THE CRIMINAL JUSTICE SYSTEM AS A PROBLEM IN BINARY CLASSIFICATION†

William Cullerne Bown*

Attempts to establish a quantitative framework for thinking about the criminal justice system have been made at least since Kaplan's influential 1968 article. Here I avoid the probabilistic approaches that Kaplan inspired and instead characterize the law's underlying problem as one of measurement. I then exploit statistical techniques developed in recent years in other disciplines to evaluate systems that also face the challenge of "binary classification" to solve it. This approach entails the mathematization of the criminal justice system's core epistemic concern of distinguishing the guilty from the innocent with Van Rijsbergen's F-measure and empirical measurements of effectiveness. Once one adopts the perspective of a sovereign, it yields a meta-meta-epistemology that allows traditional arguments like those that refer to Blackstone's ratio to be made rigorous. This provides a clearer relationship between values and policies and, in a narrowly epistemic sense, a complete answer to questions of evidence and procedure.

If we define *n* as the ratio of false acquittals to false convictions and allow X to be a policy in criminal law, then we can see that arguments in the form "because *n*, X" have been important in Anglo-Saxon jurisdictions since Blackstone. Consider for example Blackstone's dictum, that "...the law holds that it is better that ten guilty persons escape than that one innocent suffer". This is part of a longer statement in the *Commentaries*, insisting that presumptive evidence should be admitted only cautiously and, in particular, that two rules of thumb should be adhered to, one of which is to never convict of murder or manslaughter unless the body can be produced. Thus the dictum is evidently part of a longer statement that is in the form "because *n*, X". [1]

---

† William Cullerne Bown, *The Criminal Justice System as a Problem in Binary Classification*, International Journal of Evidence and Proof. http://journals.sagepub.com/home/epj. Submitted 19 April 2018; accepted for publication and scheduled to appear in October 2018. This is the version accepted for publication can be downloaded from my personal webpage at https://quantitativejurisprudence.com.

* https://quantitativejurisprudence.com.

I would like to thank Clive Freeman, David Mond and Meade McCloughan for early and patient encouragement, advice and criticism. Without them this article would not have been written. Larry Laudan, Cornelis van Rijsbergen, Michael DeKay, David Hand, Paul Edelman, Martin McQuillan, Joe Mazur and Federico Picinali were kind enough to read lengthy drafts and

Although his preference for *n* is vaguely defined, Justice Harlan's concurring opinion concerning the standard of proof in *In re Winship* in 1970 took the same form:

> "I view the requirement of proof beyond a reasonable doubt in a criminal case as bottomed on a fundamental value determination of our society that it is far worse to convict an innocent man than to let a guilty man go free."[2]

Today, Epps reports that the same kind of reasoning is widely relied on in the United States to justify many of the most fundamental policies of the criminal justice system. He articulates the reasoning in the form of what he calls "the Blackstone principle":

> "Blackstone's ten-to-one ratio and its variations can't be taken literally. There's no way to measure the exact ratio between the false convictions and false acquittals our system creates, and no one seriously advocates that it is critical to strive for exactly ten false acquittals for every false conviction. Instead, the ratio serves as shorthand for a less precise – but still important – moral principle about the distribution of errors: we are obliged to design the rules of the criminal justice system to reduce the risk of false convictions – even at the expense of creating more false acquittals and thus more errors overall."[3]

The starting point is again Blackstone's ratio; the conclusion, although lying outside this text itself, we understand is a range of policies, or Xs.

The importance of this kind of reasoning is suggested by Shapiro's account of the crisis in English law in the early modern period when it lost access to divine insight through the medium of Christian conscience. The sticking point then was proof, and a vital part of the response was the development of policies that allowed the law to convince society at large that the jury retained the divine spark. [4] "Because *n*, X" arguments assist in this task by providing a kind of justification for policies that has one foot in a modern, quantitative sensibility that was emerging then and, outside the law, has become ever more central.

The question is, by what steps do we get from *n* to X? For example, from Harlan, all we get is "bottomed on". Since Kaplan, attempts have been made to provide a formal answer, at least

---

offer insightful comments. Oliver Noble was a valuable source for analysis discussions and the raised eyebrow of my wife Alice helped me cut through many confusions. All mistakes are my own.[1] William Blackstone, *Commentaries* *352.

[2] *In re Winship*, 397 U.S. 371 (1970).

[3] Daniel Epps, *The Consequences of Error in Criminal Justice,* 128 Harv. L. Rev., 1073 (2015).

[4] Barbara J. Shapiro, *"Beyond Reasonable Doubt" and "Probable Cause": Historical Perspectives on the Anglo-American Law of Evidence*, University of California Press (1991), 241.

with regards to the standard of proof, through probabilistic methods. A useful summary is provided in Walen's recent account of what he calls the "consequentialist" approach.[5]

There are two important points about all this work. First, consideration of all four possible outcomes in Walen's equation 5 – a true positive (rightful conviction), false positive (false conviction), true negative (rightful acquittal) or false negative (false acquittal) – is narrowed to just the false negatives and false positives (equation 6).[6] This *narrowing move* can be achieved by assuming that the true positives and true negatives have either no value or the exact opposite value of their false counterparts.[7] The justifications given for this simplification vary from author to author. Kaplan's original paper, still cited without comment by for example Stein, considered the move of no substantive consequence.[8] [9] Walen, who in his turn makes the same move, both describes principled reasons and alludes to the difficulty of making the problem tractable otherwise.[10] Second, the end result is that the standard of proof is to be set in order to achieve a pre-determined ratio of the risk of false negatives to false positives, an objective that often now goes under the heading of "the distribution of errors".

However, DeKay already concluded in 1996 that, "To the extent that jurors', judges' and legal scholars' notions of correct standards of proof are based on desires to bring about particular error ratios, such notions are founded on presumptions that are fundamentally invalid."[11] DeKay's argument remains unrefuted and, for separate reasons that this article is too short to contain, I consider his conclusion correct.[12] I therefore believe that the law continues to lack an adequate quantitative – that is to say, modern – basis for the "because *n*, X" arguments that have been of such central importance. Although DeKay and I arrive at such a view through purely statistical reasoning, the conclusion itself is today unremarkable in jurisprudence. With its explicit recourse to morality to bridge the gap in reasoning between

---

[5] Alec D. Walen, 76(2) *Proof Beyond a Reasonable Doubt: A Balanced Retributive Account*, Louisiana L. Rev., 355-446 (2015).

[6] Walen *supra* note 5, at 359. This yields an equation in a form that will be familiar to anyone who has read Kaplan or the work of his followers: $SOP = 1/[1 + V_{AG}/V_{CI}]$, where V_{AG} is the value of acquitting the guilty, and V_{CI} the value of convicting the innocent.

[7] Michael DeKay, *The Difference between Blackstone-Like Error Ratios and Probabilistic Standards of Proof*, 21 L. & Social Inquiry 116 (1996).

[8] Alex Stein, *Foundations of Evidence Law*, Oxford University Press (2005), 172.

[9] What Kaplan says is: "For convenience we will deal not directly with utilities but disutilities, since the problem is more easily phrased in terms of avoiding certain consequences than in terms of achieving others.". John Kaplan, *Decision Theory and the Fact-finding Process*, 20 Stanford L. Rev., 1071 (1968).

[10] Walen *supra* note 5, at 407.

[11] DeKay *supra* note 7, at 132.

[12] See *Killing Kaplan and Kaplanism*, in preparation.

n and X, the Blackstone principle denies the need for a quantitative framework, but also acknowledges the lack of it. Epps argues that the Blackstone principle lacks justification and even some of his opponents are unwilling to come to its defence, so that Appleman describes it as a "hoary old Blackstonian koan".[13] Here I provide a new way to get from *n* to X that starts with a characterization of the law's underlying problem as one of measurement.

When we talk of measuring something, we usually think of making observations. However, the observations are meaningless if they are not embedded in a computational framework that allows them to be evaluated. Often this framework is so simple that we don't even notice it. For example, suppose you are investing $10 and with god-like insight know that policy X would result in the first set of results in the table below and policy Y the second set. Which outcome is better?

	$
X	100
Y	1000

Figure 0.1

Answer: Y. Furthermore, we have a general rule that we can apply to any two options: the bigger number is better.

Now let us turn to the law. Suppose you are trying cases and with god-like insight know that policy X would result in the first set of results in the table below and policy Y the second set. Which is better? More importantly, what is the rule that would allow you to decide between *any* two sets of results?

	True Positive	**False Negative**	**True Negative**	**False Positive**
X	560	1300	7913	385
Y	288	1572	8197	101

Figure 0.2

We don't know. Clearly the law would have great difficulty in making suitable observations, but this shows that its most fundamental difficulty is deeper, the lack of a computational framework for evaluating such observations. Tribe precisely delineated this problem in 1971,

[13] Laura I Appleman, *A Tragedy of Errors: Blackstone, Procedural Asymmetry, and Criminal Justice*, 128 Harv. L. Rev. F. 91, (2015).

and there it has rested.[14] The main result of this article is to provide a solution to this problem by drawing on statistical techniques developed after he wrote his paper. To be precise, given a value for *n*, we can always determine whether X or Y is the better policy.

Most of the policy questions at stake in criminal law come in the form of a dilemma in which we would like to reduce the false negatives but fear increasing the false positives. The central question then is how to handle the trade-offs involved. We may surmise that this is what makes the topic difficult. But if we can glimpse, as Blackstone perhaps did, that there is intrinsically a kind of numbers game in here, then the starting point Tribe has given us is that we still do not know what sort of game it is. What we might hope for is a way to generate a single number for each policy option that, like an amount of money, would then allow us to order the options from best to worst. To a mathematician this is a question of measurement and there is a substantial body of theory dealing with it. Tribe borrowed his tools, the best available at the time, from the economists. Since then however the explosion in computerized data analysis in a spread of disciplines has prompted the development of many new techniques for measuring the performance of systems that, like the law, have to make yes/no-type verdicts. It is this work on measuring the *effectiveness* of such *binary classification* that I shall draw on.

There is no innately right answer to the question of how to do the measuring. It is a question of appropriateness. Thus the meat of this article is an argument that one particular *measure* is appropriate for the criminal justice system. When Kaplan and Tribe were writing on this topic, the problem was to find any such measure at all. Today the explosion of data, computing and interest in binary classifiers means we can pick from a smörgåsbord. DeKay has already introduced signal detection theory, which is a basis for this kind of work in many fields. Hand provides a convenient overview, emphasizing the difficulty of choosing the right one with the warning that, "adopting the wrong measure can lead to incorrect conclusions" [15].

