The Objective Bayesian Conceptualisation of Proof and Reference Class Problems

James Franklin*

Abstract

The objective Bayesian view of proof (or logical probability, or evidential support) is explained and defended: that the relation of evidence to hypothesis (in legal trials, science etc) is a strictly logical one, comparable to deductive logic. This view is distinguished from the thesis, which had some popularity in law in the 1980s, that legal evidence ought to be evaluated using numerical probabilities and formulas. While numbers are not always useful, a central role is played in uncertain reasoning by the 'proportional syllogism', or argument from frequencies, such as 'nearly all aeroplane flights arrive safely, so my flight is very likely to arrive safely'. Such arguments raise the 'problem of the reference class', arising from the fact that an individual case may be a member of many different classes in which frequencies differ. For example, if 15 per cent of swans are black and 60 per cent of fauna in the zoo is black, what should I think about the likelihood of a swan in the zoo being black? The nature of the problem is explained, and legal cases where it arises are given. It is explained how recent work in data mining on the relevance of features for prediction provides a solution to the reference class problem.

I Introduction: Bayesianism, Law and the Nature of Evidence

The encounter between Bayesian theory of probability and the law has been an unhappy one. The debate in the legal world in the 1980s and 1990s on Bayesianism and the law focused on cases like *Adams*[1] where excitable counsel urged jurors to pluck numerical prior probabilities of guilt out of the air and to literally apply a numerical version of Bayes' theorem to update them. Bayesianism was

* Professor, School of Mathematics and Statistics, The University of New South Wales.

1 *R v Denis Adams* [1996] 2 Cr App R 467 and *R v Denis Adams* (No 2) [1998] 1 Cr App R 377.

presented to lawyers as essentially a matter of using formulas with numbers in the way statisticians do.[2] A great number of reasons have been advanced as to why it is not a good idea to do that, including the impossibility of reliably eliciting any such priors and the computational infeasibility of calculating with them.[3] Those reasons are sound.[4]

Nevertheless Bayesianism has something fundamental to contribute to the law of evidence. It has two main points to make: first, to advance a view on what evidence is, and second, to recommend two main forms of inference, the confirmation of theories by their consequences, and the proportional syllogism.

Its main aim is to supply a theory of what evidence is. Evidence scholars ought to want to know what evidence is, for the same reason that ornithologists ought to want to know what birds are; and astrology ought to be concerned by claims that it is not about anything. Bayesianism has a story about what the subject-matter of evidence is — namely, logical relations between evidence and conclusion. It is of fundamental philosophical interest whether that is right — or if not, what alternative is right. And naturally, debates at the foundational level have consequences more widely, as we will see.

The (objective) Bayesian theory of evidence (also known as the logical theory of probability)[5] aims to explain what sort of thing evidence is. It holds that the relation of evidence to conclusion is a matter of strict logic, like the relation of axioms to theorems in mathematics but less conclusive. Given a fixed body of evidence — say in a trial, or in a dispute about a scientific theory — and given a conclusion, there is a fixed degree to which the evidence supports the conclusion.[6] It says, for example, that if we could establish just what the standard of 'proof beyond reasonable doubt' is, then, in a given

---

2 David H Kaye, 'Introduction: What is Bayesianism?' in Peter Tillers and Eric D Green (eds) *Probability and Inference in the Law of Evidence: The Uses and Limits of Bayesianism* (Kluwer Academic Publishers, 1988) 1.

3 Andrew Ligertwood and Gary Edmond, *Australian Evidence* (LexisNexis Butterworths, 5th ed 2010) [1.28]–[1.34]; Paul Roberts and Adrian Zuckerman, *Criminal Evidence* (Oxford University Press, 2nd ed, 2010) 153–63; Ronald J Allen, 'Rationality, Algorithms and Juridical Proof: A Preliminary Inquiry' (1997) 1 *International Journal of Evidence and Proof* 254; David Hodgson, 'Probability: The Logic of the Law: A Response' (1995) 15 *Oxford Journal of Legal Studies* 51.

4 Unlike those considerations that amount to 'Wouldn't it be awful if lawyers had to learn mathematics?'

5 Kaye, 'Introduction: What is Bayesianism?' above n 2, 4; fine distinctions considered in Darrell P Rowbottom, 'On the Proximity of the Logical and 'Objective Bayesian' Interpretations of Probability' (2008) 69 *Erkenntnis* 335.

6 Introductions in James Franklin, *What Science Knows: And How It Knows It* (Encounter Books, 2009) 5–21; J Maynard Keynes, *Treatise on Probability* (Macmillan 1921); Edwin T Jaynes, *Probability Theory: The Logic of Science* (Cambridge University Press, 2003); a fuller defence in Jon Williamson, *In Defence of Objective Bayesianism* (Oxford University Press, 2010).

trial, it is an objective matter of logical fact whether the evidence presented does or does not meet that standard, and so a jury is either right or wrong in its verdict on the evidence. Bayesianism thus provides an analysis of the notion of the concept 'epistemic' often used in the law of evidence, where epistemic purposes of evidentiary rules and practices are distinguished from non-epistemic (typically ethical) ones.

