ACCURACY AND STATISTICAL EVIDENCE

Abstract. Suppose that the word of an eyewitness makes it 80% probable that A committed a crime, and that B is drawn from a population in which the incidence rate of that crime is 80%. Many philosophers and legal theorists have held that if this is our only evidence against those parties then (i) we may be justified in finding against A but not against B; but (ii) that doing so incurs a loss in the *accuracy* of our findings. This paper argues against (ii). It argues that accuracy considerations can motivate taking different attitudes towards individualized and statistical evidence even across cases where they generate the same probability that the defendant is guilty.

1. Introduction

Both intuition and the actual findings of courts distinguish between the force of 'individual' and that of 'statistical' evidence on grounds that seem unconnected to their accuracy. Consider these two cases:

> **A**: The organizers of a rodeo sue Alice for gate-crashing their Saturday afternoon event. Their evidence is as follows: Alice attended the event—she was seen and photographed on the main ranks. No tickets were issued, so she cannot be expected to prove that she bought a ticket with a ticket stub. However, a local police officer observed Alice climbing the fence and taking a seat. The officer is willing to testify in court. Careful testing has shown that this officer's testimony about such matters in such circumstances is correct about 90% of the time.
>
> **B**: The organizers of the local rodeo decide to sue Bob for gate-crashing their Sunday afternoon event. Their evidence is as follows: Bob attended the Sunday afternoon event—he was seen and photographed on the main ranks. No tickets were issued, so Bob cannot be expected to prove that he bought a ticket with a ticket stub. However, while 1,000 people were counted in the seats, only 100 paid for admission.[1]

Courts that are supposed to make findings based on what is more likely than not—as, in civil cases, actual courts in the US and the UK do—would in fact find against Alice in case A but *not* against Bob in case B. And intuition seems to accord with this. But the evidence against Alice makes *her* guilt no more likely the evidence against Bob makes *his*; and in both cases the balance of evidence favours a finding against the defendant.

The relevant difference between the cases seems to be that the evidence against the defendant is in some sense 'individualized' in case A and in some sense 'statistical' in case B, as is borne out by the fact that people exhibit parallel differences in their responses to other pairs of cases that clearly fall on either side of this vaguely defined line. Slightly more precisely: the evidence in case A seems to be about Alice and the offence that she specifically is supposed to have committed, whereas the evidence in B seems to have no more bearing on Bob's alleged gate-crashing than on that of the 900 (or 899) other miscreants in attendance on that day.[2]

---

[1] Adapted from Cohen 1977: 74-81.

[2] Here are some examples: (i) A bus known to belong either to the Blue Bus Co. or the Green Bus Co. causes an accident. An eyewitness with 90% reliability claims that it was clearly a Blue Bus Co. bus that was involved. (ii) Like case (i), except that the evidence is just the fact that an expert witness testifies that an identification technique based on matching tyre tracks with tyres is correct 90% of the time, and that based on this technique he believes that the bus involved in this accident belonged to the Blue Bus Co. (iii) Like case (i), except that the evidence is just the fact that 90% of the bus traffic on the route where the accident took place belongs to the Blue Bus Co.

A substantial literature offers both vindicatory and debunking explanations of the difference in our reactions to these and similar cases.[3] Vindicatory explanations include the following: (i) that there is no good reason to infer a probability of guilt from the evidence in case B[4]; (ii) that the evidence in case A supports counterfactuals that the evidence in case B does not[5]; (iii) that finding against a defendant on statistical evidence is inconsistent with respecting his autonomy[6]; (iv) that the evidence can support a self-ascription of knowledge that A is guilty but not a self-ascription of knowledge that B is[7]; (v) that a finding against Bob could 'easily' be mistaken, whereas a finding against Alice could not.[8] Debunking explanations include the following: (vi) that our subjective probability of guilt in B-like cases fails to match the mathematically correct probability[9]; (vii) that we get the probabilities right but are misled by the form in which the evidence is presented.[10]

But there is one thing on which advocates of all these views, except possibly (i), agree: *considerations of accuracy alone cannot motivate different approaches to A and B*. Anyone who cared only about the accuracy of the court's findings should be in favour of finding against *both* Alice and Bob (given a relatively lenient standard of accuracy) or against *neither* of them (given a relatively strict standard). For instance, Enoch et al. write of 'the loss in accuracy that is always involved in ruling out probabilistically respectable evidence'.[11] Koehler and Shaviro write:

> From the standpoint of verdict accuracy, the equivalence between background and case-specific evidence is difficult to dispute. Even cases involving 'naked statistical

(iv) Like case (i), except that the evidence is just the fact that nine times more *accidents* on that route are due to buses operated by the Blue Bus Co. than are due to buses operated by the Green Bus Co. (iv) Like case (i), except that the evidence is just the fact that tyre tracks taken from the scene of the accident match 9 of the ten Blue Bus Co. buses and only one of the ten Green Bus Co. buses. Wells (1992) presents evidence that most people, including most of the practicing trial judges questioned, are prepared to find against the Blue Bus Co. in cases like (i) and (ii) but not in ones like (iii)-(v). One glaring difference between the cases that elicit a finding of liability, and those that do not, is that in the former cases the evidence explicitly involves reference to the incident itself rather than to general statistics that are equally concerned with other, similar incidents.

[3] Other B-like cases, summarized in Redmayne 2008: 282-3, include: (i) *Prisoners*: 24 of the 25 prisoners in a yard kill the prison guard – should we convict a prisoner chosen at random from those 25? (Nesson 1979: 1192-3.) (ii) *Predicting Violence*: studies show that 50% of males who are brought up in broken homes, are unemployed and addicted to drugs go on to commit violent crimes. Can we use this fact as evidence against a defendant who possesses these characteristics? (Duff 1998: 156.) (iii) *People v. Tice*: 'Two people, Tice and Simonson, both hated Summers and wished him dead. Summers went hunting one day. Tice followed with a shotgun loaded with ninety-five pellets. Quite independently, Simonson also followed, but *he* had loaded his shotgun with only five pellets, this being all he had on hand. Both caught sight of Summers at the same time, and both shot all their pellets at him… Only one pellet hit Summers, but that was enough: it hit Summers in the head and caused his death. While it was possible to tell that the pellet which caused Summers' death came either from Tice's gun or from Simonson's gun, it was not possible to tell which' (Thomson 1986: 200-1). Do we have evidence on which to convict Tice?

