence” (1999) 20 Theoretical Medicine and Bioethics 229.
stress related and treatable with a “white” diet focusing on rice and bread. In the 1980s, high acidity secretion was thought to be the cause of peptic ulcers, and hydrogen ion blockers were the preferred treatment. In 1982, Australian physicians Robin Warren and Barry Marshall first identified the link between Helicobacter pylori (H. pylori) and ulcers, concluding that the bacteria, not stress or diet, causes ulcers. It was not until the 1990s that bacterial infection by helicobacter pylori was identified as the most common cause of peptic ulcers and targeted antibiotics became the treatment of choice. 150

That medical evidence is emergent, and so apt to change with time, often creates a tension between reality and expectation where patients are concerned. Given the importance of health to any patient, that the advice they are receiving is good advice only at the time, can be difficult to accept.

Medical evidence is also incomplete because of the constraints that are placed upon it by ethics, economics and computational limitations. An example of a computational constraint is the necessary trade-off between maximum information and time spent. Medical evidence is clearly collective in the sense that it is derived from a community of inquirers.

Taken together, all these aspects render medical evidence “fallible”, in the sense used by C.S. Peirce and Karl Popper, as well as underdetermined. If a patient complains of an earache, she may be suffering from a headache. Any piece of evidence can always be misinterpreted. Part of the judgment involved in assessing whether the available evidence in relation to a particular patient’s ailment indicates ailment x or y involves accounting for prior probabilities. Medical students are told to govern themselves according to the adage that “common things are common”, or that “when you hear hoof beats, think horses, not zebras”. Young clinicians are more likely to suggest an unusual diagnosis in the hopes that they will be the only doctor to make the correct diagnosis. More experienced clinicians tend to think that the correct diagnosis is usually something common. For example, if a patient seen in North America has an enlarged

150 JP Gisbert et al, “Helicobacter pylori eradication therapy vs. antisecretory non-eradication therapy (with or without long-term maintenance antisecretory therapy) for the prevention of recurrent bleeding from peptic ulcer” (2004) 2 Cochrane Database of Systematic Reviews, Article Number: CD004062. DOI: 10.1002/14651858.CD004062.pub2; Department of Health and Human Services, Centers for Disease Control and Prevention, Division of Bacterial Disease, “Helicobacter pylori and Peptic Ulcer Disease” online: http://www.cdc.gov/ulcer/history.htm.
spleen it is rarely the result of malaria. Instead, portal hypertension and mononucleosis are the most common causes.\textsuperscript{151} What may be referred to as an experienced physician’s “good judgment” is shaped by acquiring knowledge of the prior probabilities of various ailments. Since the prior probability of a patient experiencing a tension headache is many orders of magnitude higher than that he will have a brain tumor, brain tumors are not the primary explanation for head pain (despite what a Woody Allen character would say).

Of course there is nothing unique to medicine about the problem of under determination. Rene Descartes famously sought to doubt any and all of his beliefs by positing that there was at least one other competing interpretation of his sensory experience: the existence of a deceiving evil demon. Nelson Goodman’s “New Riddle of Induction” suggests that what we take as evidence for certain inductive generalizations could also support other generalizations. Looking at all known emeralds and seeing that they are green equally well supports the finding that all emeralds are “grue”: green and examined before time t or blue if and only if examined afterwards. The consequence of this is that evidence alone can never be entirely sufficient to justify a given interpretation. There is always a place for interpretation or judgment.

At the same time, while there will always be some art involved in judgment, the avoidance of base rate neglect can contribute to its demystification. If your personal prior probability for the existence of an evil demon is close to zero, then you won’t be overly troubled by Descartes’ dilemma. It is of course possible that you are wrong, but such uncertainty is an unavoidable artifact of our far-from-complete knowledge of the world.

5.2.1 Medical Study Design: Randomization and the Importance of Priors

Clinical medical trials are typically designed to ascertain if and to what extent a particular drug or treatment procedure has certain effects: both the desired effect of symptom alleviation and the undesired (though not necessarily undesirable) side effects.

\textsuperscript{151} The American Journal of Medicine blog, online: http://amjmed.blogspot.com/2009/08/common-sense-is-not-so-common-what-we.html.
Medical trials often divide a sample of subjects all suffering from the same condition into two groups. One group, the test group, receives the treatment that is under investigation. The other group, the control group, does not. Control and randomization are considered essential features of medical trials, intended to neutralize potentially misleading confounding factors. This approach is sometimes referred to as the Fisherian or classical method. For example, if age is known to affect the course of a condition, then this factor is directly controlled for by ensuring that the test group and control group terms are indistinguishable in terms of age composition. Random assignment to one of the two groups is used to “control” for unknown confounding variables.

One standard use of a control group in medical study design is to address the existence of the placebo effect, namely, the usually short-lived beneficial effect that comes from the simple receipt of medical attention. The control group is given a placebo, that is, an inert substance that looks like the treatment medication, to control for this effect. If the group that is receiving the treatment under study does not improve any more than does the placebo group, the treatment under consideration is deemed ineffective.

While known prognostic factors such as age and the placebo effect can be controlled for, there are an almost unlimited number of possible prognostic factors that cannot be. The technique of randomization was developed to neutralize these so-called “nuisance variables”. Randomization assigns subjects to the test and control groups “at random” in an attempt to “wash out” other variables that may affect results. After subjects are assessed for trial eligibility, but before treatment begins, subjects are randomly assigned to receive one or another treatment method or no treatment at all (the placebo group). Random allocation can be done through the use of a computer-generated random sequence; conceptually, the process is like tossing a coin. The purported benefit of randomization relates to minimizing allocation bias and controlling for both known and unknown confounding factors. An example of an unknown confounding

153 Ibid.
154 Ibid. at 259-265.
variable in, say, testing the efficacy of Viagra as a treatment for erectile dysfunction may be the introduction of a sexy new neighbor. The new neighbor, visible to the subject through his bedroom window, may affect a subject’s erectile function, but cannot be controlled for since it is unknown at the time that subjects are assigned to one of the two groups. But since subjects are randomly assigned to both the test and the control group, the idea is that there is no reason for there to be a disproportionate number of sexy new neighbors introduced into the lives of the subjects of any one particular group.

Randomized controlled experiments (RCTs) are generally performed “double blind”; that is, neither the experimenters nor the test subjects know who has been assigned to the control group and who to the test group. It is, however, possible to have a randomized controlled experiment where you randomly assign subjects to the test group and the control group, but where the subjects or the experimenters know which is which. One could also have a double blind study which is neither randomized nor controlled. Here, subjects may be assigned to the control group or the test group based on non-random factors such as one clinic’s preference for a particular treatment over another, but where the patients do not know the details of the treatment they are receiving and the person collecting the data for the purposes of the study is not the treating physician.

The example of the lady tasting tea, which became the title of David Salsburg’s book about R.A. Fisher, shows Fisher’s ideas on randomization at work. The lady in question, one Muriel Bristol, claimed to have a very discerning palette. In particular, she asserted that she could correctly ascertain whether it was the tea or milk that had first been added to a cup. Skeptical of this claim, Fisher devised the following randomized test: he gave the lady eight cups of beverage, four of each variety (milk or tea added first), in random order. As the story goes, the woman correctly identified the pouring order for all eight cups. The chance of someone simply guessing all eight cups correctly is 1 in 70 (there are 70 possible combinations of 8 taken 4 at a time).\(^{155}\)

Fisher viewed statistics as the application of mathematics to observational data and as the study of populations, of variation, and of methods of the reduction of data. Fisher valued the repetition

of important experiments and of observations as necessary given a proper understanding of probability: what we are studying is always a population of possibilities and a good study is representative of this population. Fisher began his defence of randomized assignments in the context of agricultural field experiments undertaken in 1926. He based his insistence on randomization as part of good study design on the need to control or eliminate systematic error or bias. In the agricultural context, such bias might arise in the following manner. If the variable in question is the efficacy of two kinds of fertilizer, testing one fertilizer on one wheat field and another on another field would lead to inconclusive results in relation to the fertilizers’ comparative efficacy. The results would be inconclusive because of possibly confounding factors in relation to the two fields: the plots themselves might have different inherent soil fertility, or the farmers may be using different concentrations of the two fertilizers under investigation. So long as there are factors other than the fertilizer type that could explain differences in results, differing results from one wheat field to the next cannot conclusively be attributed to one factor versus another.

Fisher viewed randomization as salutary here in controlling systematic bias in experimental design: randomization would convert unavoidable systematic errors into random errors and would also reduce the prospect of the experimenter’s involvement with the implementation of the experiment contributing to bias.

It is quite generally accepted that RCTs carry special weight. As we will see in the sections that follow, systematic reviews of RCTs rank highest on the hierarchy of evidence upon which evidence-based practitioners are to rely.

The majority view among the so-called “Bayesian” theorists in this field though is that randomization does not have an essential role in good scientific experimental design. This

---


means that like eyewitness testimony in the legal arena, there is a danger inherent in fetishizing one type of evidence or method for getting at the truth. What the discussion to come will show is that no matter what the domain, there is no substitute for a basic understanding of the nature of the evidence under consideration. This means that while RCTs have much to commend them and meta-analyses of RCTs in particular may constitute the best available evidence much of the time, this isn’t always the case and an understanding of the basic principles involved is necessary to assess when and to what extent one should rely on a particular type of evidence.

The debate in medicine about randomization is also instructive because of its focus on the importance of prior probabilities and the perils of base rate neglect. Those who overvalue the benefits of randomization may do so because they misjudge its capacity to offset the effect of powerful prior probabilities. What is also interesting to note is that while in law, properly accounting for base rates is often considered contrary to common sense, medical theory presents the contemplation of priors as intuitive when explained. It is this view that I hope to encourage.

Colin Howson and Peter Urbach argue that randomization is not necessarily an essential feature of clinical trials and should be acknowledged not to neutralize all possible prognostic factors. The implications of this can be significant in that there may be some circumstances in which non-randomized clinical trials can be used instead to yield useful results. Randomization can be unethical and impractical; for example, if a treatment is thought to be helpful, withholding it from a large number of subjects in the control group for a long period of time is both undesirable and potentially unethical. And, in the case of life-threatening diseases such as heart-attack and stroke, it is not possible to create a “control” group that receives no medical intervention.

In this, medicine and law share some of the same limitations. Even if the law were to have an independent test of veracity in the way that medicine has (the patient gets better or worse, lives or dies), it would be neither practical nor ethical to test an array of decision making processes in this way. That is, we would find it unacceptable to subject some accused to a process that was believed in advance to generate less fair outcomes than another. The principle of clinical

---

159 Ibid.
equipoise, namely genuine uncertainty in the medical community about the preferred treatment, is thought to justify randomized trials. Some, however, argue that equipoise is insufficient to justify RCTs where there is still a prevailing preference for one treatment method over another.

Instead of randomized testing, Howson and Urbach argue that what they refer to as a “Bayesian approach” is more instructive in addressing nuisance variables.¹⁶⁰ More specifically, they suggest that, “the chief concern when designing a clinical trial should be to make it unlikely that the experimental groups differ on factors that are likely to affect the outcome.”¹⁶¹ This approach is “Bayesian” since prior knowledge is explicitly acknowledged and incorporated into the experimental design. That is, in order to know what factors are “likely to affect the outcome”, one must presume prior knowledge of the topic of study. According to the Bayesian, any attempt to treat scientific experimentation as proceeding from a tabula rasa and free from prior knowledge is inaccurate and conceals the prior assumptions that are necessarily in play. The Bayesian explicitly brings to bears upon the experimental design a prior knowledge of which factors may affect results. Having acknowledged these factors, they can be controlled for directly. For example, if blood type is thought to effect a drug’s efficacy, those with different blood types can be evenly and (non-randomly) distributed in each of the groups. In this way, Howson and Urbach suggest that randomization is not absolutely necessary for the construction of clinical trials nor is it always the best way to respond to what they call the “chief concern” that confounding factors will invalidate a study’s results. One practical implication of this view is that trials that use a historical control group, namely, a group of patients who have already received an alternative therapy or no therapy at all, can, in their view, form the basis for informed medical decision making.

Howson and Urbach contend that the Bayesian approach supplies a consistent theoretical framework for scientific practice in general, and medicine in particular, and that its subjective element, while offensive to some, is entirely realistic. Scientists of comparable intelligence and with access to the same evidence at any given time do hold different beliefs and evaluate theories

¹⁶⁰ Supra note 152 at 203.
¹⁶¹ Supra note 152 at 259.
differently. With this given, they contend that it is unrealistic to deny the subjective element of medical study design.\footnote{162}{Supra note 152 at 259-62.}

John Worrall also takes the position that RCTs possess no special epistemic power. That is, it is his view that while randomization does not do any epistemological harm, it also does little good in the context of medical studies. Instead, “known” factors, that is, factors that prior knowledge tells us may affect the results in question, should be balanced between two trial groups, either initially in the trial design, or after the assignment of subjects to the two groups. While randomization may also control for “selection bias” by concealing from the experimenter which subjects are assigned to which group, the known confounding variable of selection bias can also be controlled for directly. And since randomization can do practical harm for those who in the name of generating reliable evidence are denied treatments believed to be beneficial, the practical consequences of this view are significant.\footnote{163}{Supra note 158.}

Specifically, Worrall addresses the claim that randomization makes it probable that all possible confounding factors have been controlled for. According to Worrall, the only way to make sense of this claim is if we were to take the same group of subjects and randomly assign them to each of the two groups repeatedly. In this way, any confounding variable would in the long run be balanced between the two groups. But this, of course, is not how RCTs work. Instead, the random assignment of subjects to each group occurs only once and so provides no guarantee that unknown confounding factors are not in play. Random distributions do not ensure that relevant factors are evenly distributed between the two groups. For example, a random binary sequence generated by a program based on atmospheric noise may look like this: 1, 1, 2, 2, 2, 2, 2, 2, 2, 1 (Heads = 1; Tails = 2).\footnote{164}{Computer-generated random sequences are considered only pseudo-random since they are based on algorithms that are predictable in some respect. More truly random sequences are generated by atmospheric noise. This sequence was generated by the following online random generator: http://www.random.org/coins/?num=10&cur=60-cad.0001c.} If the trait of being an even number is relevant to the phenomenon under observation, this random assignment of subjects will clearly fail to control for the trait of evenness. You will have far more even numbers in this group than odd numbers. Random
distributions are characterized by clustering (in this instance, a clustering of 2s), and to randomly distribute does not spare us from the possibility that a relevant factor of which we are unaware has clustered in one group rather than the other. What’s more, people tend to dramatically underestimate the degree of clustering that is common in random distributions.