That view contrasts with, for example:

- Psychological views that deal only in people's actual degrees of belief in propositions—objective Bayesians are keen to emphasise the difference between what people in fact believe and what they ought to believe;
- Sociological views that people's degrees of belief are socially constructed (solely) on the basis of power relations, patronage and so on;
- 'Subjective Bayesianism', which allows one to have any degrees of belief one likes in propositions, provided the system is 'consistent', eg, that one's degree of belief in not-p is 1 minus one's degree of belief in p[7] (subjective Bayesianism led the revival of Bayesian statistics around the 1970s and 1980s, but Bayesianism has tended in a more objectivist direction since then);[8]
- Frequentism and propensity interpretations of probability, which believe that all probabilities are about relative frequencies (respectively physical propensities), and that there is no such thing as what one ought to believe on non-conclusive evidence; and,

In the legal context perhaps we should add:

- 'Know-nothingism', the view that it is all too deep for words and that Anglo-American law, reliant as it has always been on the bluff good sense of the English yeomanry, should avoid delving too deeply into matters that might lead to such horrors as ontology and metaphysics.

II Numerical or Non-Numerical Probabilities?

It is not essential to the Bayesian perspective that the relation of evidence to conclusion should be given a precise number, nor that it

---

7 Bruno de Finetti, *Theory of Probability* (Wiley, 1974); John Earman, *Bayes or Bust?* (MIT Press, 1992).

8 Stephen E Fienberg, 'When did Bayesian Inference Become "Bayesian"?' (2006) 1 *Bayesian Analysis* 1; S James Press, *Subjective and Objective Bayesian Statistics* (Wiley 2003).

be possible to compute the logical relation between evidence and conclusion in typical cases. It is sufficient for objective Bayesianism that it is sometimes intuitively evident that some hypotheses, on some bodies of evidence, are highly likely, almost certain, or virtually impossible.

The most central theses of Bayesianism do not concern numbers but are certain qualitative principles of evidence. The first is the simplest principle of logical probability, called by Polya 'the fundamental inductive pattern'[9] (and the main content of the celebrated Bayes' Theorem which gives Bayesianism its name). It is:

> *q* is a (non-trivial) consequence of hypothesis *p*
>
> *q* is found to be true
>
> So, *p* is more likely to be true than before
>
> (In short, 'Theories are confirmed by their consequences or predictions.')

It is hard to begin reasoning about the world without a commitment to this principle, as can be seen by trying to imagine a tribe that did not believe in it, and thought instead that agreement between theory and observation was a reason for *dis*believing the theory. (They guess there are bison in the river field and go there to hunt them. They find none. So they conclude they will probably find bison there tomorrow and the next day and they go there day after day with high hopes. You will need to imagine that tribe because you will not be meeting them. They are extinct.[10])

Many Bayesians do believe that there is in principle a number (between 0 and 1) expressing the (logical) probability P(h|e) of any given hypothesis h on any given body of evidence e; indeed, that might be called the most orthodox Bayesian position. But Keynes, whose *Treatise on Probability* of 1921[11] first clearly expressed objective Bayesianism, believed it was impossible to give every probability a number; certainly, it seems both impossible and pointless to debate precise numbers for, say:

---

9 George Polya, *Patterns of Plausible Inference* (Princeton University Press, 2nd ed, 1968) 4. The derivation of it from Bayes' Theorem is as follows. Let h stand for hypothesis, e for evidence. Then by Bayes' Theorem
P(h|e) = P(e|h) × P(h) / P(e)
Now if the evidence e is a consequence of the hypothesis h, then P(e|h) = 1. So
P(h|e) = P(h) / P(e)
If e is non-trivial, that is, not known with certainty already, then P(e) < 1. So P(h|e) equals P(h) divided by a number less than 1. So P(h|e) > P(h), that is, the hypothesis is more probable on the evidence than it was before.

10 Franklin, above n 6, 9.

11 Keynes, above n 6.

> P(Marilyn Monroe was murdered by the CIA | The moon is made of green cheese).

There is no logical relation between 'evidence' and conclusion, so there is no point looking for a number to express it.[12] More generally, in situations where the evidence is as imprecise as in most legal cases, the positing of an in-principle precise number does no real theoretical work; for example, if it is clear on a large mass of evidence that guilt is almost certain but that a precise numerical probability, even if it existed, could not be computed, one is perforce really working with an imprecise probability.