[4] Allen and Pardo 2007 sect. 3.

[5] Cohen 1986: 165; Enoch et al. 2012.

[6] Wasserman 1991: 943.

[7] Thomson 1986: sections IV-V; Blome-Tillman 2017. Blome-Tillman's approach seems to me to offer an especially promising way to recover our intuitions in so far as they are worth recovering, assuming that one buys into the 'knowledge first' program in epistemology. My aim here, as will shortly become clear, is not to compete with it but rather to attack the claim that dispensing with statistical evidence inevitably incurs a cost in *accuracy* (whether or not it incurs a cost in *knowledge*).

[8] Pritchard forthcoming.

[9] Tribe 1971.

[10] Neidermeier and Messé 1991; cf. Redmayne 2008: 304.

[11] Enoch et al. 2012: 219.

> evidence' (i.e., a base rate unaccompanied by other evidence [as in case B]) should not be treated differently from other cases if one's sole concern is verdict accuracy.[12]

Similarly, Nesson regards accuracy in verdicts as just one of the aims of a trial, to be balanced against that of producing an 'acceptable' history of the events being investigated. But he grants that rejection of 'statistical' evidence in the light of these other aims imposes a cost in accuracy:

> Because the judicial system strives to project an acceptable account about what happened, then, the [base rate] evidence is insufficient, notwithstanding the high probability of its accuracy… One who is absolutely committed to the process of ascertaining and testing the truth, and who would thus shun any concession of the search for truth to the production of acceptable verdicts, may find that he does so at the expense of other important values.[13]

Again, Brook, writing about case B, concedes that:

> If minimization of errors simply in terms of reducing the total number of wrong results is to be the only fundamental criterion of successful fact-finding in civil litigation, then traditional probability theory, properly applied and understood, points us to the right result [i.e. finding against Bob], however harsh it may sometimes seem.[14]

Brilmayer states in similar terms why a concern for accuracy demands finding against Bob:

> [T]o deny recovery [to the organizers of the rodeo] would increase unnecessarily the number of errors in the long run. Holding each rodeo spectator liable for trespass will result in [900] correct decisions and [100] incorrect decisions. Disallowing liability will result in only [100] correct decisions but [900] incorrect ones.[15]

Brook's and Brilmayer's argument applies to Alice as much as to Bob. If it establishes anything at all then it commits us to treating these cases alike.

We can express what is common to all these views by means of a formula:

> **Accuracy-indifference (AI)**: Accuracy considerations cannot by themselves justify a difference in finding between case A and case B.

On the fact of it, (AI) is highly intuitive. After all, just as Brilmayer says, if you repeatedly find against people like Bob because of a base rate exceeding 50%, you will get it right just as often as if you repeatedly find against people like Alice because of an eyewitness whose reliability exceeds 50%. Considerations of accuracy are indifferent to the distinction between individual and statistical evidence.

I'll argue here that (AI) is false. The argument involves a simple model in which accuracy considerations *do* by themselves motivate finding against Alice but not against Bob. Wherever we draw the exact line between statistical and individual evidence, that line must put the evidence against Alice on the 'individual' side and the evidence against Bob on the 'statistical' side. Since this is plausibly the only relevant difference between case A and case

[12] Koehler and Shaviro 1990: 264.
[13] Nesson 1985: 1379, 1392.
[14] Brook 1985: 322.
[15] Brilmayer 1979: 676. I have altered the numbers to fit with my own example.

B, the model is therefore also a counterexample to the claim that an exclusive concern for accuracy demands indifference to this distinction.

Section 2 describes the basic elements of the model. Section 3 gives the argument for finding against Alice but not against Bob. The model is highly idealized in the interests of clarity, in the sense that (a) it makes somewhat unrealistic assumptions about the details of the case; (b) it misses out various complicating factors that we'd expect to be present in real cases; (c) it is explicitly focused on just one type of comparison of individual and statistical evidence. Subsection 3.2 discusses what happens if we (a) relax the assumptions; section 4 looks at the effect of (b) adding some complicating factors and (c) considering a different way in which we might compare these types of evidence for accuracy. Section 5 concludes.

2. The model

This model has four elements: (i) a threshold for finding against the defendant; (ii) an option set; (iii) a distribution for the probability of guilt that eyewitness evidence produces; (iv) a distribution for the probability of guilt that base rate evidence produces.

2.1 The Threshold

The model is, as I'll say, *semi-Lockean*. This means that a person is convicted only if the evidence supports a probability of guilt that passes a certain threshold c. For instance, the requirement that guilt be ensured 'beyond a reasonable doubt' might be interpreted as meaning that the probability of guilt must exceed 95%, so that $c = 0.95$; in UK *civil* cases what matters is the balance of evidence, i.e. $c = 0.5$. The precise value of c will not affect the argument, but for convenience I'll set $c = 0.8$ as the relevant threshold.

It matters that a threshold probability of c does not imply that the proportion of convictions that are correct is c. Rather, a threshold of c implies a rate of true conviction that exceeds c; by how much depends on the kind of evidence. For instance, suppose $c = 0.8$. And suppose that our method for determining guilt is to cast a magic die with faces labelled 1-10: if the number shown is n then the probability of guilt is $n/10$. This is because over many trials it has turned out (in light of subsequent findings) that the defendant was guilty in 10% of trials in which the die showed 1, in 20% of those in which it showed 2, and so on; and the laws of magic give us every reason to expect things to continue this way. $c = 0.8$ implies that we convict if and only if the die shows 9 or 10. But if the die shows each number equally often, then this means that on this policy, the true conviction rate is 95%, not 80% (or 90%): that is, 95% of convictions are correct. The obvious point that this fictional example illustrates will be of vital importance in what follows.