Adopting a measure does not deny the importance of questions of morality, archetypally the weighing of false convictions against false acquittals. Rather, it separates concerns, dividing the morals from the mechanics of calculation and establishing a clearer kind of relation between them. In turn, this empowers our morals to become effective and to directly determine the character of the criminal justice system. Thus this work does not attempt to provide an answer to, for example, the question of whether the standard of proof should be higher or lower; it merely provides the conceptual means by which any polity can set about answering that kind of question. In this restricted sense it provides a complete answer to the questions of evidence law.

---

[14] Laurence H. Tribe, *Trial by Mathematics: Precision and Ritual in The Legal Process,* 84 Harv. L. Rev. 1329, 1386: "C. More Sophisticated Techniques" (1970-71).

[15] David J. Hand, *Assessing the Performance of Classification Methods*, 80 International Statistical Review 400-414 (2012).

I take Laudan as my starting point as his applied epistemology has the foundational rigor that I need (though his normative objective of reducing the number of false acquittals is no part of this work). In the first 12 pages of *Truth, Error, and Criminal Law* he sets out a manifesto of naturalism and it is fair to say that the ambition articulated there is pursued here. We will consider a criminal trial a kind of empirical inquiry, so that the rules that a court follows, or *policies* as I will call them, are its *epistemology*. What we are looking for are principles, or *meta-rules*, that will allow us to choose between these policy options. These principles then will form a *meta-epistemology* for the law.[16]

Of course, there are many bases on which such a decision might be made, moral for example or welfarist. The distinctive feature of an applied epistemology framing is that it narrows down the possible criteria, Laudan's extremely so. The idea is to consider only the extent to which policies are "truth-conducive" and our measure must yield a meta-rule that allows us to choose between any two. Existing "because *n*, X" arguments, we might say, aspire to the status of a meta-rule, but the lack of a cogent connection between *n* and X makes their validity questionable. Remedying that deficiency is another way to express the purpose of this article.

Ultimately, we will find that we need to go up a level and embrace a meta-meta-epistemology. The keystone of the one I put forward is the F_β-measure (pronounced "f measure") derived by Van Rijsbergen in 1974 in the context of information retrieval, the discipline of search engines. If I start by leaning heavily on Laudan, I will lean on Van Rijsbergen even more heavily later on, drawing particularly on a chapter on evaluation from a book he wrote in 1979. [17]

The argument is organized as follows. In Section 1, I set out the elementary conceptual basis in law and statistics for this analysis. In Section 2, I reason to an appropriate statistical framework for monitoring effectiveness with two separate indicators. In Section 3, I introduce the F_β-measure, which I argue is a uniquely appropriate way to combine the two indicators into one.

1. THE NATURE OF THE PROBLEM

1.1 Binary classifiers

For all its unique features, the criminal justice system is not entirely *sui generis*. From a statistical point of view, it can be viewed as one of many systems that make yes/no-type

[16] Larry Laudan, *Truth, Error, and Criminal Law*, 4, Cambridge University Press (2006) *[hereinafter cited as Laudan 2006]*. Laudan offers his own *m/n* meta-epistemology, which I will argue has inherited the flaws of Kaplanism.

[17] Cornelis J. van Rijsbergen, *Information Retrieval*, Butterworths, 1979, drawn from the online edition at http://www.dcs.gla.ac.uk/Keith/Chapter.7/Ch.7.html, Chapter 7.

decisions in a not entirely reliable way. Another example is an over-the-counter pregnancy test kit. Both the criminal justice system and the pregnancy test kit suffer from being imperfect. Just as the criminal justice system may get the verdict wrong, so may the pregnancy test kit. And in both cases there are both two kinds of success, true positives and true negatives, and two kinds of error, false positives and false negatives.[18]

A generic term for a system of this type is a *binary classifier* and another important area where we find them is in the software algorithms used in information retrieval. For example, consider an algorithm that classifies documents as being relevant to legal scholars or not. Again, there are two right answers and two wrong.

Let us begin by setting out the standard statistical basis we have here via a Venn diagram:

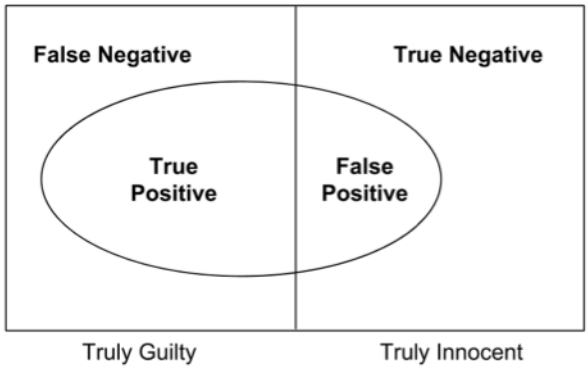


Figure 1.1

The universe is partitioned by two lines that, in the law, correspond to true (or factual) guilt and legal guilt. Our interest is not with one or the other, but the difference between them. For reasons that will become clear, I call this characteristic picture of a universe partitioned into four subsets *the 4-gram*. It can also be represented as a table:

[18] The important point with both conviction and true guilt is to define the categories so that they are binary. Thus talk of "conviction" and "acquittal", for example, is formally problematic as it is not properly binary; some cases could fall into another category, such as the case being dismissed without a verdict. For precision, we could talk of "conviction" and "not-conviction", "guilt" and "not-guilt" but this seems to gain us little while costing a lot in instinctive grasp of the language. Thus I will use familiar terms such as acquittal and innocence throughout this text and leave the reader to remember that these should always be read as indicators of binary classifications.

	Truly guilty	**Truly innocent**
Positive verdict ***Conviction***	True Positive (TP) *Rightful Conviction*	False Positive (FP) *False Conviction*
Negative verdict ***Acquittal***	False Negative (FN) *False Acquittal*	True Negative (TN) *Rightful Acquittal*

Figure 1.2

Note that the four partitions in this diagram consist of *unranked* sets, or *tallies*. Unlike a search engine for example, we are not attempting to rank our results from best downwards. Whenever an abbreviation such as TP, FN, TN or FP appears in this article it refers to a tally and all the statistics that we will develop are ultimately based purely on these four tallies.

This basic setting may seem uncontentious, even redundant. Note however that all we are doing is counting; in other words, unlike previous statistical work in this area such as Kaplan's influential paper, the theory put forward in this article is not concerned with *probability* at all. In any case, the law has not followed the other disciplines by establishing an elementary statistical framework for monitoring how good it is in distinguishing the guilty from the innocent, a quality that often goes by the name of *effectiveness*. This framework varies from domain to domain. For example, the framework used in diagnostics is different to the one in information retrieval. But as yet there is no agreed framework for the criminal justice system.

We can illustrate the nature of the challenge by turning to Tribe. Recall the initial table in Figure 0.1 and our lack of a rule that can determine which of any two outcomes is better. Tribe addresses this problem by providing a mechanism for choosing between possible outcomes via indifference curves (illustrated below with one of his diagrams) and choice sets.[19]

[19] For more detail, see Tribe *supra* note 14, at 1387.

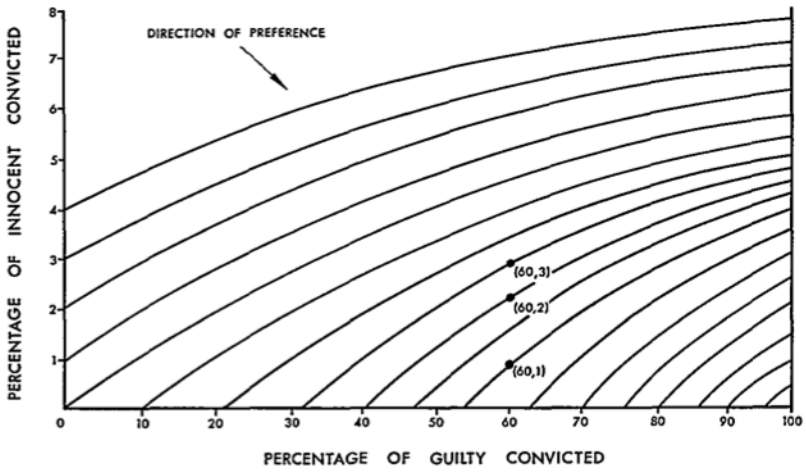


Figure 1.3

However, each indifference curve represents a series of value judgements and Tribe's method provides only an incomplete basis for drawing the curves one way or another. True, the procedure for drawing the curves incorporates the principle that a reduction in the percentage of innocents convicted has decreasing marginal utility in the face of an increasing cost in terms of the percentage of the guilty going unconvicted. But while this principle *constrains* it does not *determine* the position of any point; every point still requires its own value judgement. Hence there is no meta-rule here and this approach provides not so much a solution to our problem as a visualization of it.

1.2 The sovereign perspective

On one important point, several different strands of scholarship today have come together. Rather than looking at things from the point of view of the courts, Laudan insists on considering the criminal justice system as a whole within its social context. Epps, Walen and Kaplow avow similar approaches and I will also.[20]

My *perspective* is that of a sovereign, a concept that, as Hart avers, is in Bentham "flatly descriptive and normatively neutral".[21] Formality on this point is necessary if we are to avoid the many subtle traps, familiar to statisticians, which can arise if we allow our point of view to shift from one actor to another. For example, a diagnostic test that is, from the perspective of the manufacturer, highly accurate may nonetheless be highly inaccurate when considered from the perspective of a patient.[22]

---

[20] Louis Kaplow, *Burden of Proof*, 121 Yale L. J., 738-859 (2012).

[21] H. L. A. Hart, *Bentham and the Demystification of the Law*, 36 Modern L. Rev., 2–17 (1973).

[22] The way the numbers work out can be startling for someone who isn't familiar with the area. See Mudd Math Fun Facts, https://www.math.hmc.edu/funfacts/ffiles/30002.6.shtml.

Within information retrieval, the operator of the algorithm corresponds to the manufacturer in diagnostics. In both cases, they are the entities who have overall responsibility for the effectiveness of the binary classifier. Within jurisprudence, this best corresponds to the old-fashioned concept of the sovereign. However, in our use the concept of the sovereign will be stripped back. The legitimacy of the sovereign's power does not come into it; our interest is not in *how* law comes to be valid.

It is a question of perspective rather than existence. We may find the same doctor at one time performs calculations from the point of view of a patient, at another from the point of view of the manufacturer of a test. Look into it and the question of exactly who the manufacturer "really" is may become problematic (Which company in the supply chain? Which person in the company? Or the shareholders?), but the calculations – and their usefulness – don't.

Hence this perspective is merely that of anyone concerned with the workings of the system as a whole and for the role it plays in society including the Department or Ministry of Justice, the citizen who wants to enjoy the benefits of living in a lawful land and the judiciary when making policy.