One might toy with the idea of accepting some minimal objective notion of evidential relevance, while hoping to avoid taking on the full superstructure of Bayesian theorems. That is a vain hope — something like trying to accept mathematicians' advice on the addition of single-digit numbers while avoiding their excessively complex procedures for adding double-digit numbers. Once one has accepted the confirmation of theories by their consequences, one will be hard put to avoid accepting the next theorem, Polya's 'verification of an improbable consequence':

> *q* is a (non-trivial) consequence of hypothesis *p*
>
> *q* is very improbable in itself
>
> *q* is found to be true
>
> So, *p* is much more likely to be true than before[13]

After agreeing to sufficiently many such intuitively plausible theorems, one is on a slippery slope to full-blooded objective Bayesianism. There will be little motivation for wishing to avoid the more complex theorems, whose acceptance will make one a true believer.

III Why Believe in Objective Bayesianism?

According to logicians, the core of deductive logic does not need to give reasons for belief in itself. To give reasons already supposes logic. There is no non-circular justification for modus ponens, say, or the other basic principles of deductive logic.[14] The core of logic, at least, is just necessarily true, self-justifying, and in simple cases obvious. That is no less true when the logic is probabilistic than it is in the deductive case.

---

12 James Franklin, 'Resurrecting Logical Probability' (2001) 55 *Erkenntnis* 277.

13 Polya, above n 9, 8.

14 Susan Haack, 'The Justification of Deduction' (1976) 85 *Mind* 112.

Further, everyone appears to accept probabilistic reasoning from (approximate) frequencies in ordinary life. The reason for being fairly relaxed as one's plane takes off is just the knowledge that the vast majority of planes that take off land safely. That is to exhibit confidence in the proportional syllogism (of which more below):

> P(this flight will land safely | the vast majority of flights land safely) is high

To call that a kind of 'syllogism', of course, is to draw attention to its parallel with the ordinary syllogism of deductive logic:

> If all flights land safely and this is a flight, then this will land safely
>
> Or in the language of probability,
>
> P(this flight will land safely | 100% of flights land safely) = 1

Surely that parallel is a good one. Merely saying, as some people overtrained in deductive logic do, that logic ought to be restricted by definition to cases where the conclusion follows with certainty from the premises, is unhelpful, as it tries to do away with the cases at hand by verbal stipulation.

A further reason for believing that the principles of probabilistic reasoning are *logic* — true in all possible worlds — is that they work perfectly well with evidence for and against conjectures in pure mathematics, such as the Riemann Hypothesis. But in mathematics, there are no contingent principles such as the 'uniformity of nature' that might be thought to be needed to underpin probabilistic reasoning. There is only the conjecture and the evidence for it, and the relation between them can be nothing but logical.[15]

IV What does Objective Bayesianism do for the Theory of Legal Evidence?

By virtue of giving a foundational story on what evidence is, objective Bayesianism supplies an answer to attempts to undermine the credibility of the legal process on the basis of alternative, mistaken, views of the relation of evidence to conclusion — for example, views that the relation is 'constructed' by power relations. If mathematicians are faced with views such as that 2 + 2 = 4 is only true because of a patriarchal commitment to binary oppositions, it is water off a duck's back to them because they understand the necessity of why 2 + 2 *must* be 4. Acquaintance with such a competing view

[15] James Franklin, 'Non-deductive Logic in Mathematics' (1987) 38 *British Journal for the Philosophy of Science* 1.

will merely increase mathematicians' determination to act politically to outnumber educationists on syllabus committees. In our age of relativist currents, those defending the credibility of a system of evidence evaluation need a story as to why there is objectivity at the bottom of it. Ideally, a true story.

That is not to say that the conclusion of the exercise is necessarily conservative. As with the theory of the objectivity of ethical rights, it may turn out that current practice is in some respects not in accordance with the objective rules. That was indeed the case with witness identification evidence, where it appeared that such evidence is objectively less credible than it was normally taken to be in courts, and steps were taken to increase courts' awareness of the problem.[16] Such concerns only provide a reason for reform if one has an objective conception of evidence like the Bayesian one, according to which there can be a mismatch between court practices of belief and what the court should, as a matter of objective fact, believe on the evidence.