2.2 The policy problem

I'll start by assuming that the situation is as follows. We have the resources to punish a fixed number D of gate-crashers every year. Because we are semi-Lockean, we can only punish those against whom the evidence supports a probability of guilt of at least 80%. Every year there are many more than D visitors to the rodeo for whom the eyewitness evidence supports such a probability of guilt; this year that number includes Alice. And there are many more than D such visitors for whom statistical evidence, of the sort found in case B, supports such a probability of guilt. This year that number includes Bob. Let us suppose that the total number of such visitors in each category is K, where $K \gg D$.

In this idealized scenario, the policy problem is to choose a proportion α, $0 \leq \alpha \leq 1$, such that of the D people that we convict of gate-crashing on either eyewitness or statistical evidence, $D\alpha$ are convicted on eyewitness evidence and $D(1-\alpha)$ are convicted on statistical evidence. For instance, we might choose $\alpha = 0.3$, so that each year 30% of convicted persons

are convicted on the sole basis of eyewitness evidence that implies a probability of guilt exceeding 80%, and 70% on the sole grounds that they attended the rodeo on a day when at least 80% of those who attended did so without paying. The aim is to choose the policy with maximal accuracy i.e. the value of α for which the proportion of convictions that are *true* convictions is maximal, given that we are going to convict D people.

2.3 Distribution of a

In cases where we have a positive identification from an eyewitness, the probability of guilt is determined as follows. (i) Form an estimate p of the probability of Alice's guilt given positive identification by 'the average' eyewitness. (This is something for which we shall have to rely on pre-existing statistics, or failing that common sense.) (ii) Create a large population of photographs of persons in situations (lighting, angle of view etc.) like those obtaining at the rodeo (or wherever else the incident took place), in which a proportion p of photographs are of Alice herself.[16] (iii) Draw photographs at random from this population and ask the witness whether the person in the photograph is Alice; repeat until the witness has made some very large number N of positive identifications. (iv) If M is the number of *true* positive identifications that the witness has made, then $a = M/N$.[17] We convict on this evidence if and only if $a > c$, where c is the threshold for conviction as described above.

For instance, suppose that we think that the average eyewitness's positive identification of Alice is correct 25% of the time. So (i) we form an initial estimate that the probability of guilt, given positive ID from the 'average eyewitness', is 25%. (ii) We form a population of photographs of people climbing fences at rodeos; 25% of these are photographs of Alice herself. (iii) After seeing very many of these photographs, the eyewitness identifies $N = 100$ of these as photographs of Alice. (iv) It turns out that $N = 90$ of those 100 photographs are

[16] More generally, a proportion q of photographs have visible feature F if our initial estimate is that a proportion q of positive courtroom identifications of Alice by the average eyewitness would identify her with a person who has property F. So for instance, if Alice is more than 6' tall, none, or vanishingly few, of the photographs should be clearly of individuals who are less than 5' tall.

[17] Here is a brief mathematical justification for the procedure. Suppose that given the eyewitness testimony but prior to any testing of the eyewitness, we start out with probability p that Alice is guilty. And suppose that we show the witness a very large number T of photographs, of which N result in positive identifications, M of which are *true* positive identifications. Then the results may be tabulated as follows:

	Positive ID	**No positive ID**
Photo is of Alice	M	$pT - M$
Photo is not of Alice	$N - M$	$(1-p)T - N + M$

In the body of this table, the entries in the first column follow from the description of the case; the entries in the second column follow from the fact that for large T, a proportion p of the T photographs that the witness sees will be photographs of Alice. Now writing G for the proposition that Alice is guilty, Y for the proposition that this eyewitness has positively identified her, we can form the following estimates from the data given in the table:

(i) $\Pr(Y|G) = \frac{M}{M+(pT-M)} = \frac{M}{pT}$

(ii) $\Pr(Y|\neg G) = \frac{N-M}{N-M+((1-p)T-N+M)} = \frac{N-M}{(1-p)T}$

We can now insert these values and $\Pr(G) = p$ into the following formula, which is a theorem of the probability calculus:

(iii) $\Pr(G|Y) = \frac{\Pr(Y|G)\Pr(G)}{\Pr(Y|G)\Pr(G)+\Pr(Y|\neg G)\Pr(\neg G)}$.

This gives $a = \Pr(G|Y) = M/N$.

indeed of Alice; so our probability of guilt following testing is $a = M/N = 0.9$. We convict on this evidence if and only if $0.9 > c$.

A natural objection is that the whole procedure is highly unrealistic: it has very little to do with the ways in which we actually assess eyewitness evidence. It *is* unrealistic, but there are two things to say about that. First, the aim here is only to show that it doesn't *follow*, from the premise that accuracy is our sole concern, that we must treat cases A and B in the same way. To that end, it's enough to show that *there is* a model that distinguishes them but in which accuracy is the only concern; it isn't necessary to show that the model is an accurate description of reality. (And we knew from the outset that it wouldn't be, since as Bentham complained, some aspects of actual legal practice certainly *are* inconsistent with the aim of accuracy.[18])

Second, the procedure is rationally defensible, at least from a Bayesian perspective (for reasons outlined at n. 17). I submit that setting aside the likely costs of the imaginary procedure, something like it is rationally preferable to the overly credulous attitude towards eyewitness evidence that seems still to be prevalent in many courts.[19] It could therefore be maintained that although the model doesn't explain – and was never intended to explain – why actual courts distinguish A-like cases from B-like ones, it might still cast light on why an overriding concern for accuracy might make it *rational* to do so.[20]

Having set out the procedure that determines a, can we say anything about its distribution? Should we expect its value to be tightly bunched around a single value, so that everyone in the population has $a \approx 0.9$, say? Or should we expect the distribution to be much more spread out, so that we can (e.g.) find equally many people with $a = 0.1$, $a = 0.2$, etc? Clearly the actual distribution of a is an empirical question. But we can say enough about it $a\ priori$ to place constraints on its distribution that are strong enough to make the intended point.