So this is a Harris-like sovereign, best understood as a constructive metaphor; we choose to think *as if* the criminal justice system reflects the view of a single will.[23] Indeed, my needs are lightweight enough to be consistent with those who consider the sovereign an intrinsically flawed concept – for example, Eleftheriadis who argues instead in favor of a "supreme authority", which he makes sense of in terms of political philosophy and the social contract.[24] Looking in the other direction, as will become clear, our concern is exclusively with the making of policy and hence we have no need for a sovereign as powerful as Schmitt's "he who decides on the exception".[25]

The distinctive features of the sovereign's position that we will rely on are that it is obliged to consider the big picture of the entire criminal justice system, is complicit in it and dependent on it. It is *obliged* in that the criminal justice system has a powerful impact on the society around it to which the sovereign cannot be indifferent. It is *complicit* in that it is responsible for funding the criminal justice system and directing many elements of it. If, for example, we consider the list of cases that arrives at the door of a court, then the sovereign has agency over all of the many layers of discretion that lead police and prosecutors to construct that

---

[23] J.W. Harris, *The Concept of Sovereign Will*, Acta Juridica (Essays in Honour of Ben Beinart, Volume II), Juta & Co., Cape Town, 1–15 (1979).

[24] Pavlos Eleftheriadis, *Law and Sovereignty*, Oxford Legal Studies Research Paper No. 42/2009, University of Oxford. Available at http://dx.doi.org/10.2139/ssrn.1486084 (2009).

[25] Carl Schmitt, *Political Theology: Four Chapters on the Concept of Sovereignty* (George Schwab, trans), University of Chicago Press (2005) [1934].

particular list and not another. It is *dependent* on the actors in the criminal justice system to achieve its goals.

One advantage of using this term is that it circumscribes the ambition and applicability of this work. If Bentham and Austin have been superseded by Hart, the essence for us of *The Concept of Law* is that their command-oriented framing of the law is insufficient rather than wrong. Command thinking ignores, for example, the way the law is internalized, the expanses of the law that are enabling and the complexities of law about laws. All of these are outside the ambit of this work. My concern is exactly with the commands of the criminal justice system where the Bentham-Austin sovereign is enough.[26]

I am not adopting the perspective of a defendant (concerned only with their own predicament), a potential criminal (concerned with the risks of getting caught), nor that of any of the actors within the criminal justice system. All of these are plainly partial and hence problematic when considering matters of policy. On the other hand, within the existing discussion in jurisprudence we can identify a number of different perspectives that have laid claim to this central position and compete with that of the sovereign, which I will call *quasi-sovereign*:

- The jury perspective, which Kaplan adopted when applying decision theory to setting the standard of proof;
- The "trial system as a whole", which Tribe adopted in his response to Kaplan;
- Risinger's perspective, adopted as a defense against Laudan's arguments, of someone concerned only with those arrested or otherwise "drawn into" the system[27].

The consequences of adopting the sovereign perspective rather than any of these others are a preoccupation of this text and will emerge in the course of it. For now however we can start by merely confirming that they are indeed different. Within the jurisprudential literature, Tribe has already done this for the trial system as a whole and the jury perspective.[28] Laudan, by repeatedly emphasizing the crimes that are not noticed by the system at all, has done it for

---

[26] H.L.A. Hart, *The Concept of Law,* Clarendon Press, Oxford, 1961.

[27] "From this view, the dominant frame of reference for evaluating the justice of the performance of the criminal justice system (as opposed to its efficiency) is limited to how it deals with those drawn into it. Generally, this means those who are arrested and charged." D. Michael Risinger, *Tragic Consequences of Deadly Dilemmas: A Response to Allen and Laudan,* 40 Seton Hall L. Rev. 991, 1020 (2010).

[28] "But when one gives due weight to the costs of combining in the trier the separate functions of deciding what happened in a particular case and evaluating the anticipated consequences of alternative verdicts, one will expect the lawmaker rather than the factfinder to use a model such as the one Kaplan and Cullison propose, and one will define the decision problem to be solved not as the one-shot problem of fixing a standard of proof for a particular trial with four possible outcomes, but as the much larger problem of establishing such standards for the trial system as a whole." Tribe 1971 p1385.

the sovereign and the trial system. And Risinger himself has adopted his position precisely in order to distinguish it from the sovereign-like view espoused by Laudan.

We can also look at the differences through a statistical lens. It is clear that the extent of the universe being considered by the sovereign and the quasi-sovereigns is different. In fact, we have a sequence of proper subsets (indicated by the symbol ⊃):

Name	Sovereign ⊃	Risinger ⊃	Tribe ⊃	Kaplan
Extent	*Everything*	*"drawn into"*	*Trial system*	*One jury*
False negatives	*All criminals who are not convicted*	*All criminals the police apprehend who are not convicted*	*All guilty defendants brought to trial who are not convicted*	*One defendant, if truly guilty and not convicted*
False positives	*All innocents convicted in court*	*All innocents convicted in court*	*All innocents convicted in court*	*One defendant, if innocent and convicted*

Figure 1.4

Restricting ourselves for the moment to the first three columns, each different universe implies a different 4-gram. If, with god-like insight, we were to start counting the tallies of the 4-grams, each would yield different numbers. For example, a glance at the differences between the false negatives in the above table shows that the same system will yield different values for *n* depending on the universe we adopt for the calculation.

This rather elementary observation has repercussions. For example, if we restrict our gaze to the list of cases presented to court, we might perhaps make our policies so demanding that the prosecutors bring to us only one case of rape in the whole year. Let us assume for the sake of argument that we get the verdict right. Then we have achieved on any quantitative assessment a success rate of 100 per cent. Meanwhile, an epidemic of rape may be taking place in society at large. The point of the sovereign perspective is that it does not allow us to say, "oh, that's someone else's problem"; it obliges us to consider the criminal justice system as a whole.

This is not to say that there is a 1-1 correspondence between universe and perspective. For example, Holmes's bad man obviously does not share the sovereign's objectives but he does rationally share the sovereign's choice of universe since only its statistics will tell him how likely he is to be convicted of the crime he is considering.[29]

[29] Oliver Wendell Holmes, *The Path of the Law*, 10 Harv. L. Rev. 457 (1897).

Which is the right perspective? To investigate an aspect of the system, any perspective might be useful. The question we are concerned with however is the making of policy. One challenge then for all the quasi-sovereign perspectives is that if you decline to address the sovereign's concerns, you may perhaps expect the sovereign to lack interest in your conclusions. A second is the temptation for them to justify themselves through an unexamined reliance on the sovereign perspective.

Take Kaplan's jury-eye view for example. The most obvious utility to a juror of any outcome at a trial is zero: it has no consequences for them at all. It is only by inviting the fact-finder to consider the wider needs of society – that is, to adopt a position that is at least sovereign like – that relevance can be attempted. Implicit in Kaplan's paper, this is explicit in the later uses of it that have acquired the status of orthodoxy. For example, here is Stein in his otherwise radical book, *Foundations of Evidence Law*:

> Following John Kaplan, the utility-based criminal proof standard can be formulated as I/(I + G) where I and G denote, respectively, the social damage inflicted by convicting an innocent suspect (I) and by erroneously acquitting a criminal (G).[30]

But if the sovereign perspective is the one that matters, what is the argument for the sovereign delegating to the juror the responsibility for managing the risk of social damage? Kaplan cites the example of an employee making a decision on behalf of a manufacturer. But before we get into the psychology of that decision, there is the meta-decision by which the manufacturer invites, or not, the employee to make the decision. Contrast this with booking a hotel room where the manufacturer may leave this up the individual or instead have a central office. So, although the decision on an appropriate threshold can be delegated to jurors, the first question, as Tribe pointed out, is whether it should be.[31]

This problem is emphasized if we ponder the fact that the final column in the table above sits somewhat awkwardly with the others. We evidently cannot talk about false positives and false negatives in quite the same way there. This is because, unlike many subsequent authors, Kaplan explicitly roots his argument in what he calls a "personalistic theory of probability", a kind of reasoning that naturally falls under the heading of "subjective probability" in Dawid's

[30] Stein *supra* note 8, at 172. Kaplan's original paper, *supra* note 5, is the only reference. (I have added brackets to (I + G) that were omitted in the original).

[31] "Positing the Wrong Decisionmaker" – Tribe *supra* note 13, at 1384.

categorization of the different kinds of probability theory that may be found in legal arguments. [32] [33]

Dawid points out that subjective probability entails an irreducibly subjective element and suggests such an approach may be appropriate when one is unable "to specify, or even conceive of, some relevant sequence of repetitions of the event in question". That is, it may be appropriate for assessing the unique questions that may arise within an individual trial but not for a question of policy which is going to be applied over and over again. So, by adopting a subjective probability approach, we may conclude that Kaplan has at the outset made his work unsuitable for determining any question of policy, including specifying a standard of proof for all trials. The approach is only justified if the standard of proof is intended to vary from trial to trial, that is, if the sovereign chooses to delegate responsibility for this decision to the jury.

Ultimately, in criminal law one may argue that all the quasi-sovereign perspectives can resolve such questions of justification only by appealing up the chain of authority to the sovereign. However, to make the exercise of this article worthwhile it is not necessary to assert that the sovereign's perspective is superior to all others; it is merely necessary to allow that it is different and cannot be ignored. It is a question of separating concerns.

1.3 Epistemic policy dilemmas

In considering the possible application of mathematics to the work of the courts, Tribe distinguishes between its use in individual trials and in setting policies for the trial system as a whole, in Laudan's terms between the epistemology and the meta-epistemology.[34] My concern in this article is exclusively with the meta level.

Policies are simply any rule followed by the courts that shapes how they conduct a trial, regardless of origin. Hence they include the basic format and procedure of a trial (including specifying the range of things that will not be subject to policies), the standard of proof, whether majority verdicts are acceptable, rules governing the admissibility of evidence and doctrines of presumption. An important rider to this however is that the sovereign must concern themselves with the policies not only of the courts but of the police and prosecutors,

---

[32] Kaplan *supra* note 9, at 1066.

[33] Philip Dawid, *Probability and Proof*, appendix to "Analysis of Evidence" by T. J. Anderson, D. A. Schum and W. L. Twining, Cambridge University Press 2005, 35. Available at http://www.cambridge.org/9780521673167 (search for "appendix" under the resources tab)

[34] "In speaking of mathematical methods 'in the trial process,' I am referring to two related but nonetheless separable topics: not only to the use of mathematical tools in the actual conduct of a particular trial, but also to the use of such tools in the design of the trial system as a whole." Tribe *supra* note 4, at 1393. See particularly Section II for his discussion of policy setting. A useful review of the progress of the debate at the trial level is provided by Peter Tillers in *Trial by mathematics – reconsidered*, 10 L., Probability and Risk 167−173 (2011).

too. Policies here would include, for example, the allocation of resources to different types of crime.