The use of objective Bayesianism in setting or underpinning standards applies also to justifications of exclusionary rules of evidence. When the *Evidence Act* says:

> The court may refuse to admit evidence if its probative value is substantially outweighed by the danger that the evidence might
>
> (a) be unfairly prejudicial to a party; or
>
> (b) be misleading or confusing[17]

or:

> In a criminal proceeding, the court must refuse to admit evidence adduced by the prosecutor if its probative value is outweighed by the danger of unfair prejudice to the defendant[18]

the Bayesian has a clear story about what it means. 'Probative value' is the logical degree to which the evidence really does support the conclusion, while 'prejudice', 'misleading' and 'confusing' are ways in which people's degrees of belief might deviate from what they ought to be, on the evidence. The possibility of — indeed the persistent and predictable tendency towards — deviation between

---

16 Gary L Wells and Elizabeth A Olson, 'Eyewitness Testimony' (2003) 54 *Annual Review of Psychology* 277; Brian L Cutler and Steven D Penrod, *Mistaken Identification: The Eyewitness, Psychology, and the Law* (Cambridge University Press, 1995). Note that the basic reason for doubting eyewitness identification evidence is a straightforward statistical syllogism: Many convictions on eyewitness evidence have turned out to be false (on DNA evidence etc), so eyewitness identification evidence is unreliable.

17 *Evidence Act 1995* (Cth) s 135.

18 Ibid s 137.

logic and psychology is the reason for having the exclusionary rule.[19] It is hard to see an alternative theory of evidence substantially different from the Bayesian one making sense of that.

Those considerations apply equally well to the reasoning of juries in actual trials, properly conducted under the exclusionary rules. Bayesianism is normative as to errors of reasoning, and suggests there is cause for concern if jurors, like other human evidence evaluators, are systematically flouting the laws of probability. The legal world does not normally report trials in sufficient detail to make it possible to decide if jurors are misevaluating evidence — even if there is a transcript of everything said in court, it does not reveal the sequence of jurors' thought processes. Psychological research on mock jurors, however, suggests there are certain systematic violations of the logical principles of evidence evaluation. For example, Carlson and Russo reported that mock jurors interpret evidence so as 'to support whichever verdict is tentatively favored as a trial progresses…distortion increased with juror confidence in whichever verdict was currently leading'.[20] That, or any other distortion that means that the verdict reached can depend on the order of presentation of the evidence, is contrary to logical principles. Bayesians have a standpoint to criticise it. Those who refer evidence evaluation to the unreconstructed wisdom of the 'reasonable man' will have more difficulty in explaining what is wrong with it.[21]

V Are Evidence Evaluators Getting it Right?

There are certainly many questions as to whether naïve evidence evaluators, such as juries, judges, detectives, medical diagnosticians, political pundits and so on are evaluating evidence well. 'Naïve' here means not so much lacking in experience of making judgments on evidence, but lacking in training in statistical methods and the mathematics of Bayes' theorem, and mostly lacking also in punitive feedback on the results of mistakes. There is a great deal of work showing that 'naïve experts', if that is the right term, make systematic errors in evaluating evidence, from Kahneman and Tversky's classic psychological work of the 1970s[22] to Tetlock's recent work on the

---

[19] Supposing that protecting the jury from its tendencies to error is indeed the main reason for exclusionary rules: see debate in Lisa Dufraimont, 'Evidence Law and the Jury: a Reassessment', (2008) 53 *McGill Law Journal* 199.

[20] Kurt A Carlson and J Edward Russo, 'Biased Interpretation of Evidence by Mock Jurors' (2001) 7 *Journal of Experimental Psychology: Applied* 91, 91.

[21] Further applications in Alvin I Goldman, 'Quasi-objective Bayesianism and Legal Evidence' (2002) 42 *Jurimetics Journal* 237.

[22] Daniel Kahneman, Paul Slovic and Amos Tversky (eds) *Judgment Under Uncertainty: Heuristics and Biases* (Cambridge University Press, 1982).

errors of political 'experts'[23] — and indeed, that such errors are often resistant to training. On the other hand, the human brain, and for that matter the rat brain, is good at evaluating risks in many circumstances — otherwise there would not be any brains left — and five-year-olds and laboratory rats can often make the right decision on the basis of relative frequencies.[24]

What does 'right' mean, when we ask 'are evaluators getting it right?' The Bayesian perspective has an answer. It says, for example, that if one observes a consequence of a theory, one is correct in believing the theory more than before (not less), and that if the great majority of a pundit's predictions have proved false, it would be unwise to believe the next one. Other perspectives on evidence, for example sociological or frequentist ones, have tended to eschew pronouncements on what is objectively right, but it is impossible to avoid doing so when undertaking serious research on errors of judgment. Yet the concept of 'error of judgment' is clear enough in this research — for example, a misperception of risk on the basis of past evidence that will lead to dangerous decisions — so the credentials of the research are sound.

VI Case Study: Quantifying the 'Proof Beyond Reasonable Doubt' Standard

Should there be an effort to quantify in any way, perhaps with an imprecise number, the 'proof beyond reasonable doubt' standard of criminal law?