Consider the process that leads to the determination of a: repeated tests of visual identification. Visual identification is a *skill* that is, we normally assume, relatively stable in a single person over, say, a few weeks or months. (If we did not assume this then nobody would regard a witness's skills of visual identification as determined during the trial as indicative of his acuity at the time of the incident.) It follows that there is a correlation between the results of individual tests. To take two extreme cases: if the witness's first n positive identifications are all correct, $n < N$, then it is more likely that the $(n+1)$th positive identification is also correct. If the witness's first n positive identifications are all *in*correct then it is more likely that the $(n+1)$th positive identification is also wrong. Putting this more generally, and in terms of statistics: the relative frequency of test subjects whose $(n+1)$th positive

[18] See Jackson and Doran 2010: 178f. (commenting on Bentham 1978 [1827]).

[19] For a survey of eyewitness research and a history of its uptake within the US legal system, see Wells et al. 2006.

[20] There are two other apparent difficulties with this procedure. First, it is something of a simplification to suppose that we simply identify a with M/N without any 'smoothing' in the light of the background rate of success in the general population. Certainly, it would be absurd to think so when $N = 1$: nobody would say that if an eyewitness makes a single positive identification under test conditions, and it happens to be correct, then that witness's positive ID of a suspect would be sufficient to convict at any threshold. However, the simplification is harmless on the assumption that N is large.

Second, the reasoning might appear to involve the base-rate fallacy by not taking account of the actual incidence of gate-crashing in the population (or at any event in stadia like that into which the eye-witness identified Alice as having gate-crashed). But this is not so: we can imagine that there is no doubt that the witness saw *a* gate-crasher – perhaps because only a gate-crasher would have been in that exact location – and the question is only *which* person he saw. This form of the base-rate fallacy arises only if we mistakenly identify the *converse* proportion – that is, the proportion of guilty people whom the witness positively identifies as guilty – with the probability of guilt given positive eyewitness identification. (For a description of one such case see Bar-Hillel 1980: 211-12.)

identification is correct is higher amongst witnesses whose first n positive identifications have a higher success rate than amongst those whose first n positive identifications have a lower success rate.

Now the important point is that the tighter the correlation between successes in successive instances of the N trials that determine a, the more 'dispersed' is the distribution of a itself. Slightly more precisely: as the correlation between success in past trials and success in the next trial gets stronger, then for any given distance from the mean value of a, an increasing proportion of the population will have an a lying further than that distance from the mean.

To get an intuitive feel for this, consider a crude informal model of the test procedure. Imagine two large populations of coins, P_1 and P_2. Each coin in each population is tossed N times for some very large N and the results of each trial are compiled into a record for that coin. For a given coin i, let a_i be the proportion of its N tosses that were heads. For each $n < N$ we look at the frequency of heads on the first n tosses within each population (i.e. within P_1 or P_2), and we attempt to correlate this with the occurrence of heads on the $(n+1)$th toss in that population.

Suppose that after many tosses our findings are as follows. In both populations, half of the coins land heads on the first toss. But in P_1 there is a correlation between heads in past tosses and heads in the next. Specifically: for each $n < N$ and $m \leq n$, in the subpopulation of P_1 that scored m heads out of n tosses, the proportion of coins that land heads on the $(n+1)$th toss is $(m+1)/(n+2)$. In P_2 by contrast, there is no such correlation: for each $n < N$ and $m \leq n$, in the subpopulation of P_2 that scored m heads out of n tosses, the proportion of coins that land heads on the $(n+1)$th toss is always 0.5.

It follows from these facts that the overall distribution of the a_i is much more dispersed in P_1 than it is in P_2. In P_1 the distribution of the a_i is *uniform*, with mean $N/2$ and variance $\frac{N(N+2)}{12}$: that is, for each integer $x = 0,1 \ldots N$, the proportion of coins i such that $a_i = x$ is $\frac{1}{N+1}$, so that the distribution is flat across its support. For instance, if we toss each coin in P_1 one hundred times, then about 1% of the coins in P_1 always land heads, about 1% land tails once in these hundred tosses, about 1% land tails twice, etc.

But in P_2 the distribution is *binomial*, with mean $N/2$ and variance $\frac{N}{4} < \frac{N(N+2)}{12}$ (if $N > 1$): it is peaked around its mean and falls off rapidly as we move away from the mean in either direction. For instance, if we toss each coin in P_2 a hundred times, a negligible proportion lands heads on every toss; about 1% land tails 40 times and heads 60 times, about 8% land tails 50 times and heads 50 times, about 1% land tails 60 times and heads 40 times, and so on. The proportion of heads is therefore more widely dispersed in P_1 than in P_2. For an extreme illustration of that, note that in P_1 the proportion of coins that land the same *every* time is $\frac{2}{N+1}$; in P_2 it is $\frac{1}{2^{N-1}}$ i.e. very much less.[21]

We can put all this more generally by means of the following formal model. Let us write μ for the mean value of a across the population. And let us suppose that there is a constant real number $\lambda \in [0, \infty)$ such that for any $n < N$ and any $m \leq n$, in a sub-population of individuals for which m out of the first n positive identifications are correct, the proportion of individuals whose $(n+1)$th positive identification is correct is given by:

(1) $\mu_a^{m/n} = \frac{m+\lambda\mu}{n+\lambda}$

[21] Note that the model says nothing about the physical *chances* of any coin's landing heads, nor does it assume that the tossing of a coin is an indeterministic process. As in the overall story about statistical evidence offered here, the objective probability involved is simply frequency.