Policies are the means by which the sovereign can impose their will and we can identify two different kinds of decision that face *policymakers* acting on behalf of the sovereign, amongst whom I count the senior judiciary, legislators and senior bureaucrats in the relevant department of government, prosecuting agency and police. Some kinds of policy offer the promise of an improvement in accuracy in all cases. Introducing DNA fingerprinting is an example. Other kinds of policy involve a trade-off between the two kinds of error, for example allowing evidence of previous convictions. If we allow evidence of previous convictions then we may hope for fewer false acquittals but fear more false convictions.

Clearly, the first kind of choice is unproblematic; our difficulties arise with the second. Even though it clearly entails a trade-off, because of its special role in the system let us put the standard of proof to one side. The rest we will label *epistemic policy dilemmas.* Somewhat awkwardly therefore I will find myself from time to time referring to "epistemic policy dilemmas or the standard of proof". Consideration of these will be the crucible in which we forge our meta-rule.

The question of the trade-offs is intrinsically meta, one of deciding between policies rather than facts in any one case. A false positive involves a defendant who is truly innocent, a false negative one who is truly guilty. Thus there is no alchemy that allows us to trade one for another in any given case. It is only by changes to policies affecting many cases that we can affect the resulting tallies of the two errors.

Criminal procedure is riddled with epistemic policy dilemmas. In a second book published last year Laudan lists 33 examples, starting with the exclusion of evidence obtained by the police without a warrant and ending with the exclusion of the prior sexual history of the testifying victim in rape cases.[35] Many of these evidently can have a significant effect on the character of the system. However, his idea of what he calls "burden shifters" is narrower than my definition of epistemic policy dilemmas, and my definition drags in many of the most fundamental principles of the system.

Thus taken in total and bundling them with the standard of proof, the question of how to deal with the epistemic policy dilemmas seems the single biggest policy problem with which the criminal justice system has to grapple. They are perhaps not the holes but the cheese of policy

[35] Larry Laudan, *The Law's Flaws: Rethinking Trial and Errors?* College Publications, London, 113-120 (2016) *[hereinafter cited as Laudan 2016]*.

development, a view that has been recently articulated by Stewart[36]. Another way to put this is that finding a way to resolve the dilemmas is the central problem of evidence law.

Resolving such dilemmas is the day-to-day stuff of much algorithm development in information retrieval. For example, take the binary classifier mentioned earlier that identifies documents as being interesting to legal scholars or not. It may work by having a bag of terms that are typically involved in legal scholarship, such as "Blackstone", "jurisprudence" and so on, and identifying these in the text. Deciding on the scope of this vocabulary will have a powerful impact on the performance of the algorithm. For example, should the term "law" be included? If it is, we may find more documents that are relevant. On the other hand, it is likely to also return many documents where the law is mentioned only in passing and which legal scholars are likely to find irrelevant. The question of whether to include or exclude the word from the bag is one of information retrieval's many epistemic policy dilemmas.

2. THE CASE FOR PRECISION AND RECALL

After drawing the 4-gram, the first step in developing a statistical framework is to establish how we are going to monitor the two possible errors thrown up by the system, the false positives and the false negatives. Each error has to be compared to what I will call a *yardstick* of success, the true positives or the true negatives so as to generate a ratio. If we were actually in the habit of measuring performance, we would of course reach for such ratios quite naturally. As soon as we wanted to compare performance in two different years or two different jurisdictions we would immediately have to switch from tallies to rates.

The ratios in use vary from field to field and depend on the nature of the system being studied, its purpose and the needs of the users. For example, both diagnostics and information retrieval measure the false negatives against the yardstick of true positives. They do this by constructing the ratio TP / (TP + FN). In diagnostics this is called sensitivity and in information retrieval it is called recall, but it is exactly the same ratio.

$$\text{Sensitivity} = \text{Recall} = TP / (TP + FN).$$

A ratio constructed in this way can only vary between 0 and 1. A value of 1 (or 100 per cent) implies perfection – no errors have been made at all. A value of 0 implies utter failure – no correct results have been achieved at all.

Where diagnostics and information retrieval differ is over how to measure the false positives; each discipline constructs a different ratio using a different yardstick. In diagnostics the

[36] : "...any change in procedural rules is likely to involve some sort of trade-off between the two types of error..." – Hamish Stewart, *Concern and Respect in Procedural Law*, in *The Legacy of Ronald Dworkin*, Oxford University Press 373, 377 (2016).

yardstick of true negatives is used to create a ratio called specificity, in information retrieval the true positives are used to create an indicator called precision.

Specificity = TN / (TN + FP)

Precision = TP / (TP + FP).

It is not necessary here to go into the reasons why different ratios are used in different fields. At a high level, we could say that a pregnancy test has to be even handed (in that it is equally concerned to be useful to both women who are pregnant and those who aren't) while the user of a search engine will only ever look at one side (the things that are found). The point is that the differences derive from the different purposes of the users and the different characters of the systems being studied.

Choosing two ratios will not in itself solve our problem. For now, we are obliged to treat the two different kinds of error with two different indicators and this leaves us without any obvious meta-rule. But, if done right, it should give us indicators of the rate of each error that are meaningful to us, so that when they go up we are confident the picture is improving and when they go down we are confident it is getting worse.

If we take diagnostics and information retrieval as two alternative potential templates, then they agree to measure the rate of false negatives against the yardstick of true positives. So let us put that to one side for now. The more pressing question for us is, which is the more appropriate indicator of the rate of false positives for the criminal justice system, specificity or precision? And this boils down to the question of which yardstick is more suitable, true positives (for precision) or true negatives (for specificity).

We want to look at this from the sovereign's point of view. So we should put to one side arguments that derive from other points of view. This removes what is, in diagnostics, a critical argument for specificity. Diagnostics requires an indicator such as specificity that contains the true negatives in order to calculate the negative predictive value, which is to say the likelihood that a woman who is told she is not pregnant by a pregnancy test really isn't. Thus the use of the true negatives is a necessity.

The sovereign of the criminal justice system is not in the same position as the maker of a diagnostics test. It does not need to be able to tell someone who has not been convicted what the chances are that they truly are innocent. Thus the precedent of diagnostics does not apply in our case and we are not obliged to opt for the yardstick of the true negatives.

2.1 Achievement

We can all agree that, from an epistemic point of view at least, the criminal justice system has achieved something when it convicts a truly guilty person. But what has it achieved if it acquits a truly innocent person? All that it has done is to waste everyone's time. True, this is a better result than convicting them of a crime they did not commit, but that does not make it

in itself worthwhile. It is as if we have driven around in a circle. Certainly, it is better that we did not have an accident on the way. But we have not gone anywhere, we are no better off than we were before.

This rather simple point is perhaps easily overlooked because of the ingrained habit of looking at the issues involved from the provincial point of view of the actors involved in the system. *Given* that an innocent person has been put on trial, the court certainly achieves something by not convicting them. Similarly, *given* that the police have put forward a case where the accused is truly innocent, the prosecutors certainly achieve something when they exercise their discretion and discard the case. And the police also achieve something when, *having placed* someone under investigation, they conclude that they are in fact innocent. And again, *if I have been charged* with something I did not do then certainly the system achieves something when it does not convict me. But in all of these cases it would have been better if the innocent person had never been dragged into it at all.

This is not to say that it is possible or desirable to have a system that is not required to reliably acquit the innocent. The innocent will be falsely accused, the police will find evidence pointing at the wrong people. Such cases make plain the need for the system to be successful at acquitting the innocent-but-accused. But this does not make the acquittal of innocents the benchmark of success for the system any more than the amount of stalk collected and discarded is the yardstick a farmer would want to use to measure the success of a machine for harvesting wheat.

A court, the prosecutors and a defendant are all actors within the system whose assessment of achievement habitually starts with the premise that a specific person has been accused, whereas the sovereign presides over the entire system and that starting premise is missing. Different actor, different universe, different premise – statistics teaches us that we should not be surprised in these circumstances to arrive at a different conclusion about the appropriate way to look at the risks and performance involved.

This clear distinction in achievement between true positives and true negatives is a natural consequence of the command quality of the criminal law. When we create an offence, we generate a need for a capacity to convict those who commit the offence. To make the deterrent convincing, we must be at least capable of convicting the guilty. By contrast, we have no need to be able to do anything to those who do not commit the proscribed act.

Diagnostic classifiers, such as a pregnancy test, do not share the same lopsided purpose. As we have seen already, many of those in information retrieval do. Just as we may say that the purpose of such an algorithm is to identify all and only the relevant documents we may say that the epistemic purpose of the criminal justice system is to convict all and only the truly guilty. To truly adhere to this statement requires the acceptance of a degree of simplification implied by our epistemic frame of reference. We can perhaps agree that we would like to convict all truly guilty murderers, but do we actually want to convict all those truly guilty of driving too fast on the highway? The willingness to leave speed cameras switched off

suggests perhaps not. But the kinds of reasons that would lead us to such a conclusion are non-epistemic; for example, we do not wish to provoke widespread resentment from otherwise law-abiding citizens by revoking thousands of driving licenses. If we are concerned exclusively with truth-finding then we must allow that we want the system to establish the truth, which entails accurately distinguishing between the truly innocent and guilty, and the better it is at identifying the guilty the better it is, even if we choose for non-epistemic reasons to overlook some of its conclusions or capacity.

Consider by contrast what the system would look like if the purpose of the criminal justice system *was* to acquit the innocent rather than convict the guilty. This would be a system in which the police drove around looking for acts of innocence. On finding one, a handshake perhaps, the citizen would be un-arrested, un-charged, taken to court, un-convicted (or – bad news – not) of an innocent act and sent home (or not). This would not be a paradise of justice but a horrific hybrid of two intolerable regimes: on the one hand a totalitarian nightmare beyond all those yet imagined in which a life of law-abiding innocence would be subject to continual surveillance, prosecution and judgement, with the threat of error and punishment constantly hanging over us; and on the other hand a terrifying lawlessness in which the state had no arm with which to find, convict and deter criminals. The bizarreness of this counterfactual is evidence only of how deeply we have imbibed the underlying purpose of the criminal justice system.

Of course, the system does not have to be lopsided at all, so consider what it would look like if the police and prosecutors were equally concerned with identifying guilt and innocence. This would remove the second element of the horrific hybrid, the lack of an arm to tackle criminality. But it would leave in place the first element, the arm charged with identifying innocence, and possibly getting it wrong. This plainly is not the world we live in, or want to.

In short, it is precisely the fact that the criminal justice system *is* preoccupied with trying to convict the guilty that makes it bearable in a democracy. Thus if we consider the purpose of the criminal justice system as a whole, from the point of view of the sovereign, then it is not evenly balanced as in the diagnostics case; from an epistemic point of view the criminal justice system has a purpose, which is to convict the guilty. This suggests that the true positives rather than true negatives are the appropriate yardstick.[37]

[37] It may be objected that the question of achievement is not truly epistemic, that I am smuggling in normative values. This seems to me more a question of definition than anything else. The issue of achievement is exactly what divides diagnostic statistics from those used in information retrieval. Should we say that those choices, too, are not properly rooted in truth finding? That would seem to me perverse.