Suppose the judge in a criminal trial has directed the jury correctly on the need for it to reach 'proof beyond reasonable doubt' or some equivalent formulation. If the jury foreman returns during the jury's deliberation and asks the judge, 'Is 60 per cent okay?" the judge has a choice of three possible answers: 'Yes', 'No' and 'I am not going to tell you'. The latter answer is normally regarded as legally appropriate.[25] Nevertheless it is a problematic answer, since if the jury, lacking any assistance from the judge, were to convict on what it took to be a probability of guilt of 60 per cent, it would seem to have departed substantially from the normal and plain meaning of 'beyond reasonable doubt' and to have perpetrated an injustice.

---

[23] Philip Tetlock, *Expert Political Judgment: How Good Is It? How Can We Know?* (Princeton University Press, 2005).

[24] Vittorio Girotto and Michel Gonzalez, 'Children's Understanding of Posterior Probability' (2008) 106 *Cognition* 325; John H Holland et al, *Induction* (MIT Press, 1986) [5.2].

[25] Richard Eggleston, *Evidence, Proof and Probability* (Butterworths, 2nd ed, 1983) 114.

There are a number of objections to any quantification of the standard (over and above any distaste for numbers as such that may be endemic in the legal profession due to, for example, the lack of statistics courses in law degrees). The Bayesian perspective introduces clarity into the debate about these objections by making a clear distinction between objections based on ethical or policy considerations and objections based on conceptual problems about probability.

From the direction of policy, ethics and psychology, the problems raised include:

- There may be different standards appropriate to different cases, for example a higher standard where the punishment is heavier;
- The jury is properly left to decide the standard in the light of the facts of the particular case;
- Since there is in fact considerable disagreement as to the correct numerical value of the standard, attempts to standardise it will create only confusion, evasions and a façade of uniformity where there is no true consensus; and
- The majesty of the law and its powers of deterrence would be ill-served if the law were forced to admit the truth about the number of false convictions it allows and the number of criminals it allows to go free.

Quite different objections arise from certain, more conceptual, problems about the nature of probability:

- Some probabilities may be inherently incapable of being given a precise number;
- Evidence suitable for conviction should be 'substantial' or 'weighty', and a numerical probability expresses only the balance between favourable and unfavourable reasons, not whether those reasons are substantial; and
- A numerical standard will tend to draw attention to evidence that is quantified and logically relevant but legally inadmissible, such as proportions in reference classes containing the defendant.

Those are all substantial reasons, though none have much force when it comes to allowing a jury to convict on a probability of 60 per cent if it sees fit.[26]

---

[26] James Franklin, 'Case comment: United States v. Copeland, 369 F. Supp. 2d 275 (E.D.N.Y. 2005): Quantification of the "Proof Beyond Reasonable Doubt" Standard'

VII Frequencies and the Proportional Syllogism

While Bayesians do not insist that the relation of evidence to conclusion should always be numerical, they maintain that in many important cases there are indeed numbers expressing that relation. A paradigm is the probability assignment, sometimes called the 'proportional syllogism'[27] or 'statistical syllogism'[28] or argument from frequencies, such as:

> P(Tex is rich | Tex is a Texan and 90% of Texans are rich)
> = 0.9

Some care is needed at this point. It is initially natural to object: 'What if Tex were a philosopher? He'd hardly be likely to be rich then'. Of course that is true, but the number 0.9 is about the relation of the conclusion to the *given* body of evidence, 'Tex is a Texan and 90 % of Texans are rich', not to some *other* body of evidence, such as 'Tex is a Texan and a philosopher and 90 % of Texans are rich.' Bayesianism maintains that probability is a *relation* between evidence and conclusion — so, different bodies of evidence, different relation. This is what makes it hard to apply the Bayesian perspective in any mechanical way to the evaluation of real court cases where the jury should use as part of its evidence its general knowledge of the way the world is and how people normally behave. Formalising 'the commonsense knowledge of the reasonable man' is impossible — or at least the artificial intelligence community have been promising to do it for 50 years and have got almost nowhere, so it is not likely to be available on a CD Real Soon Now.

VIII Legal Relevance of the Statistical Syllogism

To say that the statistical syllogism is a good argument as a matter of logic is not to say that it is always a good argument as a matter of law (or of ethics). The problem is illustrated by Jonathan Cohen's 'gatecrasher paradox': Suppose 499 tickets to a rodeo have been sold, and 1000 persons are observed on the stands. If that is the sole evidence, is the rodeo owner entitled to judgment against each of the

---

(2006) 5 *Law, Probability and Risk* 159; it is true that in cases of reverse burdens, there may be conviction on a lower probability, but such cases appear to deny that 'proof beyond reasonable doubt' is the appropriate standard rather than changing the meaning of 'proof beyond reasonable doubt; see David Hamer, 'The Presumption of Innocence and Reverse Burdens: A Balancing Act' (2007) 66 *Cambridge Law Journal* 142.