As Papineau puts it,

Suppose we notice, after conducting a randomized experiment, a relevant difference between the treatment and control samples. For example, suppose that we notice that the experimental subjects who received treatment were on average much younger than those who did not. Common sense tells us that we shouldn’t then take a difference in recovery rates to show that the treatment is efficacious. But advocates of randomized experiments, like myself, seem to be in danger of denying this obvious truth, since we claim that randomization is a sure-fire guide to causal conclusions.\(^\text{165}\)

Papineau goes on to suggest that this objection needn’t undermine his defence of randomization. It is hard, however, to see how this is the case. That is, if the observation of a clustering of known confounding variables can occur in a randomized trial, then surely such clustering is just as likely to occur in relation to unknown confounding variables. It is possible to argue then on this basis that since randomization doesn’t spare us from the effect of confounding variables, we are better off facing this fact head on by controlling directly for the factors we know are confounding – that is, assigning an equal number of 2s and 1s to each group where evenness is known to affect the phenomenon under observation – and accepting that where unknown confounding variables are concerned, we simply have to accept that this is a potential weakness in any result unless we are willing to perform repeated randomized trials. In a sense Papineau’s argument presupposes a significant prior probability of eliminating confounding factors through

a randomized trial and fails to attend to the prior probability of specific known confounding factors affecting the results.

In this way, Worrall argues that a proper understanding of probabilistic principles reveals that random assignment only makes the effect of confounding variables unlikely *in the long run* and not where the random assignment is, as is the case with RCTs, done only once. He concludes that there is no practical reason for randomizing or automatically giving special weight to the results of trials that have been randomized.

All this is to say that randomization is no panacea for the ills of confounding variables. In this, RCTs share something in common with eyewitness testimony. That is, both can constitute valuable evidence in favour of a particular proposition. However, there is a real risk inherent in overestimating their power. In order to properly assess the significance of both types of evidence, one must understand the nature of the evidence and its potential shortcomings. Both RCTs and eyewitness testimony are often considered the “gold standard” in their respective domains. If “gold standard” is defined as “the ideal” or as “definitive”, then this is a cause for concern. There is no substitute for actual understanding and no type of evidence in the face of which we can abandon our powers of discerning. It is often said that a little knowledge can be a dangerous thing. To make a decision solely in reliance on a *type* of evidence that is believed reliable is never wise.

It is important to note here that while a distinction is made in this debate between “Bayesians” and “non-Bayesians”, essentially all agree in the importance of prior probabilities. That is, even the “non-Bayesians” in the context of the medical study debate accept the basic point that I advocate be similarly accepted in the law: namely, that prior probabilities must not be ignored. In the medical context, the debate has moved beyond this point to a discussion about the relative merits of different ways of addressing prior probabilities: through randomization versus direct controlling of known confounding factors.

Furthermore, the distinction between the so-called Bayesians and non-Bayesians in the medical context is largely overblown. That is, while the assertion that randomization provides no real assurance that the results of any given trial are unaffected by unknown confounding variables is compelling, and likely agreed to by most, to go on from this to say that RCTs have no special epistemic weight seems to downplay the fact that the medical community is generally guided by
systematic reviews of *multiple* RCTs. In so far as this is so, randomization would, in theory, minimize the role of unknown confounding variables in a way that only controlling for known confounders would not. Unknown confounding variables would still be an issue, since we still wouldn’t have results of the indefinite long run and we wouldn’t be repeating randomization with the same study group repeatedly, but it would make sense to think that the effects of many unknown confounding variables would be significantly minimized as the pool of RCTs analyzed grew. So while Howson and Urbach’s point that the medical community may have gone too far in idealizing RCTs is likely correct in principle, in practice, it is meta-analyses of RCTs that are generally used to guide decision making and these may well constitute the best available evidence for most cases. This given, the distinction between the two camps can be seen to have a fairly narrow focus. For practical purposes, there is a very large area of agreement.

In addition, it is important to note that in contemplating medical study design, we can see Bayesianism as “an articulation of commonsense”. That is, when presented clearly, the importance of prior probabilities is apparent. In medical study design, as I have argued is the case in presenting quantitative data to legal decision makers, Bayesianism and intuition can be brought into line and can serve as a useful guide to engaging with available evidence in a critical and meaningful way.

5.2.2 Theories about the Practice of Medicine

In addition to theoretical debates about how best to design medical research studies, there is also a theoretical debate about how best to approach the practice of medicine. The relatively recent advent of evidence-based medicine (EBM) and the surrounding debates have brought epistemological concerns to the fore.

The criticisms of EBM tend to focus on the dangers of misusing evidence. In this, the criticisms mirror concerns in the legal context about the misunderstanding and misuse of quantitative evidence by legal decision makers.

166 *Supra* note 158 at 484.
5.2.2.1 Evidence-based Medicine (EBM)

The buzz word in theorizing about medical practice these days is “evidence-based medicine” (EBM). The EBM movement began in the early 1990s and while not without its detractors, has grown in popularity ever since. It was in the early 1990s that critical appraisal became fairly widely acknowledged as an insufficient basis for making decisions about individual patients. Instead, clinicians were encouraged to view the presentation of a new patient with a new problem as a prompt for investigation into the best available evidence relevant to the particular patient’s condition or complaint.

Evidence-based medicine is about the explicit and conscientious use of the best available evidence to make clinical medical decisions. The EBM approach requires physicians to approach the treatment of patients in the following way: first, they must frame a question relating to an individual’s medical care in terms that are answerable by medical research. Next, they must efficiently search for the best evidence. Third, they must assess the evidence they have gathered in terms of its validity and usefulness in answering their question. And lastly, they must use the evidence and their assessment of it to arrive at a decision regarding an individual patient’s care.

A Health-Canada synthesis report entitled “Creating a Culture of Evidence-Based Decision Making” describes evidence-based decision making in this way:

Evidence-based decision making is the systematic application of the best available evidence to the evaluation of options and to decision making in clinical, management and policy settings.

Evidence-based medicine has received increasing attention over the last decade, as priority has been given to bringing the best available clinical evidence to the attention of practitioners, particularly


physicians. But individuals, administrators and policy makers equally need the best available evidence on which to base their decisions regarding health and health care.

Best available evidence can be obtained from diverse and extensive sources. Multi-disciplinary research brings methodologies and results from different disciplines with different limitations. Although randomized clinical trials, described as the "gold standard" of evidence, are important sources of information, they too have limitations. Other methodologies can also contribute sound and important sources of information. The key is for the decision maker to understand the limitations of the evidence at hand, and the impact and relevance it will have on decision outcomes.

It is also helpful to understand what evidence-based decision making is not. It is not tyranny over providers; it is not value free; it is not a suggestion that evidence is not being used now; it is not a methodological strait-jacket; and it is not an excuse for inaction. Nor is evidence-based decision making based solely on evidence. It is influenced by individual values, interests and judgments as well as external pressures and conditions. It is simply getting the best information in place so that people can make the best decision which is consistent with their values and circumstances.

EBM is touted as systemizing medicine and de-emphasizing uninformed intuitions:

A new paradigm for medical practice is emerging. Evidence-based medicine de-emphasizes intuition, unsystematic clinical experience, and pathophysiologic rationale as sufficient grounds for clinical decision making and stresses the examination of evidence from clinical research. Evidence-based medicine requires new skills of the
physician, including efficient literature searching and the application of formal rules of evidence evaluation the clinical literature.\textsuperscript{169}

EBM claims to provide the “current best evidence to decisions on the care of individual patients”.\textsuperscript{170} EBM relies on quality reviews of studies that assess the evidence on the basis of “good science”. This assessment is formalized through a ranking system that is widely accepted and was developed by the Oxford Centre of Evidence-based Medicine.\textsuperscript{171}

The ranking system puts forward a hierarchy of evidence. At the top of the hierarchy are systematic reviews of randomized controlled clinical trials that are free of worrisome variations (heterogeneity) in the directions and degrees of results between individual studies. Second in line are individual randomized controlled clinical trials with narrow ranges around the study’s results within which the true values are expected to lie (i.e. narrow confidence intervals). Towards the bottom of the hierarchy are case series and poor quality cohort studies. A poor quality cohort study is one that fails to clearly define comparison groups, to measure exposures and outcomes in the same (preferably blinded), objective way in both exposed and non-exposed individuals, fails to identify or appropriately control known confounders, and/or fails to carry out a sufficiently long and complete follow-up of patients. At the bottom of the hierarchy is expert opinion without explicit critical appraisal, or based on physiology, bench research or “first principles”.\textsuperscript{172}

With the assistance of an algorithm that takes into account the quality of available research data, clinicians are advised to search for highly-ranked systematic reviews and if these or other sources of evidence are unavailable, to try to locate case series, expert opinion without formal quality-assessed recommendations, and after that, case reports.


\textsuperscript{170} Centre for Evidence Based Medicine, University of Oxford, online: http://www.cebm.net/index.aspx?o=1025.

\textsuperscript{171} Ibid. See also, David Sackett, “Rules of evidence and clinical recommendations for the management of patients” (1993) 9 Canadian Journal of Cardiology 487.

\textsuperscript{172} Supra note 170.
Physicians are able to access systematic reviews of RCTs through the online Cochrane library. Cochrane reviews are systematic reviews of primary research that investigate the effects of interventions for prevention, treatment and rehabilitation in the area of human health. All Cochrane reviews assess the accuracy of a diagnostic test for a given condition in a specific patient group and setting. Each review must address a clearly formulated question such as “can orthotics help in alleviating foot and leg pain”. A Cochrane review will canvass and assess existing primary research relating to the question that meets certain criteria. Specific guidelines are then applied to assess whether the existing primary research yields conclusive evidence about a specific treatment. The reviews are updated regularly to assist physicians in making decisions based on up-to-date and reliable information:

A Cochrane Review is a scientific investigation in itself, with a pre-planned Methods section and an assembly of original studies (predominantly randomised controlled trials and clinical controlled trials, but also sometimes, non-randomised observational studies) as their ‘subjects’. The results of these multiple primary investigations are synthesized by using strategies that limit bias and random error. These strategies include a comprehensive search of all potentially relevant studies and the use of explicit, reproducible criteria in the selection of studies for review. Primary research designs and study characteristics are appraised, data are synthesized, and results are interpreted.173

There are over 4,000 Cochrane reviews currently available in the Cochrane library. As many as 2,000 protocols for Cochrane reviews are also available, providing an explicit description of the research methods and objectives for Cochrane reviews in progress.

To some, it may look as though the EBM approach runs contrary to the approach I’m advocating. That is, individual physicians are urged not to give undue weight to their prior probabilities when considering the answer to any given medical question, but instead to rely on the “new”

173 The Cochrane Collaboration, online: http://www.cochrane.org/.
information they retrieve in their research. To view the situation in this way is, however, to misunderstand. In many ways, the EBM approach can be understood as fundamentally Bayesian. Medical guidelines are based on assessments of large numbers of studies, and with each new study result, the state of medical knowledge is updated. Given the scope of the updating process that has occurred in arriving at a conclusion of a Cochrane review, a physician has to put a very high degree of weight upon the “new” evidence she discovers when searching for an answer to a question found in a review. The model for her decision making process in this case should look like multiple instances of sequential updating since the new information is not based on a single study but on a compilation of 20 or 30 studies. After updating 20 or 30 times, the physician’s initial prior probability is quickly dwarfed in significance. So the way in which EBM appropriately honours base rates is in the central reservoir of medical knowledge, not in the individual physician. That is, in the face of 30 studies concluding x, one new study concluding y will not change the ultimate recommendation of the Cochrane review. It is in this way that base rate neglect is properly avoided in the EBM approach.

5.2.2.2 Criticisms of EBM

Evidence-based medicine has garnered a great deal of debate. There is the rather minor complaint that the use of evidence in medicine is hardly a new concept – “why all the fuss about EBM?” More serious objections are that EBM misleadingly prioritizes RCTs over experience, basic clinical skills and patient preferences.

Critics also chide supporters for their staunch advocacy in the absence of true clarity about what evidence-based medicine truly entails. The juxtaposition of these two statements of Reilly amusingly illustrates the point: after referring to “the lack of consensus and clarity about what EBM is”, he proceeds to remark that “anyone in medicine today who does not believe it is in the wrong business”!


While it may be fairly uncontroversial that medical decisions should be based on the best available evidence, EBM goes further than this to articulate a general hierarchy of evidence. It is this hierarchy that is most open to criticism. The hierarchy is formed according to the design of the studies providing the evidence in question.\textsuperscript{176}

EBM guidebooks tend to suggest that RCTs always provide superior evidence:

\begin{quote}
If the study wasn’t randomized, we suggest that you stop reading it and go on to the next article in your search. (Note: We can begin to rapidly critically appraise articles by scanning the abstract to determine if the study is randomized; if it isn’t we can bin it.) Only if you can’t find any randomized trials should you go back to it.\textsuperscript{177}
\end{quote}

In so far as this is the message transmitted by proponents of EBM, the concerns about RCTs mentioned in the section on trial design above, apply. That is, while randomization can address selection bias, it is not the only way to do so, and a single randomized trial does not protect against the ills of unknown confounding factors. So, the instruction to “bin” non-randomized studies and move on may well be worthy of rebuke.