2.2 Measurement and definition

The sovereign perspective obliges us to consider afresh the basic statistical framework that we use to represent the criminal justice system, beginning with the 4-gram.

A court is presented with a list of cases over which it has no control. Thus there is no question that the 4-gram is well defined and a suitable framework on which to develop an analysis of the work of a court. But we are considering the system as a whole. And the criminal justice system as a whole is not presented with a list of cases. This elementary difference has consequences.

To begin with, let us attempt to construct the 4-gram for the criminal justice system as a whole. So... what is this?

Figure 2.1

What, exactly, is the set that is the universe that we are starting with?

Let us start with the subset that we are confident about. A true positive consists of a convicted defendant who is truly guilty. If the same defendant is convicted of two crimes, that is two elements. Equally, if two defendants are convicted of a single crime, that is also two elements. Thus an element of the true positives consists of a criminal-crime pair.

If we convert this now to neutral terminology to allow for the other tallies in the 4-gram that do not include the truly guilty, we can see that an element consists of a person-action pair. Then if we divide the 4-gram in half, between the truly guilty and the truly innocent, the truly guilty half consists of pairs where the action is a crime and the person is one of those who committed the crime. The other half, the truly innocent half, are the pairs where either the action is not a crime or the person did not commit the crime.

What are these actions that are not crimes? Just that, any action that could be a crime but in fact isn't. A handshake for example may be innocuous or it may be the signal to start shooting. Any action could be a crime and hence this set includes *all actions* in our jurisdiction that are not crimes.

Who are these people who did not commit the crimes? Again, just that, *everybody* who might have committed the crime but did not.

It is evident that this universe of action-person pairs is to us unmeasurable. Thus reliance on it can only lead to a faux meta-rule, one that is conceivable but incapable of being applied because it relies on measurements that we cannot make.

Element	Action-person pair
Universe	All pairs composed of actions that take place within our jurisdiction and people who might have performed the action.[38]
True Positive	Pair in which the action is a crime, the person committed the crime; and the person is convicted
False Positive	Pair in which either the action is not a crime or the person did not commit the crime; but the person is nonetheless convicted
True Negative	Pair in which either the action is not a crime or the person did not commit the crime; and the person is not convicted
False Negative	Pair in which the action is a crime, the person committed the crime; but the person is not convicted

Figure 2.2

In this problematic universe we should pause to consider which of the four tallies are problematic. The true and false positives, the two inside the oval defined by the guilty verdict, are within our grasp. The false negatives, crimes that are not successfully prosecuted, is clearly challenging but in principle not beyond us; household surveys are already used to estimate the prevalence of different crimes in the real world.[39] The real problem is the fourth of the tallies, the true negatives.

This is not an unfamiliar problem. For search engines tasked with, for example, monitoring the entire internet, the concept of a tally of the true negatives suffers from a lack of definition. The true negatives in this case correspond to all documents on the internet not relevant to the user's query. But "it is frequently found that this quantity is uncountable, undefined or in

[38] One might argue that this should read "actions that might have taken place within our jurisdiction"; but this only strengthens my argument.

[39] For example, the Crime Survey for England and Wales administered by the Office for National Statistics seeks to establish the incidence of some crimes through surveys of households rather than from administrative data. Such crimes include crimes of violence against children aged between 10 and 15. For the year ending March 2017, the CSEW offers estimates based on different two methodologies, each with a range based on a 95 per cent confidence interval. The *preferred* method indicates there were 359,000 incidents with a range of 292,000 to 426,000. The *broad* method indicates 660,000 incidents with a range of 570,000 to 750,000. Office for National Statistics, *Statistical Bulletin: Crime in England and Wales: year ending Mar 2017,* Table UG10, ONS July 2017.

constant flux".[40] This is one of the reasons for the reliance on the yardstick of the true positives in that field.

Returning to the law, the true positives is, in contrast to the true negatives, clearly well defined. This then is another reason for preferring the true positives to the true negatives as our yardstick.

The salience of this argument is obscured if we instead adopt the quasi-sovereign perspective of the courts. From a court's point of view, the universe is the list of cases it is presented with, a fixed number. And the number of convictions is also a known quantity. This gives us two equations in a system with four unknowns (the TPs, FPs, TNs and FNs), or degrees of freedom. Thus given any two additional equations the system is fully determined, which implies that we can then derive all four tallies and hence any ratio of them. Thus whether we use precision or specificity is merely a matter of convenience; neither tells us anything we can't derive from the other.

2.3 The 3-gram

This leads to a counter-argument. Very well, you may say, if the true negatives are not well defined, then neither is any universe that contains them. All this argumentation is based on having such a universe to start with. But when you draw the 4-gram, it includes an unmeasurable subset. You have demonstrated only that your own argument stands on flawed foundations and is unreliable.

At this point we should recognize that it is indeed a conceptual mistake to include the true negatives in the universe at all. It is of course completely conventional in this discussion to do so. And there are many statistics textbooks setting out the standard 4-gram framework that I outlined at the outset. However, those textbooks do not guarantee that the framework will be relevant or usable in a new context.

Another way to look at our problem is that we have two well-defined sets, that of criminal acts and that of those who are convicted. The two have an intersection and thus we have three tallies to consider. This is how Van Rijsbergen looked at the problem in 1974. Using the symbol $\cap$ to indicate the intersection of two sets and $\cup$ to indicate their union, he said:

> Let us now return to basics and consider how it is that users could simply measure retrieval effectiveness. We are considering the common situation where a set of documents is retrieved in response to a query, the possible ordering of this set is ignored. Ideally the set should consist only of documents relevant to the request, that is giving 100 per cent precision and 100 per cent recall ... In practice, however, this is

[40] George Hripcsak and Adam S. Rothschild, *Agreement, the F-Measure, and Reliability in Information Retrieval*, 12(3) J. Am. Medical Information Assoc. 296–298, 296 (2005).

rarely the case, and the retrieved set consists of both relevant and non-relevant documents. The situation may therefore be pictured as shown in Figure 7.11 *[Figure 2.3 here]*, where A is the set of relevant documents, B the set of retrieved documents, and A ∩ B the set of retrieved documents which are relevant.[41]

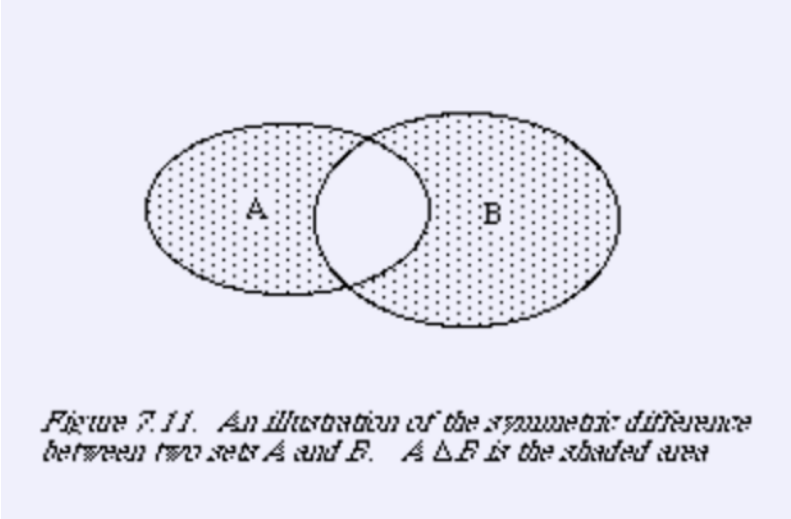


Figure 2.3

I call this Venn diagram *the 3-gram*. In it there is only one possible yardstick of success, the intersection of the two sets, and so it is no surprise that Van Rijsbergen goes on to recast precision and recall as follows:

If *A* is the set of relevant documents and *B* the set of retrieval documents, then:

$$P = \frac{|A \cap B|}{|B|} \qquad \text{and} \qquad R = \frac{|A \cap B|}{|A|}$$

The set theoretic notation should not confuse us. The | brackets indicate only that we are counting the number of elements in the specified subset and so these are simply different labels for tallies we have already met:

| A | = TP + FN

| B | = TP + FP

| A ∩ B | = TP.

[41] Van Rijsbergen *supra* note 17.

So, unlike the true negatives, the true positives survive the transition to the 3-gram, and this is another reason for preferring it as our yardstick.

2.4 Agency and incentives

This leads in turn to another counter argument. Well, you may say, this is all rather abstruse. We can sweep away these problems by working within a well-defined universe. For example, in England we could take the universe defined by the crime numbers allocated by the police to all reported crime. Now we do have a well-defined universe and a well-defined set of true negatives. This however introduces the question of the agency of the actors within the criminal justice system, and indeed the sovereign, in defining the universe.

First note that the police have a great deal of control over what acts are issued with crime numbers. If they go into the town center on Saturday night they may well find a fight; if they don't, they won't. If the police provide an online form to allow the reporting of a crime, they will get more reports than if they make people wait in a long queue.

Second, in order to compare two tallies, the elements in each set must be of the same kind. The set of cases that reaches court, the true and false positives, is composed of action-person pairs. But while they may identify a *crime* (action), crime numbers will often lack a *defendant* (person). To bring a case to court, the police in these cases must use their discretion to choose which crimes to investigate and then go on to identify a defendant. Later on, the prosecutors will also have their say.

Thus the universe that has been put forward as the basis for measuring the performance of the system, and the actors in it, is in large part created by the system, and the actors in it. In short, while this move has created a universe that comes with a number attached, and hence gets us back to a position of having four equations for the four tallies, it is a number that the system we are trying to measure is complicit in generating, which makes it and any subsequent calculations of dubious value.

Further, what incentives do the two possible yardsticks provide to these actors? The yardstick of true positives gives the actors an incentive to include cases that will likely turn out to be true positives, that is, people who really are guilty. This is desirable. On the other hand, the yardstick of true negatives gives them an incentive to include cases where the accused will likely turn out to be genuinely innocent, which is undesirable.

We can see these basic facts reflected in the two possible indicators that are competing for our affection, specificity and precision. Let us hypothesize a change to the system that involves only the tallies measured by specificity, such that the number of false convictions (FP) remains the same but more unconvicted innocents (TN) pass through the system. Specificity would rise in such a situation and tell you that your performance had improved.

As an individual court, that would make sense. *Given* that the cases arrived in front of you, acquitting more of the innocent is clearly an improvement. But from the point of view of the

sovereign, it does not make sense. It means the police and prosecution are bringing to court more innocent people. Indeed, if specificity was used to measure the effectiveness of the entire system, performance would, perversely, improve every time the police and prosecutors dragged a transparently innocent person to court. So specificity fails this test and emerges as unsuitable as a performance indicator for us.