The quantity λ is thus a measure of the correlation between a high rate of success in any sequence of trials and a high rate of success in the next. When λ is very *small*, the correlation is *strong*. As λ gets *large*, the correlation becomes *weak*. For instance, in the extreme case that $\lambda = 0$, a success rate of m/n in the first n trials implies a success rate of m/n in the $(n+1)$th trial. So in the sub-population that was successful in the first trial, the rate of success in the second (and in any subsequent) trial is 100%; and in the sub-population that was *un*successful on the first trial, the rate of success in the second (and in any subsequent trial) is zero. This is analogous to a large collection of coins that have been tossed N times, of which all land heads every time or tails every time (perhaps because all have an extreme bias). The other extreme $\lambda = \infty$ corresponds to the case where there is no correlation: *any* success rate in the first n trials implies a constant success rate of μ in the $(n+1)$th trial. If $\mu = 0.5$, this is analogous to population P_2 in the example: a large collection of coins that have been tossed N times and in which the success rate has a binomial distribution with mean $N/2$ and variance $N/4$. (Population P_1 in that example corresponds to an intermediate case in which $\lambda = 2$ and $\mu = 0.5$.)[22]

For any fixed values of λ and μ we can determine the distribution of a. Specifically, for large N, the distribution of a is roughly a *beta* distribution with parameters $\lambda\mu$ and $\lambda(1-\mu)$. Its probability density function is given by:

$$(2)\ \beta_{\lambda\mu,\lambda(1-\mu)}(x) = \frac{x^{\lambda\mu-1}(1-x)^{\lambda(1-\mu)-1}\Gamma(\lambda)}{\Gamma(\lambda\mu)\Gamma(\lambda(1-\mu))}$$

– where Γ is the gamma function $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$; for x a positive integer this yields $\Gamma(x) = (x-1)!$

These mathematical details are not necessary for a qualitative understanding of the basic argument. The important point is that the distribution of the statistic a depends, for its dispersion about its mean in a population, on the strength of correlation, for each individual in that population, between the rate of success in past trials and success in the next trial. The formalization in terms of beta functions is just a way of quantifying that point. To illustrate the quantification, consider Figures 1 and 2.

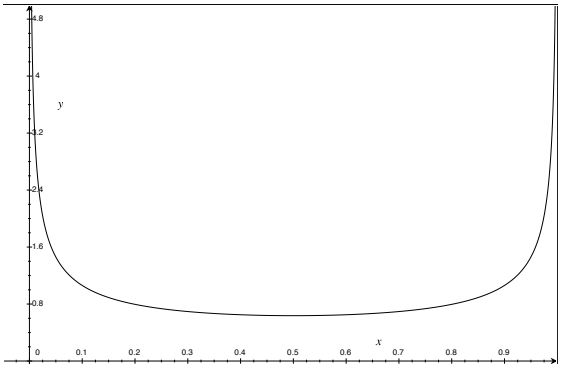

Figure 1: $\beta_{0.5,0.5}$

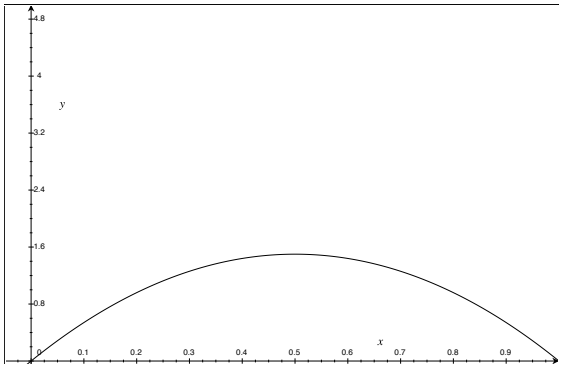

Figure 2: $\beta_{2,2}$

[22] There is a little bit of fudging here. If the average rate of success amongst those who have made m correct positive identifications out of n positive identifications under test conditions is $\frac{m+\lambda\mu}{n+\lambda}$, then the average rate of success amongst those who have made M correct positive identifications out of N positive identifications is $\frac{M+\lambda\mu}{N+\lambda}$, and so a should be taking *this* value and not the value M/N. What makes the fudging harmless is the assumption that N is large, for as N becomes large M/N and $\frac{M+\lambda\mu}{N+\lambda}$ approach one another because both tend to μ.

Figure 1 represents the distribution of a in a population where an individual's past success is strongly correlated with her performance at the next trial, here illustrated by $\mu = 0.5$ and $\lambda = 1$. Figure 2 represents the distribution of a in a population where this correlation is weak, here illustrated by $\mu = 0.5$, $\lambda = 4$.

What I want to take forward from this discussion of the distribution of a is therefore the following. First, because visual acuity as measured by a is a relatively stable skill, we should expect a correlation between the results of the sequence of trials by means of which we determine a. Second, the existence of such a correlation implies a degree of dispersal in the distribution of a across the population. It is to be expected that more people will score high values and more people will score low values than we should expect if an individual's track record and future performance were (statistically) completely independent.

2.4 Distribution of b

The fourth part of the model is the distribution of the second statistic, which I'll call b: the number of people who did not pay for entry as a proportion of the total number of people who attended. To be clear on what we are now asking about: for every rodeo or rodeo-like event, there is a certain proportion of people amongst those that attended who did not pay for a ticket: at one event, it may be 0 (if everybody pays for entrance on that day); at another, it may be 1 (if nobody pays for entrance); at a third, it may be 0.2 (if 80% pay); and so on. By 'the distribution of b' I mean the record of how frequently each such value turns up in a very long sequence of rodeos. For instance, if b has a roughly normal distribution with a mean of 0.5 and a very small variance, then this means that at almost all Sunday events, close to half of those that attended paid for their ticket. If b has a roughly exponential distribution, then there may be (e.g.) relatively many events that nobody entered without paying for entrance, half as many at which 10% did, a quarter as many at which 20% did, and so on.[23]

Now we know what is meant by the distribution of b, can we say anything about its shape? Again, some a priori constraints on it follow from the manner of its determination. The statistic b is not determined by testing, for each audience member one, over whether he or she has paid for entry. (If we could do that then there would be no need to rely on statistics in Bob's case, as we could simply apply the test to Bob directly!) Rather, we extract it from aggregate data concerning gate receipts and a head-count of spectators. We know from receipts that M people paid for entry; and we know from a head-count that N people were actually present at the Sunday event. So $b = M/N$.