Specifically, the categorical direction to base decisions on the results of randomized trials and meta-analyses is too simple. Critics have suggested that for EBM’s hierarchy to be justified, it must be limited to therapeutic decisions, and not applied either to the basic sciences or even to medical decisions such as those relating to prognosis or unsuspected side effects of drugs.\textsuperscript{178}

It is also well to remember that without sufficient understanding, evidence can be misused to yield absurd results. While this may form the basis of a criticism of EBM, it can also be used to support education focused on the meaningful interpretation of results. Just as statistical evidence

\textsuperscript{176} A La Caze, “Evidence based medicine can’t be…” (2008) 22:4 Social Epistemology 3553.


\textsuperscript{178} \textit{Supra} note 176.
is prone to abuse in law, so too can it yield absurdities in the medical context. But it is not necessary to conclude from this that such evidence should be eschewed. Instead, increased awareness of potential pitfalls would appear advisable.

One colourful example of the potential for misunderstanding is found in an article by Leonard Leibocivi on the effect of retroactive intercessory prayer. In 2001, Leibocivi reported on a study of the records of 3393 patients who developed blood infections at the Rabin Medical Centre from 1990-1996. Prayers were performed after patients left the hospital and features of the patients’ hospital stay were examined. Leibocivi found that mortality, length of stay in the hospital and duration of fever were all significantly improved by prayer after the hospital stay. In conclusion, it was stated that “remote, retroactive intercessory prayer was associated with a shorter stay in hospital and a shorter duration of fever in patients with a bloodstream infection.”

Leibocivi’s study, a satire on the believed power of prayer, is also an illustration of the dangers inherent in blindly interpreting data. His parody is in fact a compelling illustration of the perils of ignoring prior probabilities. Prior probabilities help us to distinguish causes from correlations and to ignore irrelevant associations that would otherwise be suggested by raw data.

Admittedly, there are many who would place a high prior probability on the power of prayer. Even those individuals, however, would not likely have a high degree of belief in the power of prayer to act retrospectively. There is, of course, a sense in which differences in prior probabilities pose a real problem. This is in fact, as discussed in earlier chapters, one of the main criticisms leveled against Bayesianism as a normative model for truth. Descriptively, however, it is actually a point in its favour: it supplies a plausible explanation for the continuing existence of fundamental discord.

The absurdity of Leibocivi’s “results” is also in part explained by our natural tendency to underestimate clustering in random distributions. Contrary to our initial intuitions, random distributions are typically characterized by small “streaks” or “clusters”. The so-called clustering illusion refers to the tendency to mistakenly attribute significance to small clusters or

---

to overestimate the uniformity of random distributions. That is, people erroneously expect to find global phenomena present at the local level: for example, they are apt to judge a sequence of 20 coin tosses that features 4 head tosses in a row as less likely to be the product of randomness than a 20-toss sequence of alternating heads and tails. Similarly, Thomas Gilovich found that most people judge that the sequence, “OXXXOXOXXXOXXOOXXOO” is non-random, when it is actually highly characteristic of a random sequence; it features an equal number of “X”s and “O”s and an equal number of adjacent results with the same outcome for both possible outcomes. Sportscasters sometimes refer to a successful shooting streak among basketball players as having a “hot hand”. The idea here is that a successful shooting streak is predictive of subsequent successful shoots within the same game. A study by Thomas Gilovich, Robert Vallone and Amos Tversky found that there was in fact no such phenomenon. People’s intuitive sense of randomness seems to diverge considerably from the laws of chance. That is, people erroneously expect “global” characteristics of chance to be represented “locally”. So if it is a “global” characteristic of a random sequence of heads and tails that heads and tails appear with equal frequency, they expect a short sequence of tosses to reflect the same even distribution. This attribution of global characteristics to local phenomena leads to the so-called gambler’s fallacy whereby it is believed that a losing streak is bound to be followed by a win and similarly causes people to judge a truly random sequence as non-random where any “streak” occurs. Since 4 straight tosses of heads in a row in a sequence of 10 tosses (an actually highly likely outcome in a random series of 10 tosses) is not representative of global randomness, people are quick to mistakenly judge the streak as non-random. It is this same representativeness bias that leads to the impression of a “hot hand” phenomenon. That is, people tend to mistakenly perceive what is actually part of a perfectly random sequence as extraordinary because they expect random sequences to contain far more frequently alternating results (shoot, miss, shoot, miss, shoot) than actually occur by chance.

This tendency to underestimate the likelihood of “clusters” in random distributions makes “findings” such as the one reported in Leibocivi’s study possible. In any group of patients admitted to a hospital, some of them will recover more quickly than others. There are almost an unlimited number of possible explanations for why one patient, or group of patients, recovers more quickly. It is also possible to find correlations between the group that recovers more quickly and a high number of irrelevant features of the individuals in this group: the colour sweater they wore when admitted, whether their name began with a letter from the beginning or end of the alphabet, whether they had birthdays in the early or latter part of the year, and so on. Given the almost unlimited number of such features, it will almost always be possible to find a correlation between some irrelevant feature and quick recovery.

These examples help show the ills caused by base rate neglect. You ignore what you already know at your peril. The hypothesis under investigation must have some initial plausibility before results are analyzed. This “initial plausibility” is informed by our starting knowledge about the situation under inspection. And it is the lack of “initial plausibility” or any prior probability of a causal relationship between retroactive intercessory prayer and recovery time that allows us to judge Leibocivi’s results as an artifact or a fluke.

In a similar vein, the article entitled “Why Most Published Research Findings are Wrong” reports that a significant number of highly cited clinical trials are not replicated and that the likelihood that the results of such trials are false increases with the improbability of the hypotheses tested.

Far from serving as a critique of evidence-based medicine, the possibility of studies yielding whacky results illustrates the importance of prior probabilities in the interpretation of such results. Study results that support hypotheses with a sound scientific basis are more likely to be true than those with no such prior support. Evidence-based medicine is successful because it employs prior probabilities by favouring meta-studies of randomized controlled trials that do not allow the results of any one trial to guide medical decision making. That is, with the accumulation of each new result from a study that meets certain baseline criteria, the state of our

---

best knowledge is updated. It is this essentially Bayesian character of meta studies that makes
them so much more reliable than the results of a single study and accounts for many of modern
medicine’s successes.

In this way, both law and medicine would benefit from educating its participants about the role
of prior probabilities and the proper interpretation of common types of evidence. While the law
has chosen thus far to view the risks inherent in the interpretation of statistical evidence as reason
for avoidance, at least at the level of theoretical debates, medicine provides an example of an
alternative approach.

5.2.3 Actual Medical Practice

5.2.3.1 A Model of Clinical Medical Reasoning

The preceding discussion has focused on debates at the level of theory about both medical
research and medical practice. In addition, there is the issue of how physicians conduct
themselves day in and day out. Bayes’ theorem is often cited in medical textbooks as a useful
decision-making tool. At the same time, evidence suggests that physicians pay little if any
attention to the meaning of probability and do not distinguish between circumstances of
ignorance and of under-determination. Discussion of the origin of prior probabilities is notably
absent from many clinical epidemiology texts.184

Proponents of EBM suggest that clinicians’ knowledge of probabilities should come from
published scientific literature. Those unsympathetic to EBM tend to argue that, to the contrary,
there is an unstated dimension to physicians’ knowledge and that medicine is an art and not just a
science. It may be possible to reconcile these views to a certain extent by viewing medical
diagnosis as a two-step process: the first stage involves judgment that is honed by experience
and allows a physician to identify most likely causes of disease; the second step calls upon the
physician to refer to the most up-to-date knowledge on the area now identified based on meta-
reviews of randomized trials.

5.2.3.2 Teaching Doctors to be Better Bayesians

The sections above discuss some of the ways in which medical theory – at the level of research, guideline creation, and the application of guidelines to practice – is still developing. What is especially clear is that even if particular theoretical commitments, for example a commitment to Bayesianism and an evidence-based approach, are taken as givens, practicing clinicians are prone to reasoning errors when faced with statistical evidence in the same way as are judges and juries.

A look at common approaches to teaching physicians how to be better Bayesian reasoners can be informative in guiding educational approaches that could benefit legal decision making when it comes to integrating statistical evidence.

Currently, graduate medical curricula are characterized by journal clubs or similar so-called “critical appraisal” approaches to new evidence. Journal clubs involve a small group of medical students, led by a faculty member, who get together to discuss recent articles reporting new medical discoveries. The research is evaluated on the basis of such factors as timeliness and relevance, and faculty members often highlight or clarify relevant points and put the research into clinical context. While such critical appraisal is not without its value, it fails to foster a number of skills essential to the effective practice of evidence-based medicine. That is, skills relating to the design of focused questions, tracking down the best evidence and then assessing the evidence and applying it to patient care, are not developed by the critical appraisal approach.

Standard medical epidemiology texts, such as David Sackett et al.’s *Clinical Epidemiology: A Basic Science for Clinical Medicine* 185 counsel physicians to approach, for example, a prognosis of cardiomegaly in the following way: first, consult an expert who may lead you to a recent study showing that enalapril saves lives in congestive heart failure. As a next step, you might consult your personal library, followed by an electronic literature search. In terms of reading reviews, overview, and meta-analyses, clinicians are advised to ask the following questions for the purposes of assessment: (i) were the questions and methods clearly stated?; (ii) were the search methods used to locate relevant studies comprehensive?; (iii) were explicit methods used

---

to determine which articles to include in the review?; (iv) was the methodological quality of the primary studies assessed?; (v) were the selection and assessment of the primary studies reproducible and free from bias?; (vi) were differences in individual study results adequately explained?; (vii) were the results of the primary studies combined appropriately?; and (viii) were the reviewers’ conclusions supported by the data cited?\textsuperscript{186}

Where cost-effectiveness of health care is in issue, these guiding questions are recommended: (i) was the economic question properly posed?; (ii) were the alternative programs adequately described?; (iii) has the program’s effectiveness been validated?; (iv) were all important and relevant costs and effects identified?; (v) were credible measures for these costs and effects selected?; (vi) was an appropriate analysis carried out?; (vii) were comparisons between programs properly adjusted for time?; (viii) were the presence and magnitude of bias identified?\textsuperscript{187}

While the pedagogical tactic described above is still most typical, specific EBM-based components of curricula are becoming more common, and are advocated in the literature on medical pedagogy.\textsuperscript{188} EBM training involves providing medical students with access to on-site electronic resources, using secondary evidence-based summaries, tracking students’ EBM behaviours and using situation-specific EBM techniques. Training clinicians in EBM requires training in articulating clinical questions, searching for information, appraising the evidence found, and then applying the evidence to the individual clinical scenario.\textsuperscript{189}

In a recent edition of *Sapira’s Art and Science of Bedside Diagnosis*\textsuperscript{190}, alongside chapters on the order in which to conduct an examination, how to handle a patient who does not hear well,
and appropriate bedside manner, there is a section devoted to the evaluation of diagnostic signs that focuses on the proper interpretation of quantitative data. The text counsels physicians as follows: “…the physician who is not overawed by the mystique of objectivity that emanates from data expressed in three or more significant figures will not be betrayed by uncritical dependence on them. Moreover, if one understands certain definitions, one can avoid some common errors of inference”. This is a crucial idea and one to which I will return in the next chapter. That is, for physicians and legal decision makers alike, a little knowledge is necessary to avoid both overawe and some common errors of inference. The text then proceeds to define terms such as “prevalence”, “positive predictive value”, “specificity” and “sensitivity”. An explanation of how a test with both high sensitivity and specificity can yield a result of low positive predictive value in the face of a low prevalence in the population (i.e. a low prior probability) is provided:

In the example given above (a test of 100% sensitivity and 90% specificity applied to a population with 9% prevalence), the positive predictive value was so low that you would be slightly better off guessing that a patient with a positive test did not have the disease, a conclusion that may seem to be contrary to common sense. However, a positive test raises the probability of disease from 9% (prior probability) to 49.7% (posterior probability, i.e. the probability after the test). Unfortunately, this is the kind of result that makes skeptics laugh and exclaim that the test is about as good as flipping a coin. This last statement is not true because the majority of patients without disease will have a negative test. The 49.7% is a conditional probability: If the result is positive, the probability of disease is 49.7%. Furthermore, there is a chain or reasoning in clinical problem solving. The figure of 49.7% may now become the input (the prior probability) for the next test ordered [on the same patient]. Sequential positive tests with a positive predictive value of about 50% would make the sequential probabilities 50%, 75%, 87.5%, 93.75%, and so forth, tending asymptotically toward 100%. Because questions can be asked quickly and also because physical maneuvers can

---

be done quickly in sequence, many skillful physicians can thus come to a rapid diagnosis, even though the individual maneuvers may be likened to flipping a coin. There is an even more important consideration. In the example above, the predictive value of a negative test or negative predictive value is excellent. In fact, it is an incredible 100% (because of the 100% sensitivity).\(^\text{192}\)

The passage above begins by stating that the low positive predictive value given a test with such high sensitivity and specificity “may seem contrary to common sense”. But, when the result is reframed in terms of an increase in likelihood of the disease after receipt of a positive result, i.e. from 9% to 49.7%, the result looks considerably less surprising. And when it is explained that sequential tests with only a 50% positive predictive value quickly yield very high positive predictive values, intuition and probability realign.