27 Peter Forrest, *The Dynamics of Belief* (Blackwell, 1986) ch 8.

28 Carl G Hempel, *Aspects of Scientific Explanation* (Free Press, 1965) ch 2; Merrilee H Salmon, *Introduction to Logic and Critical Thinking* (Thomson Wadsworth, 3rd ed, 1995) 99–100; William Gustason, *Reasoning from Evidence: Inductive Logic* (Macmillan, 1994) 49–51.

attendees for the cost of admission on the basis of this statistical syllogism?:

> 501 of the 1000 attendees are gatecrashers
>
> A is an attendee
>
> Therefore, on the balance of probabilities, A is a gatecrasher[29]

The normal legal answer is 'no'. The analysis of such cases has been much debated. One possibility is that the law regards it as unfair to individuals to reach decisions solely on the basis of statistical evidence, and requires that there should be at least some evidence that somehow bears directly on, or is caused by, the individual case — is 'case-specific'.[30] If that or something similar is the correct analysis, it simply means that the argument is correct as a matter of probabilistic inference, but the law chooses to require more for the sake of justice. Which, of course, it is free to do since the purposes of the law of evidence, like any other part of the law, include non-epistemic ones such as the respecting of rights and justice, as well as the epistemic one of reaching the probable truth.[31]

It is somewhat better established that the rule of exclusion of similar fact evidence in criminal trials is a matter of policy or justice rather than of probability. The argument:

> The accused has committed several robberies
>
> The crime at hand is a robbery
>
> Therefore, the accused is more likely than a random person to have committed it

is admitted to be probative as a matter of logic, and its exclusion is based on a moral consideration about its potential to unreasonably influence the jury to the prejudice of the defendant. Similar fact evidence is generally considered to be admissible 'provided it possesses sufficient probative value to outweigh the risk of

---

[29] L Jonathan Cohen, *The Probable and the Provable* (Clarendon Press, 1977) 74; David Kaye, 'The Paradox of the Gatecrasher and Other Stories' [1979] *Arizona State Law Journal* 101; Stephen E. Fienberg, 'Gatecrashers, Blue Buses and the Bayesian Representation of Legal Evidence' (1986) 66 *Boston University Law Review* 693.

[30] Alex Stein, *Foundations of Evidence Law* (Oxford University Press, 2005) 70–2, 79; William L Twining and Alex Stein (eds), *Evidence and Proof* (New York University Press, 1992) xxi–xxiv; the difficulties of explaining the matter in non-ethical, purely decision-theoretic terms are exemplified in Daniel Shaviro, 'Statistical-probability Evidence and the Appearance of Justice' (1989) 103 *Harvard Law Review* 530.

[31] Hock Lai Ho, *A Philosophy Of Evidence Law: Justice in the Search for Truth* (Oxford University Press, 2008).

prejudice'[32] — a saying that makes little sense without a robustly objectivist view of the probabilities involved in both 'probative value' and 'risk of prejudice'.

IX Reference Class Problems

Legal theorists have had to consider the proportional syllogism because of the natural occurrence of what has come to be called in law and philosophy the 'reference class problem'. It is all very well to argue

> P(Tex is rich | Tex is a Texan and 90% of Texans are rich) = 0.9

But what if Tex is a member of several classes, with differing frequencies of wealth? Suppose the evidence is

> Tex is a Texan philosopher and 90% of Texans are rich and 10% of philosophers are rich[33]

What should one then think about the probability of Tex being rich?

In general, as Venn pointed out in the nineteenth century, 'It is obvious that every individual thing or event has an indefinite number of properties or attributes observable in it, and might therefore be considered as belonging to an indefinite number of different classes of things', leading to a general problem as to how to assign probabilities to a single case on the basis of frequencies, for example the probability that John Smith, a consumptive Englishman aged 50, will live to 61.[34]

Reichenbach gave the name 'reference class problem', arguing:

> If we are asked to find the probability holding for an individual future event, we must first incorporate the event into a suitable reference class. An individual thing or event may be incorporated in many reference classes, from which different probabilities will result. This ambiguity has been called the *problem of the reference class.*[35]

Plainly, the problem will appear in any case where there may be doubt as to what class containing an instance is most relevant to determining its probability of having some attribute. Philosophers, as is

---

[32] Eg, *Evidence Act 1995* (Cth) ss 97, 101; David Hamer, 'Similar Fact Reasoning in *Phillips*: Artificial, Disjointed and Pernicious' (2007) 30 *University of New South Wales Law Journal* 609.

[33] Example from Stephen F Barker, *Induction and Hypothesis* (Cornell University Press, 1957) 76.

[34] John Venn, *The Logic of Chance* (MacMillan 1866) 176.