Even though we do not have data on this, we can say something about our expectations concerning the correlation between the proportion of non-payers in the first n audience members and whether the $(n+1)$th audience member is a non-payer. Specifically, suppose that we order the members of the audience at each event in some arbitrary way, for instance by the order in which they passed through the gate.[24] The question is whether for each n we should expect a correlation between the proportion of non-payers amongst the first n audience members to have passed through the gate and whether the $(n+1)$th audience member is a non-payer. More explicitly: suppose we look at the data across a very large number of events.

[23] These would have to be *truncated* normal and exponential distributions, because the proportion of people that did not pay at a given rodeo-like event must lie between 0 and 1.

[24] That the method of ordering is irrelevant follows from the fact that if we treat the results on any Sunday as a sequence of random variables $\chi_1, \ldots \chi_N$, where $\chi_i = 1$ if on some specific ordering the ith audience member is a non-ticket-holder, and otherwise $\chi_i = 0$, then the $\chi_1, \ldots \chi_N$ are *exchangeable* i.e. for any $J \subseteq \{1,2 \ldots N\}$ and any $k \leq |J|$, the frequency with which $\sum_{j \in J} \chi_j = k$ is a function of $|J|$. That the $\chi_1, \ldots \chi_N$ are exchangeable is a weaker condition than that they are independent and identically distributed (i.i.d.), although as we'll see there is reason to expect something like independence in this case.

Do we expect that, amongst the events in which a *larger* proportion of the first n people to pass through the gate didn't pay, the $(n+1)$th person to pass through the gate is a non-payer more often than amongst the events at which a *smaller* proportion of the first n people to pass through the gate didn't pay?[25]

The answer, I claim, is *no*. Whether one person chooses to pay has no bearing on whether another person does. Given a particular base rate of such behavior x in the population in general, the existence of a proportion $y \gg x$ of non-ticket-holders in the first n audience members on some occasion does nothing to move the rate of non-ticket-holding by the $(n+1)$th audience member much higher than x; and if we found such a high rate in the first n people that we examined then we should be inclined to say it was bad luck, rather than that it was indicative of a strain of criminality amongst people attending on those days, people whose behavior cannot have had a causal influence on anyone else who attended that day. (At any rate, we can stipulate that this is how the story is supposed to go.)[26]

This relative weakness of correlation has consequences for the distribution of b itself, just as the corresponding stronger correlation had consequences for the distribution of a. Specifically, if we model the strength of correlation in accordance with (1), then the relevant value of λ must be high. Since it follows from (1) that b has a beta distribution with parameters $\lambda\mu$ and $\lambda(1-\mu)$, where μ is the overall mean rate of non-payment, the shape of b's distribution must look something like that in Figure 2: that is, tightly bunched around its mean.

Briefly to repeat the intuition behind this. We are supposing that the payment or non-payment behaviour of individuals is close to being independent, in the sense that a high rate of non-payment among the first n individuals arriving at the stadium does not correlate with a large rate of non-payment amongst later-arriving individuals. In this sense, testing individuals for non-payment is like repeatedly tossing a coin that is known to be fair: a long run of heads is no indication that the next toss will result in heads – it is just a run of good (or bad) luck.

The basic philosophical distinction that drives the difference in the shapes of the distributions of a and of b can be put like this. Whether a *single individual* is good at identifying witnesses is a relatively enduring trait that we can identify from testing that individual's time-slices for a property like *having made a correct positive identification*: doing well in the tests

[25] Note: I am *not* saying that this procedure is being carried out in the model as a means of ascertaining anybody's guilt; rather, I am eliciting your intuitions about what would happen if we did carry it out, as a means for motivating the claim that the proportion of non-paying spectators at a rodeo is a random variable that is distributed relatively tightly about this mean value.

[26] The situation would be different if mass non-payment generally took the form of a *cascade*, in which each participant's decision to enter without paying depended on the number of participants who had already entered without paying, in such a way that non-payment by a high proportion of her predecessors tended to encourage an individual to enter without paying. (For discussion a similar case, see Blome-Tillman 2015). In that case, a high rate of non-payment amongst the first n people to enter the stadium *would* correlate with non-payment by the $(n+1)$th entrant. But even then, it is implausible that the correlation would be anything like as strong as exists between earlier and later successes in trials of the sort that determine statistic a. The reason for this is that a great variety of facts about an individual's experience and psychology are relevant to the determination of her decision to pay entrance or not, almost all of which are causally independent of – and share no common cause with – what her predecessors did. By contrast, we should expect that a sequence of tests of an eyewitness's visual acuity will reflect in its outcomes the impact of all important factors that were relevant to the correctness of the identification of a suspect.

It is perhaps worth mentioning here that it may be this, and not any Kantian conception of freedom as autonomy, that mediates the relevance to these sorts of cases of the assumption that Alice and Bill acted *freely*. That assumption has been thought, on a roughly Kantian interpretation of freedom, to imply that we cannot use the incidence rate in a group as evidence against any of its members, although the connection between these things is somewhat unclear (Wasserman 1991). It may be that it is not any radical autonomy of her decision, but rather the fact, that an unsurveyable variety of facts about an individual's experience and psychology go into determining it, that bears on the propriety of using statistical evidence against her, ultimately for reasons that my main argument is going to outline.

suggests that the person is discerning, and so we can expect there to be a correlation between a high rate of possession of that property by earlier time-slices, and its possession by the next time-slice that makes a positive identification.