If this trend continues, the next generation of clinicians will be better able to search and critique medical literature when addressing problems arising out of their clinical practice. Both online and paper journals such as ACP Journal Club and Evidence-Based Medicine are increasingly available to clinicians, as are access to online journal collections such as Medline. In addition, physicians already have access to summarized evidence sources such as clinical practice guidelines, systematic review and the Cochrane Collaboration.\(^\text{193}\)

However, while proponents of EBM make bold claims such as that, “it is absurd to suggest that the use of evidence could be detrimental”,\(^\text{194}\) this is clearly not the case. A misuse or misinterpretation of evidence by clinicians ill-equipped to understand the evidence with which they are faced could easily lead to worse results than if they proceeded on the basis of expert guidance or clinical experience. Part of the skill of appraising the evidence requires clinicians to have a certain facility with the manipulation of probabilistic data. In this, people are worryingly prone to error.

\(^{192}\) Ibid. at 9-10.


\(^{194}\) Ibid.
It is in part this concern around error that has led the law to shy away from the use of quantitative evidence in the decision-making process. But examples of ways in which medical pedagogues have endeavored to overcome this obstacle may be highly instructive in the legal context as well.

There is a small literature devoted to teaching medical students about the role of prior probabilities in the differential diagnosis process. There is strong evidence that people in general problematically ignore base rates in certain contexts and doctors are not notably exceptional in this regard. 195

Differential diagnosis is often taught as the memorization of lists of possible or likely causes of symptoms with the goal of exhaustiveness. These lists do not necessarily incorporate relative probabilities of the competing causes and if they do, these relative probabilities are not always well-understood.

Two techniques for enhancing clinician understanding of prior probabilities and their use have been put forward. 196 These two techniques are the “probability pie” and “the dollar”. Both represent simple heuristics that present probabilities in a way that is more easily grasped. The effectiveness of these presentation techniques has not itself been studied, but it is fair to assume that their prima facie perceived accessibility is a good indicator.

One of the techniques is the probability pie, which works in the following fashion. Instead of presenting a list of possible diagnoses related to symptoms presented, clinicians are asked to present the possible diagnoses in the form of a pie chart where each diagnosis is weighted according to probability. Along similar lines, the clinician can be told that she has one dollar of probability to spend. The dollar technique in particular reinforces the idea that probabilities range from 0 to 1 and must add up to 1 in total.

The example Upshur uses to illustrate this pedagogical technique is of a man who presents with chest pain. His symptoms could be caused by angina, gastroesophageal reflux disease,


196 Ross Upshur, “Two Techniques for Teaching the Estimation of Prior Probabilities” (200) 12:3 Teaching and Learning in Medicine 141.
musculoskeletal pain or aortic aneurysm. Clinicians are asked either to divide a pie chart up among these possibilities, or “spend” their dollar on the four competing diagnoses. Using the pie chart technique, clinicians could rate angina as 60%, gastroesophageal reflux disease (GERD) as 25%, Aortic Aneurysm as 10% and musculoskeletal (MSK) pain as 5%. The guiding principle that “common things are common” referred to above should figure in assigning these initial probabilities. That is if angina and GERD are very common while aortic aneurysm and MSK pain are relatively rare, this would account for their respective assignments in the pie chart. In this way, the pie chart is used to represent the relevant base-rate statistics. The pie chart presentation would look something like this:

![Pie Chart Example](image)

**Figure 2: Example of a Probability Pie**

The corresponding “dollar” presentation requires physicians to assign 60 cents to angina, 25 cents to GERD and so on.

Different diagnostic tests are then recommended based on these preliminary (base rate) assessments: a stress test for angina, endoscopy for GERD, ultrasound for aortic aneurysm and no diagnostic test for MSK pain. Once the clinician has the results of these diagnostic tests, she can update her diagnostic picture. Use of the pie or dollar tools in this way results in the physician considering base-rate evidence and then “updating” her assessment of the relative
likelihood of each cause under consideration given each new piece of information provided in the form of test results.

The pie and dollar heuristics are particularly helpful in illustrating the way in which prior assessments are properly updated to reflect new information. For example, if the test for aortic aneurysm comes back “positive”, but has a high false positive rate, such as 50%, it is relatively easy to see that ½ of all the non-white area (not aortic aneurysm) is larger than the whole white region (it is aortic aneurysm). This means that in the face of a “positive” aortic aneurysm test result, it is still more likely that one of the other diagnoses applies. That is, even in the face of a positive aortic aneurysm result (assuming that the test has a false positive rate of 50%), in light of the base rate probabilities of the other causes for the condition, it is still considerably more likely that the patient has one of either angina, GERD or MSK and *not* an aortic aneurysm.

The pie and dollar heuristic approaches have the benefit of making prior probabilities explicit. That is, the clinician’s assessment of the relative likelihood of competing diagnoses, before the performance of further tests, are clearly articulated. In this way, the source of the priors can be examined – they should be evidence-based – and new information gathered through additional testing can be properly used to update the assessment. This approach can also help to identify when additional diagnostic testing would be of limited use, such as in a case where the prior probabilities are heavily weighted against a particular diagnosis and the available diagnostic test is of very limited accuracy.

Upshur identifies a number of limitations inherent in such techniques – that the source of prior probability assessments may be unknown and so not easily subject to scrutiny, that likelihood ratios may not be available for tests, that patient complaints often come in groups, and so on. But one of the added benefits of these heuristics, in addition to assisting with making prior probabilities more explicit and helping with the proper approach to updating priors, is that it makes the uncertainty in decision making explicit. This is clearly an advantage in medical practice, but some may argue that it is less desirable in legal decision making where the perception of certainty may strengthen the legal system’s goal that results appear fair and encourage public confidence in “the system”. This line of inquiry will be pursued in the chapter that follows.
Medicine has advanced beyond anecdotal evidence. And while there is still debate at the theoretical level about the best way in which to rank evidence and exactly which constraints are necessary in the development of studies, at the level of theory, prior probabilities are acknowledged. Overreliance on RCTs may be a problem in medicine just as is overreliance on certain kinds of evidence in the legal domain. But this suggests only that some basic level of understanding of nature of the evidence, instead of blind reliance on evidence-type, is always beneficial. And no matter what the area of interest, base rate neglect will always be problematic. Even those who argue against a Bayesian updating approach are themselves presupposing it; that is, they are relying on their own high prior probability belief against Bayesianism in the face of new arguments in its favour to justify retaining their initial anti-Bayesian position.

What I have also tried to show is that physicians, like judges, lawyers and jury members, have difficulty in applying even the most basic of probabilistic principles to their practice. But in medicine, unlike in the law, there is wide-spread acceptance at the level of theory that this is a problem in need of a solution and that a small amount of education can be a great help. In this, it seems promising to suggest that we may be able to learn from the medical profession and adapt some of the pedagogical techniques used in that context to the legal sphere.

The chapter that follows will focus on applying some of the lessons learned in the medical context to the law. These lessons include appreciating the importance of prior probabilities, the “commonsense” nature of these priors, and the helpfulness of particular techniques in incorporating knowledge of priors into the decision-making process. These ideas will be applied to legal decision making to generate suggestions for improvement.
Chapter 6
Lessons to Learn

When you get the dragon out of his cave on to the plain and in the daylight, you can count his teeth and claws, and see just what is his strength. But to get him out is only the first step. The next is either to kill him, or to tame him and make him a useful animal.\textsuperscript{197}

6 Lessons to Learn

In the previous chapter we saw that medicine has advanced beyond anecdotal evidence. While there is still debate at the theoretical level about how to rank evidence and exactly which constraints are necessary in the development of studies, at the level of theory, prior probabilities are acknowledged. In this, it seems promising to suggest that the law may be able to learn from the medical profession and adopt some of the pedagogical techniques used in that context.

I argued in chapter 4 that Bayesianism is not at odds with many prevalent legal intuitions. If it is possible to bring probability and intuition largely in line, then Bayesianism can offer a helpful normative framework for correcting errant intuitions and adopting new kinds of statistical evidence into legal discourse.

In the preceding chapters I have tried to encourage acceptance of a Bayesian approach to evidence as a normative framework for working with statistical evidence – by showing that it aligns with existing foundational legal intuitions and need not generate troubling paradoxes. I have also illustrated that while modern medicine has largely accepted a Bayesian approach at the theoretical level, practitioners are still challenged by errant intuitions and medical education has become alive to this in a way that can guide legal educators.

I hope to have shown that while mathematical formulas formalize and explain how to work with statistical evidence, complex mathematics is not required to conceptually grasp how to reason

\textsuperscript{197} Oliver Wendell Holmes Jr, “The Path of the Law” (1897) 10 Harv L Rev 457 at 469.
properly with probabilities. This is critically important since the law demands the transparency of justice and calls on ordinary citizens to evaluate evidence.

The first task of this chapter is to revisit the problems described in chapter 1 and consider what a Bayesian, and at the same time intuitively accessible, approach entails. This will assist in illustrating the practical application of the Bayesian-intuit approach that I am promoting.

I will discuss some of the implications and applications of the conclusions drawn in previous chapters. That is, having suggested that probabilistic evidence should be embraced, that it is possible in most cases to bring sound reasoning involving probabilities and intuition into line, and that certain modes of presenting evidence – in terms of natural frequencies and pictographs – are preferable, I will now consider what further conclusions may be drawn in relation to some more pressing questions of concern within the world of legal evidence.

Lastly, I make some specific recommendations as to the type of education that could be provided (a) to legal professionals such as judges and lawyers; and (b) to jury members in the form of a jury charge or introductory remarks by an expert presenting statistical evidence in court.

6.1 Methods to Adopt

We saw in chapter 5 that the medical profession generally accepts Bayesianism as a normative model but that as a matter of practice, unsurprisingly, medical physicians make the same kinds of errors when reasoning with statistics as do legal decision makers.

The “probability pie” and dollar approach to understanding probabilities have been recommended by Ross Upshur and I suggest that these techniques would be equally valuable in the context of legal education and the presentation of statistical evidence to juries.\(^\text{198}\) The probability pie technique will be illustrated in the treatment of one of the cases below.

6.1.1 The Power of Modes of Presentation

I already illustrated in a previous chapter the way in which presenting probabilities in terms of natural frequencies can make it far more intuitively accessible. But there are other techniques that require great conceptual clarity on the part of the evidence presenter that can help to render statistical evidence more intuitively accessible. Take the following example: suppose you have a test for determining whether a fetus is a boy or a girl. The probability of concluding that it is a boy \((b)\) given that it is actually a boy \((B)\) is greater than the probability of it concluding girl \((g)\) given that it is a girl \((G)\). If it’s a boy, the test will read “boy” 80% of the time while it will only read that it is a girl 70% of the time that it’s a girl. Which test result should we find more believable? Should we be more confident that it is a boy, given a test result showing that it is a boy, or that it is a girl, given a test result showing that it is a girl?

Presented in this way, the answer appears to be that we should be more confident in boy results. After all, 80% is better than 70% and this test gets boys right 80% of the time. But the correct answer is that we should be more confident in the girl result. You can calculate this from the conjunctive probabilities (with base rate \(p(B) = 0.5\), and \(p(G) = 0.5\)):

\[
\begin{align*}
B \text{ and } b &= 0.4 \\
B \text{ and } g &= 0.1 \\
G \text{ and } b &= 0.15 \\
G \text{ and } g &= 0.35
\end{align*}
\]

Note that we are only able to generate all these conjunctive probabilities from the limited information provided in the question because there are only two options here: boy or girl. It is because of this that we can deduce that the probability of \(B \text{ and } g\) is 0.1 (since we are told that the probability of \(B \text{ and } b\) is 0.4) and that the probability of \(G \text{ and } b\) is 0.15 (since we are told that the probability of \(G \text{ and } g\) is 0.35). Were there other possibilities – chickens, aliens, what have you – we would have insufficient information to answer the question.

Given the conjunctive probabilities listed above, \(p(B|b) = \frac{0.4}{0.4 + 0.15} = 0.73\), while \(p(G|g) = \frac{0.35}{0.35 + 0.1} = 0.78\). Since \(p(B|b) < p(G|g)\), that is, the chance of actually having a boy given
that the test reads “boy” are less than the chances of actually having a girl given a reading of “girl”, you should have more confidence in a test reading of “girl”.199

But this seems counterintuitive and the math takes some skill and effort to sort out. However, if the problem is presented in the following way, rather than in the way it was presented at the outset, the correct result is far more intuitively accessible:

Suppose you have a test for determining whether a fetus is a boy or a girl. Boy and girl are the only two options and there are an equal number of boys and girls conceived. The probability of the test concluding that it is a boy (b) given that it is actually a boy (B) is greater than the probability of it concluding girl (g) given that it is a girl (G). If it’s a boy, the test will read “boy” 80% of the time while it will only read that it is a girl 70% of the time that it’s a girl. At the same time, the chance of a false reading of “boy” is higher than the chance of a false reading of “girl”. We know this because a 70% chance of reading “girl” when it’s a girl (which is what we’re told) means that the other 30% of the time it’s reading “boy” when it’s a girl. Since there’s an 80% chance of getting a “boy” reading with a boy, there’s only a 100% - 80% = 20% chance of getting a false girl reading. So in effect, this test is boy-biased. It reads “boy” more often when there’s a boy, but it also gives more “boy” false positives. When a boy-biased test gives the relatively rare reading of “girl” you can have more confidence in the result.

As you can see, presentation of statistical evidence does matter. And what the above example illustrates is that a clear grasp of the underlying concepts by the person presenting the evidence, as well as a focus on how best to present it so that it is intuitively accessible, can make a significant difference.

199 Robyn M Dawes, Everyday Irrationality (Boulder, CO: Westview, 2001) at 85-86.
6.2 Approaches to Problems from Chapter 1: Addressing Common Fallacies

At this point it would be helpful to revisit the case studies from chapter 1 to see what lessons can be drawn from the discussion that followed their introduction. What would it mean for decision makers to approach the statistical evidence involved in these cases in a way that adopted intuitive conceptual approaches based on sound Bayesian reasoning?