[35] Hans Reichenbach, *The Theory of Probability* (University of California Press, 1949) 374.

their way, have written at length on the ubiquity and difficulty of the problem, without offering a solution.[36] Artificial intelligence researchers on commonsense reasoning have also come across the problem, and they too have found it intractable.[37]

In law, it has come to be appreciated that any case involving statistical evidence could be infected by reference class problems, leading to potentially endless argument between counsel on the relevance of different classes that include the case at hand.[38] The 'prosecutor's fallacy' is in the first instance a reference class problem (though there is more to the fallacy than that). The prosecutor invites the jury to consider the proportional syllogism:

> The vast majority of innocent people do not match the perpetrator (in DNA or whatever characteristic has been identified)
>
> So the defendant, if innocent, would very probably not match the perpetrator (hence, as he does match the perpetrator, he is probably not innocent)[39]

A defence against this fallacy involves looking at another reference class, that of all persons matching the perpetrator. In that class, the proportion of innocent people may not be low. Base rate fallacies more generally involve neglect of proportions in a reference class that is in fact relevant to the problem and needs to be combined with more specific information.[40]

Another example is the much-discussed *Shonubi* case. In *United States v Shonubi*,[41] sentencing guidelines required an estimate of how much heroin Charles Shonubi, a Nigerian drug smuggler, had carried through New York's John F Kennedy Airport ('JFK') on seven previous trips during which he had been undetected. The estimate was

---

[36] Henry E Kyburg, 'The Reference Class' (1983) 50 *Philosophy of Science* 374; Alan Hájek, 'The Reference Class Problem Is Your Problem Too' (2007) 156 *Synthese* 563; Mark Colyvan, Helen M Regan and Scott Ferson, 'Is it a Crime to Belong to a Reference Class?' (2001) 9 *Journal of Political Philosophy* 168.

[37] Raymond Reiter and Giovanni Criscuolo, 'On Interacting Defaults', *Proceedings of the 7th International Joint Conference on Artificial Intelligence* (1981) 270.

[38] See articles in 'Special Issue on The Reference Class Problem' (2007) 11 *International Journal of Evidence and Proof* 242.

[39] The problem is often put in terms of conditional probabilities: P(match|innocent) is low, which does not imply that P(innocent|match) is low. That is essentially equivalent to the above, since P(match|innocent) is low because of the low proportion of matches in the class of innocents, that is, 'The vast majority of innocents do not match'; while P(innocent|match) refers to the proportion of innocents in the class of matchers, which may not be low.

[40] Maya Bar-Hillel, 'The Base-Rate Fallacy in Probability Judgments' (1980) 44 *Acta Psychologica* 211.

[41] 895 F Supp 460 (EDNY 1995), discussed in Peter Tillers, 'If Wishes Were Horses: Discursive Comments on Attempts to Prevent Individuals from Being Unfairly Burdened by Their Reference Classes' (2005) 4 *Law, Probability & Risk* 33.

based on the average amounts of heroin found on Nigerian drug smugglers caught at JFK airport in the time period. Why should that be used as the reference class relevant to the case rather than, say, George Washington Bridge tollbooth collectors (Shonubi's day job)? Or, take a more typical case involving valuation: Valuing a house for sale involves estimating its price from the sale records for 'similar' houses. No other house is exactly similar to the given one, so how widely or narrowly should one choose the reference class of 'similar' houses, and on what criteria? Number of bathrooms? Age? Street number? Ethnicity of owner?

X Solution to the Reference Class Problem

There is some consensus that the reference class problem is inherently unsolvable; that 'there is no principled way to establish the relevance of a reference class'.[42] That is hard to believe. Human life, and for that matter animal life, requires continual judgments of risk on the basis of frequencies — the risk of lions behind rocks, of being waylaid on the way to the shops, of rejection of tenure, and so on. To stay alive and in the game, one must evaluate a good proportion of risks well, which is impossible if one cannot distinguish the few relevant reference classes from the many irrelevant ones. We solve reference class problems every day. Surely it is possible to say how.

The solution is, in principle, straightforward. In summary:

- A reference class is defined by its features (for example, the houses in Centreville of a given age and number of bathrooms), so the problem reduces to explaining the relevance of features;
- For statistical evidence, *relevance is co-variation*: a feature A (such as 'age') is relevant to a prediction B (such as 'value') if A and B co-vary (or are correlated); and
- The ideal reference class for an outcome B is the class defined by the intersection of all the features relevant to B.

To clarify: One must distinguish between a set or class — the actual members, such as houses — and the features defining it such as 'houses in Centreville, 30 years old with two bathrooms'. It is the members of the class that must be counted but it is the defining features that are or are not relevant to prediction.

---

[42] Mark Colyvan and Helen M Regan, 'Legal Decisions and the Reference Class Problem' (2007) 11 *International Journal of Evidence and Proof* 274, 275.