Now let us look at the yardstick of true positives by asking, what happens to precision under this hypothesis? Everything else being equal, nothing. Precision is not interested at all in how many unconvicted innocents pass through the system. In this scenario, precision does not wrongly indicate to the sovereign that the performance of the system is improving, and it gives no perverse incentives to the other actors in the system.

Now let's try the converse test with only the tallies measured by precision. Let us hypothesize a change in the system in which the number of false convictions (FP) remains the same but more truly guilty people are convicted (TP). Specificity is not interested in how many guilty people are convicted and, all else being equal, would be unmoved by this change. By contrast, precision would rise in such a situation and tell you that your performance had improved. As an individual court, that would still make sense. But from the point of view of the sovereign, it would also make sense. It would mean the police and prosecution are bringing to court more guilty people. Indeed, if precision was used to measure the effectiveness of the police and prosecution, their performance would, rightly, improve every time they brought a genuinely guilty person to court and secured a conviction.

Thus it is clear that the yardstick of true negatives provides perverse incentives that would undermine its use in the real world. By contrast, the yardstick of true positives provides welcome incentives.[42]

2.5 Conclusion

Although presented in a narrative, the three grounds for rejecting specificity do not depend on one another. To reject this argument therefore, it is necessary to reject the case based on achievement and that on measurability and that on incentives.

If we compare the case for precision in the criminal justice system with that in information retrieval where it has long been adopted, the case for the criminal justice system is the stronger, for the question of incentives impinges strongly on the criminal justice system while it is completely missing from information retrieval.

[42] This does not however mean that we have in this way removed *all* problems of incentives. For example, a police chief might try to raise the score on the indicator by questionably limiting the cases that end up being counted merely to those that are relatively easy to resolve.

The choice is a true dichotomy and the fact that the case is stronger than in another discipline that made this choice and lived with it happily for 40 years while working with mountains of data and many different decision-making systems (different algorithms) should weigh heavily with us.

The above arguments in favor of the yardstick of the true positives apply as strongly to measuring the rate of false negatives as they do to the false positives. So we can endorse our provisional decision to use recall to monitor those.

We set out to develop statistics that could be used by the sovereign to monitor the effectiveness of the system in its entirety. We have established that from that viewpoint the true positives are the better yardstick. Although consistent with the sovereign's concerns, this is quite a departure from established scholarship. The conventional parsing of concerns into the categories of "error reduction" and "error distribution" in evidence law[43] mean that "because *n*, X" reasoning has been quarantined from consideration of both the true positives and true negatives. For Laudan the two indicators that matter are *n* and *m* and the one tally that is *not* involved in those two is the true positives. Equally, the true positives are missing from Kaplan and hence the consequentialist approaches to the standard of proof and Stein. And of course they are not mentioned in Blackstone's dictum.[44]

The three distinctive features of the sovereign's position have come into play. Its obligation to consider the impact of its decisions on wider society underlies the question of achievement and leads to the question of measurement. Its complicity and dependency underlie the question of incentives.

We can now go forward with the two indicators of precision and recall, one to monitor the rate of false convictions, one to monitor the rate of false acquittals, and both using the yardstick of rightful convictions. These are the same indicators that are used in information retrieval. The reasons for using them are not the same, but the end result is and that will make information retrieval more relevant to our work as we go forward than diagnostics.

As a kind of sanity check and to feed our intuition, here is an example of what the curves for precision and recall typically look like for one binary classifier when plotted against the algorithm's confidence threshold, which corresponds to the standard of proof. The data is

---

[43] Laudan 2006 *supra* note 16, at 1.

[44] An exception to this general rule is Lillquist, who has said, "It is not enough to weigh only the costs of erroneous convictions and acquittals; *accurate* convictions must be weighed in reaching any decision about the proper standard of proof." Erik Lillquist, Recasting Reasonable Doubt: Decision Theory and the Virtues of Variability, 36 U.C. Davis L. Rev. 85, (2002) 91.

taken from an algorithm using the open source Maui[45] document classification engine; the table of data on which the curves are based will be presented and used later:

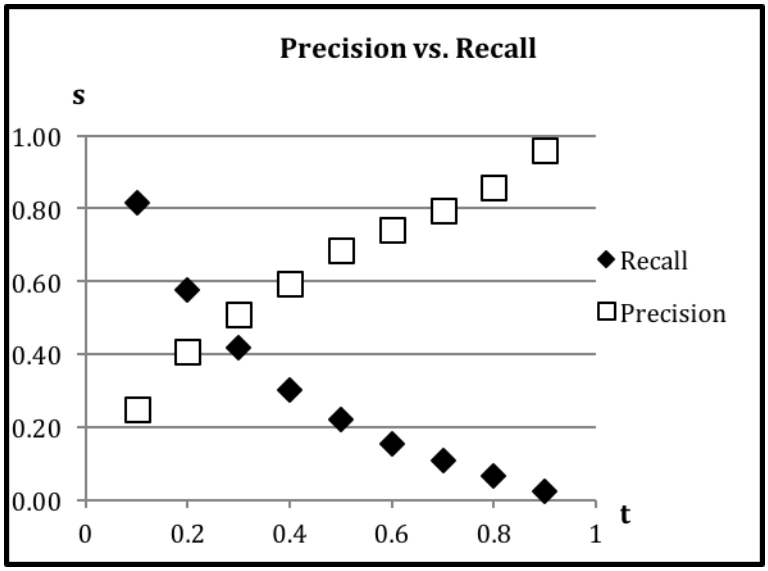


Figure 2.4

Following Hand, let us designate the vertical axis *s* for score and the horizontal axis *t* for threshold. As we raise the confidence threshold (increase *t*) precision rises while recall falls – we would expect nothing else given the evident trade-off between the two errors that we are struggling with.

We can now, in principle at least, identify epistemic policy dilemmas easily. They are any policy question where we lack an option that, compared to the other options, maximizes both precision and recall.

We don't yet have the measure or meta-rule we want because we have two indicators, and – even if we had measurements of them – it's not yet clear how to make a decision that resolves the tension between them.

3. THE F_β-MEASURE

The necessity for a single measure and meta-rule can be illustrated by attempting to derive a strategy for managing the standard of proof with only the tools we now have. For example, what about *maximizing precision*? That is in a sense too easy. For example, we could raise the standard of proof to something like "beyond even the tiniest bit of doubt". A side effect however would be that we stop caring about recall; fewer and fewer people would be

[45] Alyona Medelyan, *Human-competitive automatic topic indexing (Thesis)*, University of Waikato. Retrieved from http://hdl.handle.net/10289/3513. Software at http://www.medelyan.com/software.

convicted of anything and the system would cease to fulfil any useful purpose. Thus "maximizing precision", for all its rhetorical attractiveness, is not a viable strategy for the sovereign when we insist on giving the word "maximize" a rigorous meaning.

A second possible approach would be to set *a precision threshold*. When we are over the specified figure, we are happy for precision to fall and can focus on recall; but when below we are committed to doing all in our power to raise precision.

In this case, below the threshold, the problems we saw with maximizing precision return. In order to achieve our objective, we abandon any concern for recall with potentially unlimited consequences for the saliency of the system. Equally nastily, above the threshold we could find ourselves sacrificing an enormous amount of precision for a tiny amount of recall.

Thus if we start, like the sovereign, with a concern for both kinds of error then precision and recall in themselves are inadequate to solve our problem.

I now have a presentational difficulty to get over. We want to derive our measure from first principles but lack empirical observations and thus any sense of the hurdles that need to be overcome. And when we look for inspiration to the binary classifiers in information retrieval we see that to overcome the hurdles requires a degree of mathematical expertise that is beyond most of those in jurisprudence. The clearest way through I can see is to begin by describing *how* to construct the measure, move on to illustrate *what* it does, and finally establish the reasons for its *applicability* to the criminal justice system.

3.1 How

The *F-measure* is the harmonic mean of precision and recall, that is:

F = 2 * (Precision * Recall) / (Precision + Recall).

Or, more concisely:

F = 2PR / (P + R).

It's a kind of average. So it always lies in between the values of precision and recall, and in the special case when precision is the same as recall, the F-measure is also the same.

If in the equation above we replace recall and precision with their definitions in terms of tallies we can rewrite the above equation as:

F = 2TP / (2TP + FP + FN)

This is reminiscent of the structure of precision and recall. With the F-measure you can improve your performance by reducing false positives, reducing false negatives or increasing true positives. There are three tallies involved and an improvement in any of them will lead to an improvement in the F-measure.

However, the F-measure is not the only way to combine precision and recall into a single indicator and each different method reflects different choices, which are not value free. The harmonic mean gives equal weighting to shifts in precision and recall, and hence to the two kinds of error that are possible in the system, conviction of the innocent and escaping of the guilty. In other words, all else being equal, a reduction of 1 in either error gives the same improvement to the harmonic mean.

On the other hand, to give false convictions more weight than false acquittals we need to give precision more weight than recall. The standard way to do this in information retrieval is through the following formula for the F_β-measure, or simply F_β:

$$F_\beta = (1 + \beta^2)PR / (\beta^2 P + R)$$

We can think of F_β as giving recall β times as much weight as precision. Values of $\beta < 1$ emphasize precision, while values of $\beta > 1$ emphasize recall. So we might give β the value 3 if recall is to be emphasized. Thus if we have a criminal justice system that prizes precision more than recall, we should expect a value of β that is less than 1.

Once again in the special case when precision equals recall they both also equal F_β. If we set β at 1, then we have the F-measure or what we can now call the F_1–measure.

Again, we can rewrite this in terms of tallies to get:

$$F_\beta = (1 + \beta^2)TP / ((1 + \beta^2)TP + \beta^2 FN + FP)$$

Here we can see that values of $\beta < 1$ emphasize false positives so that an additional false positive does more damage than an additional false negative, while values of $\beta > 1$ emphasize false negatives.

3.2 What

If we set β at $1/\sqrt{10}$ (one over the square root of 10), then we get the following:

$$F_{1/\sqrt{10}} = 1.1 PR / (0.1P + R).$$

Notice that the balance between the weight given to recall and precision in the denominator is 10:1 so that precision is 10 times more important to the final result than recall.

Now let us recast this in terms of the tallies:

$$F_{1/\sqrt{10}} = 1.1TP / (1.1TP + 0.1FN + FP).$$

In this case, we see that the balance between the weight given to false positives and false negatives in the denominator is again 10:1. So we see that a single additional false positive will do 10 times as much damage to the score as one false negative and we can say that false positives are 10 times as important to the final result as false negatives. However – and this is

a critical difference – the value of the indicator is *not* fully determined by false positives and the false negatives; the true positives also have their say.

Thus we have eventually found our way back to n and can say that $n = 1/\beta^2$. However, we do not in the F_β-measure have an indicator of "error distribution" that deals purely with the false positives and false negatives. We have an indicator of overall epistemic effectiveness in which the two errors are considered together with the achievement of the true positives. This reflects the nature of our predicament. We can only trade a false positive for a false negative in our totals by making changes to policies, and this will impact also on the true positives.