Two of the most common conceptual errors made in reasoning with statistical evidence are as follows:

1. avoiding statistical evidence that is actually highly relevant to a factual issue in dispute; and
2. “failing to think of the opposite” in the consideration of statistical evidence.\(^{200}\)

As a reminder, Bayes’ rule tells us that with the following three pieces of information we can assess the likelihood of a hypothesis (H) given some data (D):

1. the prior probability of the hypothesis/the base rate: \(p(H)\)
2. the probability of the data if the hypothesis is true: \(p(D|H)\)
3. the probability of the data if the hypothesis is not true: \(p(D|\sim H)\)

\[
p(H|D) = \frac{p(H)p(D|H)}{p(H)p(D|H) + p(\sim H)p(D|\sim H)}
\]

\(^{200}\) Given my concession to legal considerations beyond verdict accuracy, these are only “errors” if they stem from ignorance about statistics as opposed to a knowing decision to exclude otherwise relevant evidence for other reasons of policy.
6.2.1 Statistical Evidence: Addressing Problematic Avoidance

6.2.1.1 How Statistical Environmental Evidence Could Have Been Considered

In *Massachusetts v. EPA*, a number of states brought a legal suit against the United States Environmental Protection Agency (the “EPA”) to force the EPA to regulate emissions from automobiles. In order to prove their case, they needed to prove harm. The reader may recall that at each level of court, the question of the background probability of global warming was avoided even though it would be an essential consideration to the question of whether EPA’s inaction in regulating automobile emissions could be proven to have caused harm and if so, how much.

There is vigorous debate about the nature, causes, and consequences of global warming. Debates concern the proportional contributions of causes to the increased global average air temperature and whether the warming trend is unprecedented or within normal climatic variations. Whether or not there is scientific consensus on this issue, the courts should have explicitly acknowledged that the question of EPA’s negligence *could not be assessed* absent a finding on this question.

Had the defence accepted a Bayesian normative model for decision making, they might have made this explicit from the outset as an alternative argument in case they did not succeed on the issue of standing.

Bayes’ rule makes it clear that the probability of EPA’s responsibility given data about increased global average air temperature could not be assessed absent an established probability of increased global average air temperature in the event that EPA had regulated automobile emissions: \( p(\sim H | D) \).

Using the pie chart pedagogical tool, the legal inquiry into whether the EPA caused harm in failing to regulate automobile emissions could have begun with the following presentation of necessary background information. This chart is presented for purposes of illustration only and is not based on known contributions to global warming:

---

201 *Massachusetts v EPA* (2007), 127 S Ct 1438.
Figure 3: Probability Pie – Factors Affecting Global Warming

Using this chart as a baseline for prior knowledge of the relative weight of factors affecting atmospheric warming globally, certain inferences could be drawn about assessing the relative contribution of automobile emissions (a subset of the category of burning of fossil fuels for energy) based on certain data. For example, it would become apparent that a test that purported to distinguish atmospheric warming due to atmospheric aerosols versus the burning of fossil fuels that was 70% accurate would be of very little help; it is easy to see that 30% of the black slice of the pie (which would represent atmospheric aerosols wrongly attributed to the burning of fossil fuels) is bigger than 70% of the white sliver (which would represent the correctly identified attributions to fossil fuels). In this way, it becomes relatively apparent that given these base rates, relying on the results of such a test alone would lead to more false attributions to fossil fuels than correct ones.

The point is simply that had this case not ultimately been dismissed on the preliminary issue of standing, it would have been necessary for the courts to consider statistical data about atmospheric warming. Avoiding such data, as the court was inclined to do in this instance,
would not have been best. Instead, a pie chart presentation of such data would have increased the chances that such evidence would have been properly considered to reach a just result.

6.2.1.2 How Statistical Evidence on Seatbelt Efficacy could have been Considered

In the case of *Beazley v. Suzuki Motor Corporation*, the judge’s rejection of statistical evidence on the effect of seat belt use during car crashes demonstrated a similar failing. The judge was faced with a question of the contributory negligence inherent in the plaintiff’s failure to wear a seatbelt that could only be evaluated against knowledge of the effect of such seatbelts in crashes of a similar nature. That is, the judge would have had to consider the question of the probability (and extent) of harm had a seat belt been worn. In order to assess that hypothesis of no injury (H) given the data that a seat belt was worn (D), we now know that we could use the following information:

1. the likelihood of no injuries of this type in car accidents in general: \( p(H) \)
2. the probability that a seat belt is worn where there are no injuries of this type: \( p(D|H) \)
3. the probability that a seat belt is worn where there *is* injury of this type: \( p(D|\sim H) \)

It may be that my common sense is somehow lacking, but I would have a difficult time estimating any of the three probabilities above drawing solely on my common sense. And this is so even though I could say with the same confidence as did the judge in that case that seatbelts clearly significantly reduce the likelihood of injury (in general) in an accident.

All this is to say that assuming a finding of some negligence on the part of both parties, the judge would have benefited significantly from considering the expert report on the role of seatbelts in reducing injuries (or injuries of a particular kind) when contemplating the extent of contributory negligence and apportioning damages accordingly. That is, in concluding that the knowledge that seat belts prevent injury was sufficient, he seems to have confused knowledge that

---

\[ p(H|D) > p(H|\sim D) \] with knowledge of all the elements necessary to assess the probability that there is no injury where seatbelts are worn:

\[
p(H|D) = \frac{p(H)p(D|H)}{p(H)p(D|H) + p(\sim H)p(D|\sim H)}
\]

As mentioned in chapter 1, the Ontario Court of Appeal confirmed in 2005 that a plaintiff’s recovery for damages can be reduced by up to 25% where s/he failed to wear a seatbelt.\(^{203}\) The question of where, on the spectrum of a 0% to 25% reduction in damages, a particular plaintiff lies, rests on a factual question. Would wearing a seatbelt have made no difference to this kind of injury, have prevented it completely, or somewhere in between? This is a question that is distinct from whether seatbelts generally reduce the risk of injury in car accidents – the piece of information that the trial judge so confidently possessed.

In a case such as this, it may well be argued that for the purpose of limiting time and expense, an expert report on the effects of seatbelt use on injury should not be admitted into evidence. I note that in this particular case, the report already existed and would only need have been reviewed by the judge. There are many cases though in which objections based on judicial economy would have significant weight. I emphasize, however, that my only quarrel is with the rejection of such evidence on the basis that it would not in fact improve decision accuracy.

6.2.2 Remembering to “Think of the Opposite”

When statistical evidence is considered, one error in particular is highly common: the error of failing to “think of the opposite”. The ubiquity of this error is demonstrated by a test used by Doherty and Mynatt (1990) in which subjects were asked to imagine that they were a doctor examining a patient with a red rash. The subjects were shown 4 pieces of evidence and asked to choose which pieces of information they would need in order to diagnose a disease, “Digirosa”.

(i) The percentage of people with Digirosa.
(ii) The percentage of people without Digirosa.
(iii) The percentage of people with Digirosa who have a red rash.
(iv) The percentage of people without Digirosa who have a red rash.

\(^{203}\) See Snushall v Fulsang, [2005] 78 OR (3d) 142.
These pieces of information correspond to \( p(H), p(\neg H), p(D|H) \) and \( p(D|\neg H) \) in Bayes’ rule:

\[
p(H|D) = \frac{p(H)p(D|H)}{p(H)p(D|H) + p(\neg H)p(D|\neg H)}
\]

The only piece of information that you *do not* need from the list is (ii) if you have (i) or vice versa. This is because once you know one of those two you know the other since the percentage of people with and without Digirosa must equal 100. But people frequently make the error of thinking that you also need not know (iv). This is the fallacy of “failing to think of the opposite”.\(^{204}\)

To simply illustrate this fallacy as well as the ease with which it may be avoided, I will assume in analyzing the cases below that the hypothesis at issue is whether the accused were/was at the scene of the crime and that the prior probability of this hypothesis is 0.5; that is that the decision maker goes into the consideration of the evidence at trial with no preconceived notions relating to this factual question. This accords with model instructions to juries that they only consider the facts as presented in coming to their conclusions and not rely on external information:

To decide what the facts are in this case, you must consider only the evidence presented in the courtroom. Evidence is the testimony of witnesses and things produced as exhibits. It may also consist of admissions.

The evidence includes what each witness says in response to questions asked. The questions are not evidence unless the witness agrees that what is asked is correct. Only the answers are evidence.\(^{205}\)


Once we have assumed a value of 0.5 for the prior probability, Bayes’ rule is considerably simplified since $p(H) = p(\sim H)$ and these terms cancel each other out on top and bottom:

$$p(H|D) = \frac{0.5 \times p(D|H)}{0.5 \times p(D|H) + 0.5 \times p(D|\sim H)}$$

$$p(H|D) = \frac{p(D|H)}{p(D|H) + p(D|\sim H)}$$

There are now only 2 terms with which we must concern ourselves: $p(D|H)$ and $p(D|\sim H)$.

6.2.2.1 People v. Collins

In *People v. Collins*\(^{206}\), the reader may recall that the statistician, having identified a number of characteristics as relevant to the couple culpable in a robbery – partly yellow automobile, man with mustache, negro man with beard, girl with ponytail, girl with blond hair, and interracial couple in car – assigned an individual probability to each of these characteristics and applied the product rule. He then concluded that the chance of a couple fitting all the characteristics was 1 in 12 million and equated the probability to the probability of the accused couple’s innocence.

There is a whole host of errors in the way the “expert” presented this evidence, not least of which is that he assumed, without any grounds, that these characteristics were independent. That is, he assumed that if you’re a bearded man you’re not more likely to have a moustache when in fact it would seem that in this direction at least (if you have a beard, then you have a moustache) these two traits are almost always dependent! As discussed in an early chapter, the product rule for working with probabilities only applies where the probabilities are independent. But for the purposes of this discussion, my focus is not on the proposition that the chance of a couple fitting all the characteristics was 1 in 12 million, but rather on the reasoning from that proposition to a finding of guilt.

In particular, it is contrary to Bayes’ rule to equate the probability of finding a certain set of characteristics with the probability of innocence (and then to infer guilt from what has mistakenly been identified as a vanishingly small probability of innocence).

\(^{206}\) *People v Collins*, (1968) 66 Cal Rptr 497.
If you consider the simplified version of Bayes’ rule derived above, you have the following:

\[ p(H|D) = \frac{p(D|H)}{p(D|H) + p(D|\sim H)} \]

In this case, the hypothesis (H) is the hypothesis that the accused were present at the scene of the crime and the data (D) is that the accused’s characteristics match the characteristics observed by the witness.

Even were the 1 in 12 million statistic assumed correct, all that gives us is the probability of finding the data, namely \( p(D) \). The 1 in 12 million statistic also turns on how many people are in the population in total. This statistic is far more instructive in a population of 12 million or less where a 1 in 12 million phenomenon can be expected to occur no more than once; conversely it is a lot less help in a population where 12 million occurs many times over. In the late 1960s, when the *People v. Collins* case was decided, there were somewhere around 60 million married couples in the U.S. If you add couples who were just dating or living together unmarried, this number would be larger still. But, even on the conservative assumption that we were dealing only with the pool of married couples, the expert’s testimony that the chance of finding all the characteristics together was 1 in 12 million would look like this (since 1/12 million x 60 million = 5):

![Figure 4: Probability Pie – Couples with and without All Traits](image-url)
This tells us that we’re looking at a really small number of couples. But it doesn’t tell us anything about the relative likelihood of whether the accused couple were at the scene, i.e. $p(D|H)$ versus $p(D|\sim H)$. In this, relying on the 1 in 12 million statistic alone to draw a conclusion about the probability that the accused were at the scene of the crime is clearly mistaken. The probability of finding the characteristics would only equal the probability that the accused were at the scene if there were no possibility that $\sim H$. Reasoning along these lines commits the fallacy of failing to think of the opposite.

Once it is appreciated that both $p(D|H)$ and $p(D|\sim H)$ must be contemplated, it becomes clear that what is of interest is what is happening within the sliver of the population pie that has the featured traits. And since the population of that sliver of pie is greater than 1, in fact the 1 in 12 million statistic indicates a sliver populated by around 5 couples, the necessary information is the relative likelihood of finding all these characteristics given that this couple did or didn’t do it. If we are assuming 5 couples in the U.S. with all relevant characteristics, then the $\frac{H}{\sim H} = \frac{1}{4}$ (1 couple who did it plus 4 who didn’t = 5 couples in total):

$$p(H|D) = \frac{p(D|H)}{p(D|H) + p(D|\sim H)} = \frac{1}{1 + 4} = \frac{1}{5}$$

But really, we didn’t even need to run the calculation above. What should be intuitively clear is that if you know that there are 5 couples in the population with the same characteristics as the accused couple and all you know is that the accused couple has those characteristics, you know nothing more than that there is a roughly 1/5 chance that you apprehended the right people. And that’s clearly not enough to ground a criminal conviction.\(^{207}\) The size of the population as a whole is of utmost importance here. The phrase “you’re 1 in a million” may seem meaningful when uttered out of love, but in a population that exceeds 6 billion, this is actually rather faint praise. The statement that “there are only about 6000 people as great as you!” is entirely equivalent, if less romantic.

\(^{207}\) This line of critique does assume that suspicion fell upon the accused couple only because they possessed the characteristics identified by the eyewitness. There could of course have been other factors implicating the accused couple, though in order for them to play an inculpatory role, they would have to have been presented to the jury and explicitly accounted for in the statistical expert’s testimony, which they were not.
6.2.2.2  R. v. Johnson

In *R. v. Johnson*, Johnson appealed his conviction on the basis that the reasoning followed by the decision maker at trial instantiated the prosecutor’s fallacy. You will recall that in that case, Johnson’s DNA was found to match the DNA from the semen sample on the victim’s pyjamas and evidence was provided at trial that the probability of a “random” match was 1 in 890 billion.