But when we examine *distinct individuals* in some class – such as: the class of people who entered the stadium on this or that day – for the property of not having paid for entry, we are *not* identifying a trait that can be ascribed to the whole group. It is not as though we think of individual guilt as manifesting some sort of moral or legal miasma that was associated with the event and which somehow caused audience members on that day to enter without paying for a ticket. There *are* historical instances of this view's having been widely accepted with regards to this or that type of wrong-doing; and if the argument that follows is correct, it may be no accident that they were also cases in which *collective* punishment for such types of wrong-doing seemed to some to be appropriate. Be that as it may, what matters from the present perspective is the connection between the different distributions of a and b in this model to the rate of accuracy of convictions on the two types of evidence that they represent, to which I now turn.

3 The demands of accuracy

Recall that the policy problem was to determine the optimal value for the proportion α of convictions on eyewitness as opposed to statistical evidence, where α lies between 0 and 1 inclusive. 'Optimal' means that we are maximizing accuracy i.e. the proportion of convictions that are *correct*. To this end, we need to calculate two quantities: the rate A of true convictions per conviction on eyewitness evidence, and the rate B of true convictions per conviction on statistical evidence. The proportion of convictions that are true convictions will then be given by $T(\alpha) = \alpha A + (1 - \alpha)B$: we must select an α that maximizes $T(\alpha)$.

3.1 Solving the policy problem

We calculate A as follows. The curve in Figure 3 is the same as that in Figure 1: it depicts the distribution of eyewitness reliability amongst cases in which the eyewitness picks out a suspect (for instance, Alice) as having gate-crashed a rodeo or similar.

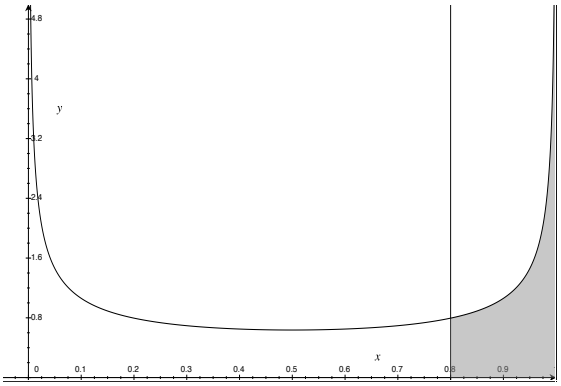

Figure 3: distribution of a

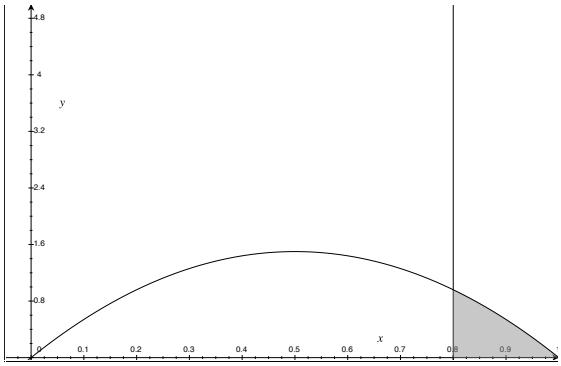

Figure 4: distribution of b

The region under the curve to the right of $x = 0.8$ is shaded: this represents the distribution of this accuracy amongst those cases in which we convict on this evidence, because these are exactly the cases in which the witness's accuracy exceeds the threshold i.e. $a > c = 0.8$. The average accuracy of all these convictions will be the mean value of the accuracy within the shaded region: this is the average rate A of true convictions given that the conviction is made on eyewitness evidence.

Intuitively we can see, just by looking at this curve, that $A > 0.9$. Amongst the cases in which we have an eyewitness whose accuracy exceeds 0.8, there are for any $\Delta \in (0,0.1]$, more cases in which her accuracy is $0.9 + \Delta$ than in which her accuracy is $0.9 - \Delta$. (It suffices for this, though it is not in general necessary, that the distribution function is *increasing* throughout $a \in (0.8,1)$.) More formally, the mean rate of true conviction is:

$$(3)\ A = E(a|a > 0.8) = \frac{\int_c^1 x\beta_{0.5,0.5}(x)dx}{\int_c^1 \beta_{0.5,0.5}(x)dx} = \frac{\int_{0.8}^1 x\beta_{0.5,0.5}(x)dx}{\int_{0.8}^1 \beta_{0.5,0.5}(x)dx} \approx 0.93$$

Given the distribution of eyewitness accuracy assumed in Figure 1 and Figure 3, approximately 93% of convictions on eyewitness testimony are true convictions.

Exactly parallel reasoning applies to statistical testimony based on gate receipts. See Figure 4, which stands to Figure 2 as Figure 3 does to Figure 1, and which we can assume to represent the distribution of the proportion of non-payers per stadium. Intuitively we can see, just by looking at *this* curve, that $0.9 > B$. For amongst the stadia in which the rate of non-payment exceeds 0.8, there are for any $\Delta \in (0,0.1]$, more cases in which this rate is $0.9 - \Delta$ than in which it is $0.9 + \Delta$. (It suffices for this, though again it's not necessary, that the distribution function is *decreasing* throughout $b \in (0.8,1)$.) More formally, the mean rate of true conviction is:

$$(4)\ B = E(b|b > 0.8) = \frac{\int_{0.8}^1 x\beta_{2,2}(x)dx}{\int_{0.8}^1 \beta_{2,2}(x)dx} \approx 0.87$$

Given the distribution of incidence rates per rodeo assumed in Figure 2 and Figure 4, approximately 87% of convictions on statistical evidence of the type facing Bob are true.

It follows straightforwardly that given a policy that selects a proportion α of convictions from cases where we have eyewitness testimony sufficient to convict, and the remaining $1 - \alpha$ from cases where we have statistical evidence sufficient to convict, the overall rate of true convictions per conviction is $T(\alpha) = 0.93\alpha + 0.87(1 - \alpha)$ i.e.:

$$(5)\ T(\alpha) = 0.87 + 0.06\alpha$$

Since the accuracy rate is strictly increasing in α it follows that the *accuracy*-maximizing policy is to convict *only* on eyewitness evidence; more generally it follows that accuracy-maximization alone is grounds for preferring eyewitness to statistical evidence, at least when the model presented here approximates closely enough to the truth.