This difference between 2-ness and 3-ness impacts on the way we should talk about the F_β-measure.

Accurate statement:

> The F_β-measure is a way of establishing a balance between precision and recall.

Inaccurate statement:

> The F_β-measure is a way of establishing a balance between true positives and false positives.

What makes the second statement inaccurate is that, unlike the first statement, it is an *incomplete* description of the F_β-measure that ignores the true positives.

If we want to talk in terms of tallies rather than ratios, then the best way to interpret β is as establishing an *exchange rate*, as between currencies. To be precise, β^2 is the rate at which we are prepared to trade false negatives for false positives. So if we use this measure we accept that 1 false negative and β^2 false positives have the *same* value to us; and, everything else being equal, we will be equally happy with systems that output errors composed of any mix of the two – so long as the sum of the two at this exchange rate comes out the same. An oddity of this exchange rate however is that it is embedded in a system that does not allow us to directly trade one currency for the other. In practice, changes to policies will not leave everything else equal; they will impact also on the true positives.

3.3 Applicability

Ultimately the adoption or not of the F_β-measure is a question first of tractability and then of appropriateness. In order to get a meta-rule we must have a measure; if we refuse to make any assumptions we cannot derive a measure and the problem becomes intractable. Then, there is no measure that is intrinsically "best". So the use of the F_β-measure stands or falls on the arguments presented here. In total, there are nine.

First, the derivation of the F_β-measure. Historically, the F_β-measure, or simply F_β, became established in information retrieval following Van Rijsbergen's derivation of it. He achieved this with some sophisticated and, at the time, new techniques from the mathematics of

measurement. The core idea was to derive the F_β-measure by deploying the principle of decreasing marginal effectiveness, which has proved so powerful in economics and was the basis of Tribe's indifference curves. His arguments themselves are too mathematically rarified to go into here but his contextualizing of the work bears repeating at length. He fully understood that the mathematics itself was not sufficient to make the case for what we now call the F_β-measure, especially given the lack of what he calls an *empirical ordering*:

> The problems of measurement in information retrieval differ from those encountered in the physical sciences in one important aspect. In the physical sciences there is usually an empirical ordering of the quantities we wish to measure. For example, we can establish empirically by means of a scale which masses are equal, and which are greater or less than others. Such a situation does not hold in information retrieval. In the case of the measurement of effectiveness by precision and recall, there is no absolute sense in which one can say that one particular pair of precision-recall values is better or worse than some other pair, or, for that matter, that they are comparable at all. However, to leave it at that is to admit defeat. There is no reason why we cannot postulate a particular ordering, or, to put it more mildly, why we can not show that a certain model for the measurement of effectiveness has acceptable properties. The immediate consequence of proceeding in this fashion is that each property ascribed to the model may be challenged. The only defense one has against this is that:
>
> (1) all properties ascribed are consistent;
>
> (2) they bring out into the open all the assumptions made in measuring effectiveness;
>
> (3) each property has an acceptable interpretation;
>
> (4) the model leads to a plausible measure of effectiveness.
>
> It is as well to point out here that it does not lead to a unique measure, but it does show that certain classes of measures can be regarded as being equivalent.

This acknowledges the unreasonable effectiveness of mathematics previously noted by physicists and allows that anyone may reject his measure of effectiveness but at the same time is explicit about the strengths of it. All of Van Rijsbergen's argument here, both virtues and caveats, will apply to us if we adopt this measure. [46]

Thus the F_β-measure is not unique and the results that come from it do not reflect an empirical ordering provided by Mother Nature. Thus it is an example of what Hand calls

---

[46] Eugene Wigner, *The Unreasonable Effectiveness of Mathematics in the Natural Sciences*, 13 Communications in Pure and Applied Mathematics 1-14 (1960).

"pragmatic measurement".[47] But this does not constitute a reason for rejecting the F_β-measure. In principle it would be bizarre for the law, an invention of men and women, to reject the F_β-measure on the grounds that it has the same origin. Measuring is what we are trying to do here; we should not object to using the province of mathematics developed precisely for that purpose.

We can also consider a simpler, more intuitive justification. This starts by noting that although we talk in terms of precision and recall, we could with equal justification refer to their reciprocals, not P and R but 1/P and 1/R, call them P' and R'. If we used these to measure performance they would reveal to us exactly the same insight as their usual cousins. Equally, we could refer to 1/F and $1/F_\beta$, which we could call F'_β. Now, it so happens that we can re-write the equation that defines F_β like this:

$$\frac{1}{F_ß} = \frac{ß^2}{(ß^2+1)} * \frac{1}{R} + \frac{1}{(ß^2+1)} * \frac{1}{P}$$

Using our alternative reciprocal versions of the indicators, this is:

$$F'_ß = \frac{ß^2}{(ß^2+1)} * R' + \frac{1}{(ß^2+1)} * P'$$

Thus F'_β is the simplest possible kind of weighted average of two quantities, the weighted arithmetic mean. In this sense, F'_β does not require a derivation; it is merely the ordinary way we routinely combine two into one with some tilting. And if that justification stands for F'_β, it stands equally well for F_β.

Thus we will take our ability to derive F_β from first principles as the first argument in favor of its adoption.

Second, an elemental point is that the criminal justice system is a binary classifier and can be treated in the same way as other binary classifiers in fields such as diagnostics and information retrieval. There are of course a thousand ways in which these systems differ from each other, but the 4-gram (or 3-gram) abstracts from them a core that they have in common. All these systems are engaged in a difficult task of discrimination. All can make mistakes of two kinds. There simply are no grounds for believing that the criminal justice system behaves in a way that makes it resistant to the techniques we apply to other binary classifiers.

Third, the F_β-measure is based on the basic statistical framework we adopted, precision and recall and the yardstick of true positives. These align with our natural concerns about the

[47] David Hand, Imperial College, personal communication.

nature of achievement in the system, the practical possibility of measurement and the incentives any indicators give to actors in the system.

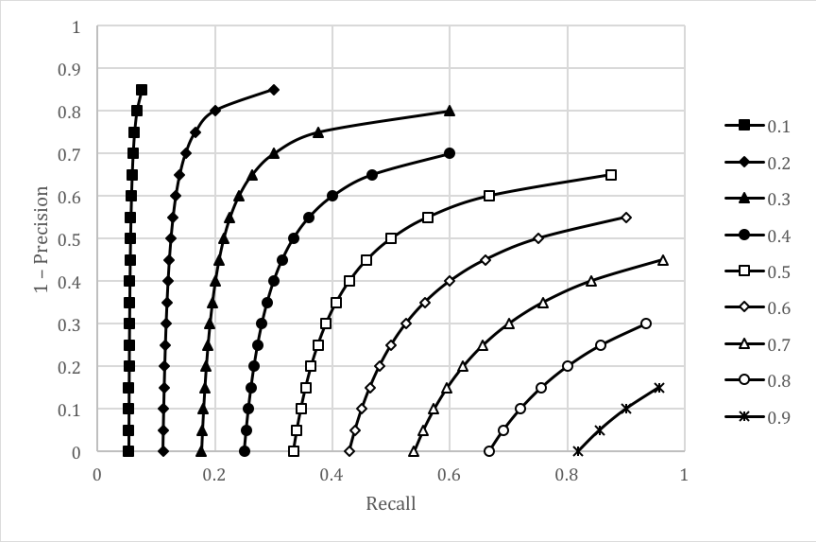
Fourth, the long jurisprudential tradition of considering the ratio of false negatives to false positives is not abandoned but secured. *n* appears as an input to the measure and the exchange rate interpretation is natural. Indeed, it is a strikingly delightful aspect of the F_β-measure that the value judgement of *n* has a natural and simple interpretation when considering the formula both in terms of ratios and tallies. This is not an approach that is alien to established legal reasoning concerning the trade-offs between false acquittals and false convictions; rather, it simply conceptualizes the trade-offs in a properly constrained way.

Fifth, this allows us to derive the indifference curves that are the basis of Tribe's approach. The two axes of Tribe's charts are the percentage of innocents convicted and the percentage of guilty convicted. [48]

Percentage of innocents convicted = FP/(TN+FP) = 1 – Specificity.

Percentage of guilty convicted = TP/(TP+FN) = Recall.

For the good reason that it uses the yardstick of true positives rather than true negatives, we chose precision rather than specificity. If we therefore consider not Tribe's space but the corresponding one defined by the axes of 1 – precision and recall, we can draw on this contours that represent equal values of effectiveness as measured by the F_β-measure. These are indeed curves of indifference to us, and all flow from a single value judgement. The chart below shows such contours for increments of 0.1 in the value of the F_1-measure.



[48] Tribe *supra* note 14, at 1388.

Figure 3.1

Sixth, it is implied by Hand's typology. A useful check on our derivation is to work through the typology of measures in his review. Firstly we can exclude the problem-specific techniques that have been developed, for example, to allow the classifier to be rapidly updated (antispam filters) or to handle very large data sets (particle physics). Of the more general techniques based purely on the informal notion of "accuracy", we only need the first section, where the confidence threshold is determined in advance, which eliminates DeKay's signal detection theory and many others. This leaves us with the Kappa statistic, the Youden index and the F-measure. Both the Kappa and Youden are ways of combining specificity and sensitivity, and so would take us back to a mistaken reliance on the true negatives.[49]

Seventh, within information retrieval Van Rijsbergen's argument that it not only improves on the previous measures but also does the job we need has been fully embraced. Today it is routinely included in textbooks as the only method for assessing unranked sets of results. [50] The verdict of the jury of this discipline is: it works. Further, where there are criticisms of the F_β-measure's usage, for example from Powers, these are accompanied by the recognition that it remains indicated for situations where we are focusing on one yardstick of true results at the expense of the other, as we are here [51].

Eighth, the F_β-measure incorporates the natural principle, already selected by Tribe, of decreasing marginal effectiveness as the basis for balancing precision and recall.

Ninth, Van Rijsbergen has shown that it is either equivalent to or superior to a wide range of alternative indicators that were put forward in the context of binary classification in information retrieval in the 1960s and 1970s.

To sum up, while the F_β-measure does not claim to be unique, for the criminal justice system it is uniquely appropriate.

3.4 As a meta-rule

We now have via F_β, as we may call the F_β-measure from now on, a viable meta-rule that can handle the trade-offs between false convictions and false acquittals and make decisions on

---

[49] Hand *supra* note 15.

[50] See for example Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, *Introduction to Information Retrieval*, online edition, Cambridge University Press (2009). See section 8.3: "Evaluation in Information Retrieval" where the F_β-measure is the only technique presented for unranked sets of results.