In this case it would appear that the fallacy of “failing to think of the opposite” was avoided. Here, the single piece of evidence that was relied upon was in fact the likelihood of “the opposite”. And it is for this reason that that case would appear to have been correctly decided and correctly upheld on appeal.

Again, what we are interested in is \( (H|D) = \frac{p(D|H)}{p(D|H) + p(D|\neg H)} \). The probability of a random genetic match is actually the probability of a match \( p(D) \) given that it was not the accused whose semen was found on the complainant’s pyjamas and quilt. So in this case, the evidence that \( p(D|\neg H) \) was 1 in 890 billion meant that the probability of the hypothesis given the data looked like this:

\[
p(H|D) = \frac{p(D|H)}{p(D|H) + \frac{1}{890 \text{ billion}}} = \frac{x}{x + \frac{1}{890 \text{ billion}}}
\]

It is true that the value of \( x \) was unknown, but the only way in which \( p(H|D) \) could fail to be an incredibly large number would be if the value of \( x \) were to be very small in comparison to 1 in 890 billion. That is, if \( x \) were, say, 1 in 890 quadrillion. But without knowing the exact value, we can be highly confident that the probability of the data (a genetic match) given the hypothesis (that the accused was at the scene of the crime) would not be a small number at all. It could only be anywhere close to such a small number if the test gave an outrageously high number of false negatives, that is, almost never identified a true match when there was one. And so long as \( x \) was a large number relative to 1 in 890 billion (which we can be confident it was) the value of \( p(H|D) \) would approximate 1 (i.e. close to certainty).

Of course, this does not take into account other possibilities of error here. That is, the chance of a random match isn’t the only way in which you could get evidence of a match absent the semen
sample actually having been the accused’s. Human errors also occur. For example, samples can get mixed up.\textsuperscript{208}

But our interest in this case is in the inference that the decision makers drew from the evidence of the likelihood of a random match to the probability of the accused’s guilt. And this was an appropriate inference to make.

In this case, it was the appellant’s counsel who appears to either have been in need of clarification on basic statistical concepts and reasoning, or perhaps he was hoping that his audience was. The argument that the jury should have been specifically instructed about the prosecutor’s fallacy in this case does not hold water.

\textbf{6.2.2.3 The Case Against Lucia de B.}

A variant on the same error that occurred in Peoples v. Collins also arose in the case against nurse Lucia de B. set out in the first chapter. As you will recall, in that case, Lucia was sentenced to life in prison for the murder or attempted murder of a number of patients in two hospitals where she had worked. The main issue at trial was whether the disproportionate number of incidents during Lucia’s shifts was merely coincidental. An expert witness testified that the probability that Lucia had experienced as many incidents as she had “randomly” was less than 1 in 342 million. It was inferred from this that the chance of her innocence was less than 1 in 342 million. Yet to draw this inference was to “fail to think of the opposite”, namely to fail to consider the percentage of cases (or in this case the probability of) \textit{no foul play} and a disproportionate number of incidents.

In this case I will use an analogy to the Digirosa example, as opposed to explicitly invoking Bayes’ rule, to illuminate the error. See Table 3. In the Digirosa example, the ultimate question was “Digirosa here?” Here, the ultimate question is “foul play here?” And while the “red rash” is the indicator giving rise to suspicion of Digirosa, it’s the “disproportionate number of deaths” that serves a potential indicator in the Lucia case.

Evidence relating to the complementary factors (a) and (b) was not presented, but this is often the case. In a sense, it is already built into the legal decision-making scheme that culpable behaviour is very rare in the population.

What is of particular interest to us here, however, is that the only real evidence that was provided was that the probability of finding this percentage of “incidents” was less than 1 in 342 million. And from this it was inferred by the decision makers that (c) was close to 100%. After all, since the chances of what had happened occurring where so remote, the only possible inference is that the nurse did something wrong; after all, we understand the mechanisms by which culpable behaviour would give rise to this extremely rare confluence of bad events.

But this line of reasoning is fundamentally misguided. And one way of seeing this is to consider that it fails to take (d) into account, as we know people are apt to do. That is, it fails to consider the percentage of incidents of innocent behaviour where there are a disproportionate number of deaths. When we’re dealing with an event of great rarity – in this case, a highly unlikely
clustering of deaths – in a sense we know as little about its relationship to innocent behaviour as its relationship to foul play.

The reasoning error in this case turns on the inference from a really rare event to the only conceived-of explanation for it. That is, the decision makers reasoned that since what happened here was super rare and there was one super rare mechanism that one could think of that would have caused it – namely the super rare mechanism of a nurse killing or neglecting babies – they had their explanation.

But had they appreciated the absolute necessity for comparing the likelihood of the culpable mechanism to the other possible mechanisms, they would likely have avoided this mistake. In particular, they would have thought to consider potentially innocent mechanisms such as the following: (i) the types of patients that this nurse had, i.e. she may have preferred caring for, or have been assigned to, sicker patients; (ii) the nature of the ward, i.e. was it a palliative care ward where those there already had a far greater likelihood of death; (iii) did she work with less experienced nurses whose incompetence could have explained the preponderance of deaths; (iv) was she only assigned to night shifts during which it is known that the incidence of death is higher; or (v) is this just a super rare clustering of events (appreciating that it is actually very common for super rare events of any description to occur; given the vast class of all super rare events, the possibility of one such event occurring is actually pretty high). 209

Once all possible “innocent” explanations are considered, it becomes quite clear that in spite of the rarity of the proportion of deaths on the nurse’s watch, the percentage of (d), incidents of innocent behaviour, may be just as high as or higher than the percentage of (c), incidents of foul play. That is, it may be more likely that there is an innocent explanation for the confluence of deaths on the nurse’s watch than a culpable one.

209 It is by operation of this principle that we are so frequently amazed by the “coincidence” of running into someone who shares an acquaintance of ours when we’re far away from home. “What are the chances that when I was in Japan of all places, I’d run into someone who knew my kindergarten teacher?!” Well, the chances that you’d run into someone who knew your kindergarten teacher are pretty darn small. But, the chances that you’d run into someone – anyone – who knows anyone that you know are actually not so small at all. When you think of all the people you’ve ever known and all the people who have ever known them, this is actually a pretty large group of people. The chances that you’d run into one of them is not so inconsiderable after all.
In considering the three cases above, I hope to have demonstrated that either by invoking Bayes’ rule explicitly or by analogizing to the “Digirosa” example, the error of “failing to think of the opposite” (or alternatively, awareness that insufficient information is available to ground a decision) can be made relatively intuitively obvious. I have demonstrated this both by explicit invocation of Bayes’ rule for those who are more comfortable with formulas and by invocation of the “Digirosa” example for those who prefer to think in sentences rather than symbols.

6.3 Implications

6.3.1 Judges vs. Juries

Having set out the kinds of errors people are apt to make when reasoning with probabilities and the ways to conceptually frame the information so as to reduce the likelihood of error, two things perhaps become clear: (i) it is possible to frame the information in such a way that the probabilistically correct conception becomes intuitively accessible; and (ii) this still requires a fair bit of work.

In working through, for example, the case of Lucia de B. above, I trust that by the end of the analysis it became clear to my reader what the decision makers may have missed. It also may have become clear that it took some time to arrive at this clarity and that some concerted effort was required.

The implications of this are both that it is possible for legal decision makers to use probabilistic evidence properly and to reason more clearly about uncertain facts, but also that this is not always a simple matter. It makes sense that this is so; if it were not, the onus would be on me to explain why the problem had not been fixed long ago.

What I am recommending therefore as a way forward is both the inclusion of some basic probabilistic reasoning concepts in legal education for lawyers and judges and the use of clearer modes of presenting such evidence to lay juries.

It would, however, be natural to think that one of the implications of my project could be to support the decline of lay juries in favour of professional decision makers who are more likely to receive some requisite training. Accuracy in decision making is not, however, the only consideration in play in such a discussion and I do not propose to fully consider the question at
present. I would however suggest that this conclusion may be appropriate when legal cases that involve more complex statistical evidence such as environmental or pharmaceutical cases are involved. In run-of-the-mill criminal cases where juries are most common, the appropriate consideration of one critical piece of statistical evidence, such as DNA evidence, can be promoted by proper presentation by a neutral and informed expert. Helpful jury charges by the trial judge that frame the evidence in the ways I have suggested may also assist.

6.3.2 Experts

One of the more pressing issues within law today concerns the use of expert evidence. The question of when experts should become involved in legal decision making and how to make use of the evidence they present when they do become involved is obviously quite closely related to the idea of embracing statistical evidence. This is so both because experts often present evidence in terms of overtly statistical information and also because experts could potentially be employed to assist in integrating probabilistic evidence into legal decision making. The kinds of evidence that have been the topic of this dissertation and which may be subject to the types of reasoning fallacies discussed are often introduced by expert witnesses or expert reports.

Expert evidence constitutes an exception to the general rule against the admission of opinion evidence:

> The opinion rule is a general rule of exclusion. Witnesses testify as to facts. As a general rule, they are not allowed to give any opinion about those facts. Opinion evidence is generally inadmissible. Opinion evidence is generally excluded because it is a fundamental principle of our system of justice that it is up to the trier of fact to draw inferences from the evidence and to form his or her opinions on the issues in the case.²¹⁰

The most common exception to the opinion rule is an opinion of a properly qualified expert. The fears and concerns that relate to the use of expert scientific evidence in the law largely overlap with those pertaining to the presentation of statistical evidence. The concern is that decision

---

²¹⁰ *R v K (A) (1999)*, 45 O R (3d) 641 (CA) at para 71.
makers will defer unduly to the conclusions of experts and will fail to sufficiently understand the testimony of experts (on matters overtly statistical or not) such that they either misuse it or are unable to identify weaknesses in the evidence as presented.

The dangers associated with expert evidence and recommendations to limit its misuse are the topic of the Goudge Inquiry Report, released on October 1, 2008. The report was commissioned by the Ontario government after it was discovered that the flawed forensic evidence of Dr. Charles Smith had led to a number of wrongful convictions in relation to the deaths of children. Many of the cautions recommended in that report, and some of the recommendations for how courts should deal with the evidence of experts in general, can be applied to the more specific instance of the introduction of statistical evidence into court.

By way of background, from 1981 to 2005, Dr. Smith worked as a pediatric pathologist at Toronto’s Hospital for Sick Children. He had no formal training or certification in forensic pathology but became involved as a witness in criminal cases involving pediatric deaths in the early 1980s. By the 1990s he was touted as the “expert” in the field. From the start, there were some who expressed concerns about his work, but it was not until 2005 that the Chief Coroner for Ontario called for a full review of Dr. Smith’s work in criminally suspicious cases and homicides in the 1990s. Five highly respected forensic pathologists external to the Ontario Coroner’s office who, unlike Dr. Smith, were actually trained and certified in forensic pathology, were asked to review Dr. Smith’s work.

The results of the review were startling. In all but one of the 45 cases examined, the reviewers agreed that Dr. Smith had indeed conducted examinations as claimed. However, in 9 of the 45 cases, the reviewers did not agree with significant facts that appeared in Dr. Smith’s report or testimony at trial. In 20 out of 45 cases, the reviewers took issue with Dr. Smith’s opinion; in 12 of those cases, there had been a guilty finding by the courts. The results of this review were released in April 2007. Six days later, a Commission was struck to assess the practice and oversight of pediatric forensic pathology in Ontario.

A number of old cases were reopened and convictions overturned. In one particularly egregious case, William Mullins-Johnson spent 12 years in jail after he was convicted of sodomizing and murdering his 4-year-old niece after Dr. Smith had concluded at trial that the girl was sexually assaulted at the time of death. Since William Mullins-Johnson was the only person with the girl
at the time of her death, a finding of sexual assault at the time of death implicated Mullins-Johnson. It also contradicted the defence’s account that the girl, who had a history of vomiting in bed, might have died of natural causes.\footnote{R v Mullins-Johnson, 2007 ONCA 720.}

Attempts to clear Mullins-Johnson’s name based on DNA testing were foiled by the unavailability of tissue samples. In 2005, however, 11 years after the trial, the missing tissue samples turned up in Dr. Smith’s office. William Mullins-Johnson was released on bail in 2005, pending a review of his case. On July 16, 2007, a report by three expert pathologists determined there was no evidence that the girl was sexually assaulted, and the Ontario Attorney General Michael Bryant made a statement that William Mullins-Johnson's conviction could not stand and that he should be acquitted by the appeals court. On October 15, 2007 he was.

Mistakes made by Smith and his superiors in the coroners’ office led to a number of wrongful convictions and prosecutions. In addition to being jailed, some parents had surviving children temporarily or permanently removed from their custody.

Nineteen bungled cases were the subject of a public inquiry, led by Justice Stephen Goudge, the final report of which was released in October 2008. The inquiry was tasked with discovering what went so badly wrong in the oversight of forensic pathology in Ontario and recommending steps to assist in preventing such gross failings in the future.

A number of the major problems identified by the Inquiry were specific to Dr. Smith or to the field of pediatric pathology. For example, the inquiry found that Dr. Smith lacked training and basic knowledge about forensic pathology and was sloppy and inconsistent in how he gathered and reported information. There were also additional respects in which he was a very poor participant in the criminal justice system. He misunderstood his role, prepared inadequately, and testified in sometimes confusing, other times dogmatic, ways. He mistakenly saw his role as an advocate for the Crown whose job it was to “make a case look good”. He failed to appreciate that he should not speculate in court. He also made false and misleading statements to the court.
One of the conclusions that came out of the inquiry was the need for oversight of professionals in terms of setting standards and holding people responsible for their actions and decisions.

In terms of the role of counsel, since probing cross-examination is an important part of the truth seeking process, it was found to be important that both defence and Crown counsel be educated in forensic pathology so as to be better equipped to observe the scope and limitations of the expert’s expertise and opinions.