The essential idea of the solution is that the relevance of one feature to the prediction of another is defined by co-variation, that is, the 'ons' of one feature by and large go with the 'ons' of the other. Why is the colour of traffic lights relevant to the decision to drive across the intersection? Because when the light is green it is safe to drive and when it is red it is not safe (almost always).[43] That is, colour of traffic light co-varies with safety of driving across the intersection. But the colour of the car ahead does not co-vary with safety, so there is no point in attending to that feature when deciding whether to drive across the intersection.

The usual measure of co-variation of two features is the correlation coefficient. That is what is normally used in the area of recent statistics where deciding on the relevance of features has been most intensively studied, the process called 'feature selection' in data mining, also known as 'variable selection' or 'attribute selection;.[44] A database is organised into many rows (the cases) and columns (the fields, attributes, properties, or features of the cases). Data mining deals with very large databases: hundreds of rows for real estate but millions for many kinds of health and gene data and financial records, with possibly thousands of features (columns). In such large cases, the great majority of features are expected to be irrelevant to the task of prediction. For example, not every feature or measurement in a gene database will be helpful in predicting cancer, and most features of financial records will be irrelevant to determining creditworthiness. The more features in a database, the harder it is to evaluate each feature's relevance. The aim of feature selection methods is to determine from large amounts of data which of the many properties or features of the individual cases are relevant to a given classification or prediction task. A feature is relevant if it gives some information about the outcome — for example, 'number of bathrooms' makes some difference to 'house price' in the sense that, on average, a different number of bathrooms goes with a different house price. Relevance is correlation. There is a standard definition of correlation and there are some alternative

---

43 Ronald K Templeton and James Franklin, 'Adaptive Information and Animal Behaviour: Why Motorists Stop at Red Traffic Lights' (1992) 10 *Evolutionary Theory* 145.

44 See generally Avrim L Blum and Pat Langley, 'Selection of Relevant Features and Examples in Machine Learning' (1997) 97 *Artificial Intelligence* 245; Isabelle Guyon and André Elisseeff, 'An Introduction to Variable and Feature Selection' (2003) 3 *Journal of Machine Learning Research* 1157; Mark A Hall and Geoffrey Holmes, 'Benchmarking Attribute Selection Techniques for Discrete Class Data Mining' (2003) 15 *IEEE Transactions on Knowledge & Data Engineering* 1437; Patricia E N Lutu and Andries P Engelbrecht, 'A Decision Rule-Based Method for Feature Selection in Predictive Data Mining' (2010) 37 *Expert Systems with Applications* 602.

measures of association to choose from, but they are all intended to measure the degree to which one variable 'goes with' another.[45]

Once the relevant features for a prediction (such as house price) have been identified, it is clear what the relevant reference class for a case is. It is the class of items that agree with the case in all relevant features.[46] If being Nigerian, being a drug mule, being at JFK, and being in the time period are all reasonably believed to be relevant to the amount of drugs smuggled, and there is no evidence that any other feature on which data is available is relevant, then the ideal choice of reference class for Mr Shonubi is the class of Nigerian drug smugglers at JFK in the time period.[47]

There is a potential problem with this ideal choice of reference class: maybe the set defined by the intersection of all relevant features is too small to be usable. It is usable if there is a sufficiently large number of cases in it for a reliable estimate of the target. A data set that is too small, or perhaps contains only the single original case, will not support reliable estimates since there is too much chance involved in which few cases happened to land in the set.[48] To know whether the data is enough to ensure reliability of the estimate, one consults standard statistical theory on the variance or standard deviation of the estimate in question.[49]

If the ideal choice of reference class is too small for reliable inference, one must use larger classes defined by some of the attributes and 'trade off' the results of each. That is a difficult problem, typified by the case above of Tex, the potentially rich Texan philosopher. Work on this problem proceeds.[50]

---

45 Further in James Franklin, 'Feature Selection Methods for Solving the Reference Class Problem: Comment on Edward K Cheng, "A Practical Solution to the Reference Class Problem"' (2010) 110 *Columbia Law Review Sidebar* 12.

46 'Relevance' of features briefly suggested in Gustason, above n 29, 50.

47 The reasonable belief on the relevance of those features could be based on explicit counts of data in a database, but it could also be based on general knowledge gained through normal experience of life; such prior beliefs can be subject to normal cross-examination in court.

48 Reichenbach said when coining the phrase 'reference class problem' that it should be 'the narrowest class for which reliable statistics can be compiled', which is correct when such a class can be defined, except that one does not narrow a relevant reference class by splitting it according to irrelevant attributes. Reichenbach, above n 36; other problems for Reichenbach's solution in Hájek, above n 37, 156.

49 See also Franklin, above n 46.

50 Ibid.