[51] David M. W. Powers, *What the F-measure doesn't measure: Features, Flaws, Fallacies and Fixes*, arXiv:1503.06410v1 (2015).

policies. The methodology is to pick a value, b, for β and then, by adjusting the policy in one way or another, to aim to maximize F_b. The version of the policy that gives F_b the highest score is the one we adopt. This is the usual approach in information retrieval where determination of the maximum value for F_b is routinely interpreted as an attempt to find the best possible compromise between precision and recall.[52] And it allows us to deal with all epistemic policy dilemmas and the standard of proof.

We can illustrate our new approach with data from a real binary classifier algorithm where the policy that we adjust is the confidence threshold that the algorithm must reach in each case before it classifies a document as a positive result.[53] This is analogous to varying the standard of proof.

To perform the calculations, let us return to the results from the Maui algorithm we looked at before in order to generate the sample curves for precision and recall. To generate those results, we compiled a test bed dataset of documents that have been manually classified as relevant or not. In this case, the universe consists of 1860 documents that are relevant and 8298 that are not.

This is an arbitrarily selected algorithm. All that is important for us however is that it is a binary classifier that shares the qualities we expect to find, particularly the tension between precision and recall such that, as we adjust the confidence threshold, one rises and the other falls.

The threshold can vary from 0 to 1 and if we increase it in increments of 0.1, then we get the following table of results, from which we can calculate precision, recall and values for F_b for any value of b. In this case we have chosen F_1.

[52] Ricardo Baeza-Yates and Berthier Ribeiro-Neto, *Modern Information Retrieval: the concepts and technology behind search,* second edition, Addison Wesley, 144 (2010).

[53] Medelyan *supra* note 43.

Threshold	TP	FN	TN	FP	Precision	Recall	F_1
0.10	1517	343	3720	4578	0.25	0.82	0.38
0.20	1069	791	6738	1560	0.41	0.57	0.48
0.30	776	1084	7554	744	0.51	0.42	0.46
0.40	560	1300	7913	385	0.59	0.30	0.40
0.50	410	1450	8109	189	0.68	0.22	0.33
0.60	288	1572	8197	101	0.74	0.15	0.26
0.70	203	1657	8245	53	0.79	0.11	0.19
0.80	121	1739	8278	20	0.86	0.07	0.12
0.90	46	1814	8296	2	0.96	0.02	0.05

Figure 3.2

We can then draw the curves for precision, recall and F_1.

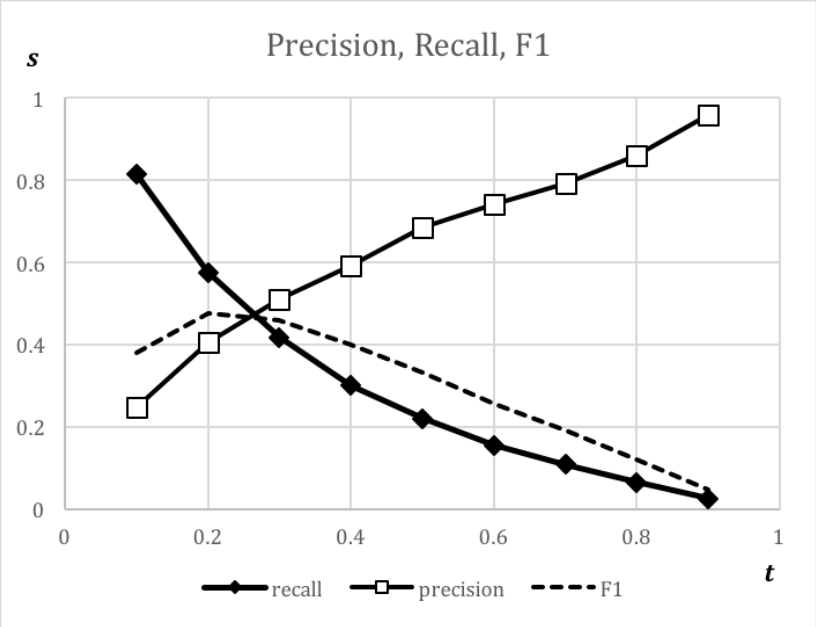


Figure 3.3

As can be seen, the confidence threshold that maximizes F_1 is 0.2. Thus if we have chosen b = 1 to represent our value judgment about the relative importance of false acquittals and false

convictions, then setting the confidence threshold at 0.2 is the policy choice we should make.[54]

Not all policy questions lend themselves to the drawing of curves in the same way as the standard of proof. We may have policy questions where it is not obvious to us that there is obviously a natural ordering of the available options. But this does not stop us picking the option with the highest value for F_b.

We can now provide a rigorous chain of reasoning to justify (or not) statements in the form "because *n*, X". To be precise, these statements will now take the form "because the sovereign has chosen to give recall *b* times as much weight as precision, policy X follows." Or, which I prefer because it is easier for the lay person to understand, we can switch to the exchange rate interpretation and say "because the sovereign has chosen to value false negatives b^2 times as highly as false positives, policy X follows", for which we can use the shorthand, "because b^2, X".

This meta-rule can be applied to any policy question, including the edifice built on the Blackstone principle, the epistemic policy dilemmas and the standard of proof. Equally, within our epistemic frame of reference, we can't get anywhere without first choosing a value for *b* and hence all policy arguments are of the form "because b^2, X".

If we return to Laudan's definition of a meta-epistemology as "a body of principles that will enable us to decide whether any given legal procedure or rule is likely to be truth-conducive and error reducing" then we see that our methodology based on F_β provides Laudan with everything he asked for. It provides not just a meta-rule to deal with some questions, but a way to deal with all of them. It is a meta-epistemology.

The methodology has three elements:

- A value judgement, b^2, which represents the exchange rate at which we are prepared to trade the two errors
- the F_β-measure as the mechanism of evaluation
- measurement of precision and recall under different policy regimes.

Three points are worth noting. First, different values of *b* will lead to different choice of policies by way of a different numerical evaluation of policies. In other words, each distinct value for *b* gives rise to a distinct meta-epistemology. So what we in fact have devised is a meta-meta-epistemology that provides a spectrum of meta-epistemologies, each defined by a value of *b*.

[54] This is a crude approach sufficient for our purposes; improvements in accuracy should be expected from interpolation.

Second, there is no reason to expect to be able to treat policies as independent. Looking to information retrieval, if the bag of words in the Maui algorithm is empty then adding the word "law" to the bag will dramatically improve effectiveness, but if the bag already contains many more precise words, it may decrease effectiveness. Consequently, each policy will need to be considered in context. Thus Laudan's original objective of assessing each policy for the extent to which it is truth-conducive or not becomes considerably harder than we might have hoped.

Third, there is no particular reason why the curves for precision and recall should be the same in different subsets of our universe, for example for different types of offence. Where differences in the curves are large, we cannot both keep the policies of the system and the standard of proof the same *and* continue trying to maximize F_b for the same value of *b*. In information retrieval, we might well tune the algorithm differently for different kinds of documents, distinguishing for example between a repository of scholarly papers and newspaper articles. If we want to keep the policies and standard of proof the same, then we must accept we are abandoning *b* and with it our value judgement about the relative importance of precision and recall. Alternatively, if we want to continue with *b*, we will be obliged to vary either the standard of proof or the policies.

4. FURTHER WORK

This work prompts a number of questions.

Can both the established quantitative approaches inspired by Kaplan and this be right at the same time? No, the thrust of this article is that, from the sovereign perspective, true positives and true negatives cannot be treated equally, and hence the characteristic narrowing move which treats the two as equal is evidently invalid. The underlying difference is that between output and input. By aiming to manage the distribution of errors, today's followers of Kaplan treat *n* as an objective to be achieved while this treats it as a value judgement, highly consequential but not the consequence itself.[55]

Given that it has been asserted that the effectiveness of the courts cannot be measured, it is clearly essential now to establish whether, with our new understanding, we can overcome this hurdle. Of course, we agree in fact that the system's effectiveness has in the past been unmeasurable – the absence of a computational framework such as this one is proof of it. But now we have dealt with that problem, will that be enough? After all, we still need to obtain trustworthy observations of outcomes, including both kinds of error, for our four tallies.

[55] I look into Kaplan's impact further in *Flawed Methodologies, Kaplan and the Standard of Proof: Killing Dahlman, Hamer, Kaplow, Laudan, Lippke, Stein, Stewart and Walen*, submitted for publication.

If we look to the natural sciences, then it is clear that our situation in the law is different. Just as Mother Nature has provided no empirical ordering for our computational framework (as noted above by Van Rijsbergen), so she declines to provide us with what we might call an "empirical verdict" on the decisions our system makes, something she does offer the maker of a pregnancy test. However, this deficiency is not unique to the law. For example, if I say a book about sovereignty belongs in the law section of the library and you say it belongs in international relations, who is to say that you are right and I am wrong? So indexing is an example of another domain that suffers from the same problem. I will argue that by borrowing further techniques from this other domain of information retrieval we can indeed obtain trustworthy measurements.[56]

Then, we have here in our chosen meta-epistemology a way of measuring the performance of the entire system. If we want to make a decision on any one policy question, we could try to hold everything else steady and try out the different options. In systems as large, complex and delegated as our criminal justice systems currently are, that is a challenging prospect, to say the least. We will need ways to decompose the system into distinct elements where performance can be measured separately and decisions taken independently. I will set out a preliminary approach.[57]

Finally, the upshot of this work can be seen as allowing "because *n*, X" arguments to finallly plant both feet in the modern world. There are philosophical, constitutional, moral and political questions that arise, to which I intend also to return. I think the enduring incoherence of evidence law[58] is a consequence of a failure to democratize the law. This might usually be seen as a choice but I view it as a necessity, forced upon the law by its inability to mathematize its core epistemic concern and hence embark on the distinctively modern project of empirical measurements. Ultimately, instead of contenting ourselves with consistency of process we could aim for better outcomes, with potentially dramatic consequences for "dark cases"[59] that usually occur behind closed doors, such as rape.

The ambition then is to re-establish criminal law on new foundations, which entails abandoning some of the old foundations. Freed from the necessity of relying on the old foundations, we would be at liberty to consider the extent to which we really admire them.

---

[56] See *Measuring Justice* (working title), in preparation.

[57] See *Epistemic Management of the Criminal Justice System* (working title), in preparation.

[58] For example Hamer reviewing Stein says the book "... provides a telling demonstration of the difficulty of bringing coherence to evidence law. Discretion being the better part of valour, this is an area from which the law may continue to remove itself." David Hamer, *The Truth Will Out? Incoherence and Scepticism in Foundations of Evidence Law*, 70(2) Modern L. Rev., 318-338, 326 (2007).

[59] Shapiro *supra* note 4, at 214.

Nonetheless, until such an appreciation emerges, some may consider this objective, in Tribe's words, "more dangerous than fruitful". [60]

[60] Tribe *supra* note 14, at 1393.