In addition, it was recommended that judges should consider expert scientific evidence in light of the following:

1. the reliability of the witness, including whether the witness is testifying outside his or her expertise;
2. the reliability of the scientific theory or technique on which the opinion draws, including whether it is generally accepted and whether there are meaningful peer review, professional standards, and quality assurance processes;
3. whether the expert can relate his or her particular opinion in the case to a theory or technique that has been or can be tested, including substitutes for testing that are tailored to the particular discipline;
4. whether there is serious dispute or uncertainty about the science and, if so, whether the trier of fact will be reliably informed about the existence of that dispute or uncertainty;
5. whether the expert has adequately considered alternative explanations or interpretation of the data and whether the underlying evidence is available for others to challenge the expert’s interpretation;
6. whether the language that the expert proposes to use to express his or her conclusions is appropriate, given the degree of controversy or certainty in the underlying science; and
7. whether the expert can express the opinion in a manner such that the trier of fact will be able to reach an independent opinion as to the reliability of the expert’s opinion.  

In terms of the role of the court, the following recommendations were made:

1. When a witness is put forward to give expert scientific evidence, the court should clearly define the subject area of the witness’s expertise and vigorously confine the witnesses’ testimony to it.
2. Trial judges should be vigilant in exercising their gatekeeping role with respect to the admissibility of expert evidence.
3. In determining the threshold reliability of expert scientific evidence, trial judges should have regard to the most germane tools and questions.
4. The trial judge’s gatekeeping function may be facilitated, in some cases, by written descriptions in the expert reports of the nature of the relevant discipline and how it engages with the various criteria of reliability.
5. Judges should consider whether there are parts of the proposed expert evidence that are sufficiently reliable for admission and others that should be modified for admission.
6. The National Judicial Institute should consider developing additional programs for judges on threshold reliability and the scientific method.
8. A code of conduct for experts in criminal proceedings should be created and incorporated into the criminal justice system. The code should provide that experts have a duty to assist the court on matters within their expertise.
9. In cases where expert evidence is important, trial judges should make use of model charge language provided by the Canadian Judicial Council model instructions.

10. Judges should remind jurors that they should apply their common sense to expert testimony and that it is up to them to decide whether to accept all, part, or none of the expert’s opinion.

11. In addition, judges should, in appropriate cases, provide structured questions to assist the jury in determining the ultimate reliability of the expert’s opinion.\textsuperscript{213}

6.3.3 Identifying the True Locus of Danger

In many ways, what went wrong in the Dr. Smith cases sheds light on the dangers inherent in the use of statistical evidence. It is my hope that these dangers can be identified and limited so as to ensure that the value of meritorious uses of such evidence may be preserved.

The failure of the legal system to properly limit and define the role of the expert and the over eagerness of decision makers in abdicating their decision-making authority exacerbated the fall-out of Dr. Smith’s mistakes. This abdication of decision-making authority to a partisan expert is highly problematic. And the risk is that the presentation of statistical evidence encourages such abdication since it is too complex for decision makers, and even the cross-examining lawyers, to properly engage with it and assess its weaknesses.

One way to address this is by making the evidence more understandable through the intuitive presentation of statistical evidence, along with some basic statistical education for those in the legal profession. Had Dr. Smith presented evidence about the number of cases he had seen and in which percentage of them he had noticed such-and-such features, this would have been less misleading than simply presenting his “expert opinion” that, for example, “the child was sexually abused”.

In addition, it may be helpful to address head on the tendency of decision makers to shift decision-making authority to others. For while the methods I recommend will go some distance to improving understandability, there will of course still be cases in which the evidence presented is complex. It is, however, the combination of complex evidence with an eagerness to rely on another’s opinion that is most problematic.

The eagerness of decision makers to rely on the judgment of another, and not just consider information that s/he presents is highlighted by a fascinating series of studies conducted in relation to the red/green cab/bus problem discussed in chapter four.

In a series of studies, Gary Wells of Iowa State University tested subjects’ treatment of probabilistic evidence that was easily and intuitively understood. That is, unlike the studies of Daniel Kahneman, Paul Slovic and Amos Tversky, his focus was not on eliciting subjective probability assessments that diverged from mathematically correct probabilities, but rather on accurate subjective probabilities that none-the-less gave rise to surprising legal assessments.214

Wells starts with the standard blue bus problem already described in chapter four. As a reminder, in that problem, the blue bus company owns 80% of all buses in a given area. There is an accident involving driver negligence and the injured plaintiff sues the blue bus company. The only evidence put forward by the plaintiff at trial is that the blue bus company owns 80% of the buses in the area. Since the standard of proof in civil suits is proof on a balance of probabilities, it is sometimes expected that the blue bus company should be found liable since the probability that it was their bus is greater than 50%. However, the common legal intuition in this case (and the result actually reached by the courts in the analogous case of Smith v. Rapid Transit, 1945) is that there is insufficient evidence here for a finding of responsibility.

If you will recall, it was suggested in chapter four that one possible explanation for this result is that the legal intuition in that case rests on the idea of insufficient evidential weight. That is, that a probability of responsibility based on market share alone is not enough to ground a finding of legal responsibility.

Wells makes minor alterations to the nature of the problem in order to test various explanations for the divergence of subjective probability (which is 80%) and the legal assessment, which falls short of a balance of probabilities. (He also uses a “grey” bus company instead of “green” or “red” to contrast with the blue bus company. For no apparent reason, except perhaps to underline the unreliability of colour assessments and/or recollections, the literature on the “bus” problem

seems unable to agree on the colour of the second bus company’s buses). Wells stops short of drawing any general conclusions about the results from this series of experiments, but I wish to suggest that they in fact provide strong evidence for the eagerness of decision makers to shift responsibility onto someone else, whether an expert or non-expert witness. This inclination, once recognized, can assist the legal system in avoiding the potential for certain errors. That is, the risk lies not in the use of experts or statistical evidence per se, but in the over-willingness of decision makers to give up their responsibility. This tendency can however be offset by making statistical evidence (and other evidence put forward by experts) intuitively accessible and requiring special jury charges that are responsive to potential probabilistic errors in reasoning with the evidence presented.

In all versions of the case used by Wells, a dog is hit by a bus that was driving recklessly (and negligently) down the road, and evidence is presented that there are only two bus companies that travel in the area. In all these cases, it is assumed as established that there was a duty of care, a breach of this duty, and that the breach of this duty caused the dog’s death. The only issue to be determined is which bus did it. What differs from scenario to scenario is the way in which the additional piece of evidence relating to the identity of the bus is presented. In all cases, the “standard” scenario is contrasted against an altered scenario in which the probabilistic evidence is presented differently. The “standard” scenario is the one with which we are already familiar: a witness testifies that the blue bus company owns 80% of the buses and accounts for 80% of the traffic on the road in question while the other bus company owns only 20%.

In experiment #1, the alternate scenario features an error-prone witness who testifies that he entered “blue bus” in his log book 10 minutes before the accident took place at a location 10 minutes from the weigh station where he entered the information. There is also evidence that the entries in weigh station attendant’s log book are 80% accurate.

One group of eighty psychology students read the “standard” version and another group of 80 the weigh attendant version. The two groups answered the question of the probability that the blue bus had hit the dog in the same way: 80-85% of both groups said that the probability that it was the blue bus was 0.8. But interestingly, there was a considerable difference in the verdict assessments: only 8.2% found the blue bus liable in the standard version while 67.1% found the blue bus company liable on the weigh attendant version.
These results suggest a couple of things: (i) they suggest that refusal to find a defendant responsible where there is a >0.5 probability of causation need not mean that the probability has been miscalculated (here the majority of subjects said that the probability was 0.8 and at the same time did not return a verdict of liability); and (ii) they support my hypothesis that there are considerations in legal verdicts that go beyond probability assessments. The subsequent variations on the experiment aim to suss out what these extra-probabilistic considerations are.

In experiment #2, causal relevance is tested by presenting the evidence in terms of accident responsibility. That is, a transportation official testifies that the blue bus company is responsible for 80% of all the accidents involving buses in the area. The results for both subjective probability assessment and the verdict of liability were the same as with the standard version.

In experiment #3, the experiment is modified to test whether it is some notion of distributional fairness that explains the discrepancy between a subjective probability assessment of 0.8 and a refusal to find liability on a balance of probability standard. That is, it is postulated that what subjects are reacting to in both the “standard” and base rate “causal” accounts in experiments 1 and 2 is that the use of such base-rate evidence would result in a finding against the one defendant company in all such cases. That is, once it was determined that one company owned more than half of the buses or was responsible for more than half of the accidents, in the absence of other evidence, they would always be found liable. This approach would give rise to shifting the onus of proof to the defendant in all such cases. In contrast, in experiment #3, the scenario was altered as follows: both companies own the same number of buses. All the buses have one of two types of tires. A witness testifies that the tire track prints on the dead dog matched the tires on 80% of the Blue Bus Company’s buses and only 20% of the other company’s bus tires. Unlike the “standard” or the weigh attendant versions of the problem, here the onus is not shifted to company A in similar future cases. On the contrary, if both companies are equally negligent, in the long run, they can expect to be found liable an equal number of times: that is, if tire-track evidence is used in all such cases, and a match with the tracks of 80% of your buses leads to a finding of liability, if both companies are equally negligent, they will each be found liable in 80% of the cases brought against them. In this way, this version of the problem addresses the possibility that subjects could be rejecting the base-rate evidence out of a concern that its use would result in findings against one defendant in all cases; this possibility is addressed here by evenly distributing liability between the two companies in the long run.
But once again, while most subjects reported their subjective probability that it was company A at 0.8, the decision to return a verdict of liability was only slightly higher than for the other two accounts and considerably lower than for the weigh attendant case.

One subset of the subject group was comprised of 37 practicing trial judges. They were asked to provide an explanation for their answers. The most common explanation was to dismiss the evidence on the basis that it was “naked” or “mere probability”. Some noted the absence of information in relation to sample size: 80% responsibility for all accidents is more meaningful where there have been 1000 accidents than where there have been 5. While this is an excellent point, all who made it appeared to miss the fact that there should be a perfectly symmetrical concern in relation to the weigh attendant evidence; it is unknown whether his 80% reliability in log-book entries was assessed over a small or large number of entries.

What is of greatest concern is that 7 of the 27 judges invoked the argument that since it was either company A or company B that was responsible, the real probability that either company was responsible is 50/50 regardless of the other evidence. This line of thinking is shocking in its failure to understand the value of statistical evidence – would these people also accept an even-odds bet on whether or not the sun will rise tomorrow? After all, there are only two possibilities: either it will or it won’t. If so, I think I’ve found my road to riches.

On top of this, the 50/50 odds line of reasoning fails to distinguish between the weigh attendant’s evidence and any of the other versions of the evidence. If you think that any degree of uncertainty yields 50/50 odds, then you have no reason to prefer an error-prone log to tire-track evidence. You should find in favour of the defendant any time the plaintiff has failed to convince you of the certainty of her claim.

Wells ultimately concludes that we could benefit from more psychological research on this topic. I would agree. I would also suggest that a plausible hypothesis for the discrepancy between the weigh attendant version and all the others relies heavily on the fact that the source of uncertainty in the former is another human being. This fact may affect decision makers’ assessments in at least two ways: (i) people assume that they understand the mechanism of human error and that it can be probed (i.e. if the weigh attendant were truly stupid or dishonest or weird or colour blind, this information would be put to them such that in its absence they can assume that they may
place a high degree of weight on the 80% reliability of his logs); and (ii) people feel more comfortable transferring responsibility onto someone else.

The second hypothesis is particularly relevant for our purposes since it does much to explain the risk inherent in expert evidence and the reason that the Dr. Smith cases went so horribly awry. This hypothesis is also supported by two additional follow-up studies performed by Wells in which the tire-tracks version of the scenario was presented but in experiment #4, a county official testified that the tire-tracks found matched all of the Blue Company’s buses (and none of the other company’s) and that this matching technique was 80% reliable. The witness proceeded to testify that, “based on this technique, he believed that the bus that ran over Mrs. Prob’s dog was a Blue Bus Company bus.” A liability verdict was returned here by over 5 times as many subjects as in the original tire-tracks evidence version and by approximately the same number of subjects as in the weigh attendant versions.

Since what is common to the weigh attendant and tire-tracks + belief endorsement (experiment #4) versions is that the source of the evidence and the loci of uncertainty becomes another human being, it seems plausible to infer that there is a preference for relying on another person to take responsibility.215

One possible implication of this preference is that we should take great care in allowing witnesses to express opinions or beliefs. It is in acknowledgement of this need for care that the general rule against opinion evidence exists and that the law tries to limit expert evidence. But limiting the use of expert evidence is a crude mechanism for addressing this concern. In fact, what Wells’ studies suggest is that statistical evidence properly presented by experts and unaccompanied by belief statements are unlikely to inappropriately sway decision makers. This supports my belief that the real danger lies in ignoring relevant statistical evidence rather than in overemphasizing it. Focusing the loci of danger in this way – namely on any expression of opinion or belief – allows the law to embrace useful statistical evidence without risking that decision makers will be inappropriately swayed.

6.4 Recommendations

Many of the considerations and recommendations made in relation to the use of expert evidence are also helpful for dealing with statistical evidence, whether it is introduced by an expert witness or otherwise.

In particular, decision makers should consider statistical evidence in light of the following:

1. the reliability of the evidence;
2. the reliability of the scientific theory or technique on which the evidence draws;
3. whether the risk of error will be apparent to the decision maker;
4. whether the language used to present the statistical evidence is appropriate; and
5. whether the evidence can be presented in a manner such that the trier of fact will be able to reach an independent opinion as to its reliability.

In relation to considerations 4 and 5, the pie chart conceptual technique put forward earlier in the chapter is clearly pertinent. That is, the more understandable the evidence is to the decision makers, the greater are the chances that they will be able to make their own assessment of the evidence without relying inappropriately on either the views of the witness introducing the evidence or the superficial indications of the evidence itself. It is also apparent, however, that a greater degree of numerical literacy or comfort with manipulating certain basic probabilistic concepts will assist in reaching sound conclusions.

In terms of recommendations for the court, I would import the following recommendations into the context of the use of statistical evidence:

1. Law schools should consider adding a short course in basic numerical literacy as part of the required curriculum;
2. The National Judicial Institute should consider developing additional programs for judges on basic numerical literacy;
3. A code of conduct for experts could be created that instructs experts on helpful modes of presenting statistical evidence;
4. When statistical evidence is put forward, the court should clearly define the limits of the evidence and possibly warn against common misunderstandings;
One recommendation in the Goudge Report with which I would take issue is the recommendation that judges should, “remind jurors that they should apply their common sense to expert testimony”. As we have seen, common sense can be misleading where evidence is improperly presented. While I have argued that intuition and sound Bayesian reasoning can be brought into line, this assumes proper presentation and in some cases some limited additional explanation. What I think is actually meant by this recommendation, and reframed in this way would be a recommendation that I could full-heartedly endorse, is that decision makers should be reminded to keep their critical reasoning faculties engaged in relation to all the evidence that is presented and not to defer to “experts”. A good expert should present evidence in such a way so as to make it understandable to the decision makers. If it is not understandable, then the expert’s “opinion” should not be relied upon.

Recommendations #1 and #2: Law schools and the National Judicial Institute should consider including basic numerical literacy to their curriculums

A basic course in numerical literacy generally and principles of probability in particular should be offered at all Canadian law schools and required for trial lawyers. This course could build on the literature used for training doctors to be better Bayesians and referred to in chapter 5 of this dissertation. Basic concepts of probability such as the following would be covered:

- What is probability (mathematical probability, different theories of probability)
- How to combine probabilities (including the product rule and the concept of independence)
- The limitations of probability (a.k.a. why the weatherman can be right even when he’s wrong)
- Common fallacies involving probability (e.g. base rate neglect, failing to think of the opposite)
- Examples of common fallacies in the legal context
There are a number of texts that could be relied on for this purpose of covering these topics, but using legal cases (such as the ones discussed in this work) as illustrations could also be helpful.216

Recommendation #3: A code of conduct for experts could be created that instructs experts on helpful modes of presenting statistical evidence

As of January 1, 2010, there is a new “Duty of Expert” rule (rule 4.1) in Ontario’s Rules of Civil Procedure that is aimed at educating experts about their prevailing duty of objectivity. The creation of this rule is in part a response to the startling declaration by Dr. Smith that he saw his role as helping the Crown and that other experts are too often seen to be partisan participants in the court process. The new rule requires experts to confirm that they understand their duties to the court and their prevailing duty of objectivity by signing an “Acknowledgment of Expert’s Duty” (Form 53) and appending it to their reports.

The “Acknowledgment of Expert’s Duty” could be amended to include a duty to present statistical evidence in a manner that is easily understood.

I would also recommend a thorough consideration of the role of the expert and whether expert evidence should in fact constitute an exception to the rule against opinion evidence. That is, it may be possible in many if not most cases for experts to limit their testimony and reports to an explanation of their particularized knowledge of a fact in dispute and/or a description of the testing process they took. The results of any such investigation could be shared with decision makers without an expression of opinion or belief on the ultimate fact in question on the part of the expert. That is, instead of stating that in her opinion the death was caused by strangulation, an expert coroner could limit her testimony to something like the following:

I have viewed 10,000 corpses in my professional career. Out of these, I attributed 100 to strangulation with more than an 80% degree of probability. I based these assessments on red marks on the neck that were consistent with the shape of a hand.

print. Here, I observed red marks on the neck that when compared with the other cases I attributed to strangulation, I would attribute a 90% probability of strangulation.

The expert could stop short of expressing her opinion that *this* was a case of strangulation. While it may seem that the difference is merely semantic, Wells’ studies suggest that this may not be so. By forcing experts to stop short of stating their opinion or belief with certainty about the particular fact in dispute, and sticking to the language of observation and probability, decision makers are forced to keep the responsibility of the ultimate call squarely on their shoulders. This distinction may well be important in spite of the fact that the language of observation is often highly belief-laden. That is, to “observe” that two marks are “similar” is value laden in a way that the simple observation of a red mark is not. That is, when the cross-examining lawyer asks the coroner whether there is any independent test of whether her 100 prior assessments of strangulation were in fact accurate and the coroner is left without a response, the decision maker must squarely confront this weakness in the evidence and is denied the comfort of a witness who “sticks to her belief” that it was strangulation even in the face of revealed sources of uncertainty.

Uncertainty is as unpleasant as it is ubiquitous. But it is the role of the ultimate decision maker and not the expert (particularly a partisan one) to shoulder this burden. This should be made clear to experts, lawyers, judges and juries so that all within the system can do their part to enforce this clear differentiation in roles.

Recommendation #4: When statistical evidence is put forward, the court should clearly define the limits of the evidence and possibly warn against common misunderstandings.

6.4.1 Clearly Define the Limits of Expert Evidence

Just as experts should be reminded of their limited role in the acknowledgement they sign, decision makers should also be reminded of this limitation.
The current model jury charge on expert evidence published by the Canadian Judicial Council reads as follows:\textsuperscript{217}

\textbf{FINAL INSTRUCTIONS}

\textbf{10. Types of Evidence}

\textbf{10.3 Expert Opinion Evidence}

(General Instructions) 73

(Last revised February 2004)

[1] You heard the evidence of NOW, an expert witness. S/he gave an opinion about some technical matters that you may have to consider in deciding this case. S/he is qualified by his/her training, education and experience to give an expert opinion.

[2] Remember, the opinions of experts are just like the testimony of any other witnesses. Just because an expert has given an opinion does not require you to accept it. You may give the opinion as much or as little weight as you think it deserves. You should consider the expert’s education, training and experience, the reasons given for the opinion, the suitability of the methods used and the rest of the evidence in the case when you decide how much or little to rely on the opinion. It is up to you to decide.

[3] NOW was asked to assume certain facts. What an expert assumes or relies on as a fact for the purpose of offering his or her opinion may be the same or different from what you find as facts from the evidence introduced in this case.

[4] How much or little you rely on the expert’s opinion is up to you. But the closer the facts assumed or relied on by the expert are to the facts as you find them to be, the more helpful the expert’s opinions may be to you. How much or little you rely on the expert’s

opinion is entirely up to you. To the extent the expert relies on facts that you do not find supported by the evidence, you may find the expert’s opinion less helpful.

While a charge of this nature may be required where there appears to be no way that the actual reasoning process of the expert could be laid bare to and understood by the jury, I suggest that this should be more the exception than the rule. Instead, expert evidence should be presented as I have suggested above, and stop short of conclusive opinion. While this may be easier said than done, identifying overemphasis on the opinion aspect of expert opinion evidence as problematic is a good start. In this way, another model charge should be designed that emphasizes this. Such a charge may look something like this:

**FINAL INSTRUCTIONS REVISED**

[1] You heard the evidence of NOW, an expert witness. S/he provided information about some technical matters that you may have to consider in deciding this case. Her training, education and experience may help you in your decision-making process.

[2] NOW was asked to assume certain facts. What an expert assumes or relies on as a fact may be the same or different from what you find as facts from the evidence introduced in this case.

[3] Remember, experts may be mistaken. You do not have to accept anything that the expert said and may give it as much or as little weight as you think it deserves. You should consider the expert’s education, training and experience. You should also focus on the reasons given for the opinion, the suitability of the methods used and any weaknesses in the expert’s evidence that came up in the examination-in-chief or cross-examination. Did what the expert say make sense to you?

[4] There is uncertainty in all the evidence you have before you. It is ultimately your job (and not the job of the expert) to assess the extent of this uncertainty and then to draw your own conclusions.
6.4.2 Consider Warnings Against Common Misunderstandings

Assuming that recommendation #3 is adopted and that judges become more educated themselves about the possible pitfalls of reasoning with probability, it may be appropriate to warn juries about common misunderstandings in response to the presentation of particular types of statistical evidence.

For example, in the case of *People v. Collins*, the judge, having been educated on this very matter, could have followed the presentation of statistical evidence by placing this evidence in its proper context and warning the jury against “failing to think of the opposite”. Preferably, the experts, or lawyers in cross-examination, should perform this role, but the trial judge could appropriately step in to fill any gaps in this regard. In so doing, s/he would of course have to take care not to appear to be favouring one side over the other.

For example, in the case of *R. v. Terceira*, 1998 CanLII 2174 (ON C.A.) the trial judge instructed the jury as follows on the use of DNA evidence:

> The DNA tests in this case were conducted soon after The Centre of Forensic Labs opened itself for DNA case work. And there is evidence that challenges the conduct of the DNA tests and evidence that challenges the results. These are reasons for you to take a good close look at the DNA evidence, yourself, all the evidence you've heard from the Crown and the defence, and scrutinize it to see if you consider it reliable as a piece of circumstantial evidence. You have obviously followed it closely, spent a lot of time in court looking at the autorad projections, various aspects of the bands and their measurement and interpretation. You've seen those from both sides. I don't intend to repeat all of that. You have part of it, or small parts of it, anyway, in some of the material in front of you. You followed it very closely. I am confident you will use your common sense, you won't be overwhelmed by any aura of scientific authority advanced by any of the DNA witnesses. The assessment of the evidence really does boil down to a common sense assessment of the evidence, of the various opinions that you have heard, your assessment.

While the phrase “common sense” may not be ideal (since as we have seen, depending on how information is framed, our common sense can sometimes lead us astray; I would encourage use
of the phrase “your own judgment” instead), the message conveyed in the jury instruction above is essentially appropriate. It urges the jury to engage with the evidence themselves, to consider its strengths and weaknesses, and to make their own evaluation. The Supreme Court of Canada approved this instruction to the jury in *R. v. Terceira*, [1999] 3 S.C.R. 866.

### 6.5 Concluding Remarks

It is stated in the Goudge Report that, “no justice system can be immunized against the risk of flawed scientific opinion evidence. But with vigilance and care, we can move toward that goal.” An analogous direction applies to the use of statistical evidence. We cannot immunize the justice system against the risks of misunderstanding or misusing statistical evidence. At the same time, such evidence should not be held to a higher standard of risk avoidance than other types of evidence. Furthermore, we can take steps towards minimizing the risk of error. It is my hope that this dissertation constitutes one such step.

In this final chapter I have identified two common statistical errors in legal cases that should be overcome: (i) the avoidance of helpful statistical evidence based on a misconception of irrelevance; and (ii) failing to “think of the opposite” when confronted with evidence in statistical form.

I suggest that the first error is related to the second in that so long as statistical evidence is commonly misused, courts cannot be criticized for avoiding it. So the solution to both problems is to limit the misuse of such evidence.

I have made four ultimate recommendations for curtailing misuse:

1. Law schools should consider offering a short course in basic numerical literacy for lawyers;
2. The National Judicial Institute should consider developing additional programs for judges on basic numerical literacy;
3. A code of conduct for experts could be created that instructs experts on helpful modes of presenting statistical evidence

218 *Supra* note 212 at 28.
4. When statistical evidence is put forward, the court should clearly define the limits of the evidence and possibly warn against common misunderstandings.

This dissertation takes a first step in accomplishing the first two educational goals by hopefully bringing to the attention of the legal community the existence of a problem and illustrating that a little education would go a long way to address it.

Bayesianism provides a good normative model for working with statistical evidence and can be used to inform its proper presentation. Presented properly, plenty of statistical evidence can be understood intuitively. And understood intuitively, statistical evidence can be integrated into decision makers’ reasoning about the case, thereby avoiding problematic abdication of decision-making authority to experts.

As scientific knowledge continues to advance, more and more evidence should be presented to courts in an overtly statistical manner if the law is to keep up with new knowledge: “We are in a historic moment of horse-versus-locomotive competition, where intuitive and experiential expertise is losing out time and time again to number crunching”.

I have tried to show that statistical evidence (or number crunching) and intuition can be brought in line, but to do so requires proper presentation of such evidence, informed by good psychological research, and is greatly aided by some basic numerical education.

If the law does not wish to be left in the dust with horse-drawn carriages, it must hop onto the train and embrace the use of statistical evidence. I hope that my treatment of this topic has helped to show that this is both possible and welcome.

Bibliography

JURISPRUDENCE

Beazley v Suzuki Motor Corporation, 2010 BCSC 480 (CanLII).
City of Los Angeles Department of Water and Power v. Manhart (1978), 435 U S 702.
Froom v Butcher, [1975] 3 All E R 520.
People v Collins, (1968) 66 Cal Rptr 497.
R v JA, 2010 ONCA 226.
R v JAA, 2011 SCC 17.
R v Johnson, 2010 ABCA 230 (CanLII).
R v K (A) (1999), 45 O R (3d) 641 (CA).
Smith v Rapid Transit Inc (1945) 58 N E 2d 754.
Snushall v Fulsang, [2005] 78 OR (3d) 142.

MONOGRAPHS

Dawes, Robyn M, Everyday Irrationality (Boulder, CO: Westview, 2001) at 85-86.


Hume, David, *An Enquiry Concerning Human Understanding* (New York: Collier & Son, 1910) at section IX.


**ARTICLES**


Holmes Jr, Oliver Wendell, “The Path of the Law” (1897) 10 Harv L Rev 457.


Kahneman, David & Amos Tversky, "Subjective probability: A judgment of representativeness" (1972) 3 Cognitive Psychology 430.


———, “Two Techniques for Teaching the Estimation of Prior Probabilities” (200) 12:3 Teaching and Learning in Medicine 141.