We can now return to the widely-held view with which I began, namely that accuracy considerations alone *cannot* motivate different treatment of Alice's case and of Bob's case. What we have seen is that in the present model, accuracy considerations alone *can* motivate different treatment. More precisely: suppose that we are given a fixed threshold for conviction and a fixed number of convictions, and that we must choose how many convictions are based on eye-witness testimony and how many are based on statistical evidence. Then *accuracy considerations alone* can motivate preferring to convict based on eye-witness testimony in every case. According to this model, therefore, the target position is simply false. Concern for accuracy alone can motivate a policy, one effect of which is that we convict Alice but not Bob.

3.2 Permissible relaxations of the model

We derived this result from quite specific assumptions. It is worth asking to what extent we can relax the assumptions of the model whilst preserving the result. The answer is that the

model can be relaxed in three ways. These correspond to relaxations of some of the assumptions at 2.1 concerning the threshold, at 2.2 concerning the objective, and at 2.3 and 2.4 concerning the mean and the mathematical form of the distributions for a and b.

3.2.1 Variation in c

First: the superiority of eyewitness evidence holds at *any* plausible level at which we fix the threshold c: however strict or lenient we are about the standard of proof, convictions based on eyewitness evidence will convict fewer innocent ones than convictions based on aggregate incidence data. This follows from the fact that for any $c \geq 0.5$ we have:

$$(6)\ \frac{\int_c^1 x\beta_{0.5,0.5}(x)dx}{\int_c^1 \beta_{0.5,0.5}(x)dx} > \frac{\int_c^1 x\beta_{2,2}(x)dx}{\int_c^1 \beta_{2,2}(x)dx}$$

This inequality fails if c falls close enough to 0; but a threshold for conviction well *below* the base rate, in this case 0.5, seems unlikely.

On the other hand, it is consistent with my semi-Lockean assumptions that we operate with a *higher* threshold $c' > c$ when the evidence is statistical. If we do, then it is possible to ensure that the expected safety of convictions on statistical evidence matches or even exceeds the expected safety of convictions on eyewitness evidence, assuming that the latter type of conviction is still responsive to the lower threshold c. In the model, this follows from the fact that if $c < 1$ then there is some $c' \in (c, 1)$ such that $\frac{\int_{c'}^1 x\beta_{2,2}(x)dx}{\int_{c'}^1 \beta_{2,2}(x)dx} > \frac{\int_c^1 x\beta_{0.5,0.5}(x)dx}{\int_c^1 \beta_{0.5,0.5}(x)dx}$; that this result holds independent of the particular beta distributions in this model follows from the fact that for *any* two probability distribution functions f and g with support exclusively in $[0,1]$ and $c \in (0,1)$, there is some $c' \in (c, 1)$ such that $\frac{\int_{c'}^1 xg(x)dx}{\int_{c'}^1 g(x)dx} > \frac{\int_c^1 xf(x)dx}{\int_c^1 f(x)dx}$.[27] So on this variant of the model, dispensing with statistical evidence *does* incur a cost in accuracy.

This is true; and in fact our actual intuitions about this case would seem to fit a version of this variable-threshold policy. (After all, we are willing to find against Bob if it turns out that *everyone* on the relevant day entered without paying.) But conceding that does nothing to help Accuracy-Indifference, which said that a concern for accuracy cannot motivate a difference in our treatment *of cases A and B* i.e. of cases in which the evidence makes guilt *equally* likely. What we saw was that there can be a variable-threshold policy on which statistical evidence can produce convictions that are as accurate in expectation as those based on 'individualized evidence'. So given that we are adopting the variable-threshold policy, accuracy considerations can't motivate treating statistical evidence *on which we are prepared to convict* any differently from 'individualized' evidence *on which we are prepared to convict*. But it is also true, on the variable-threshold policy, that we are treating (a) cases where individualized evidence generates a probability of guilt x *differently* from (b) cases where statistical evidence generates the *same* probability x, for any x in the range $(c, c']$. So if accuracy considerations can motivate a variable-threshold policy then they can motivate an invidious attitude towards cases A and B, contrary to Accuracy-Indifference.

3.2.2 Minimizing false non-convictions when $K \gg D$

[27] Proof: choose $c' = E(f(x)|x > c) = \frac{\int_c^1 xf(x)dx}{\int_c^1 f(x)dx}$.

Second: at section 2.2 we made some assumptions about the form of the prosecutor's optimization problem took. More specifically, we supposed that the objective is simply to maximize the proportion of these convictions that are true convictions.

Suppose that instead we have two objectives: to maximize the proportion of convictions that are true convictions, *and* to minimize the proportion of non-convictions that are false *non*-convictions. That is, suppose that there are altogether $K \gg D$ cases in which we have either eye-witness or statistical evidence against a person that exceeds the threshold c. We are aiming to select α in such a way that, of the D convictions that we secure, a maximal proportion are true convictions, and in such a way that, of the $K - D$ non-convictions, a minimal proportion are false non-convictions: i.e., so that of those who are not convicted, the proportion that are guilty of gate-crashing is as low as it can be.

If we are given that $K \gg D$, we can assume that the rate of incidence amongst the $K - D$ *non-convicted* persons against whom we have evidence of either type that exceeds the threshold, is the same as the base-rate amongst *all* people against whom we have evidence of either type that exceeds the threshold: call this μ'. It follows that the rate of false non-convictions, for *any* value of α, is $1 - \mu'$. More precisely, *whatever* value we choose for α, we will secure $T(\alpha)$ true convictions per conviction and $(1 - \mu')$ true non-convictions per non-conviction. Therefore, choosing $\alpha = 1$ remains the optimal choice because it is 'Pareto optimal': it secures a higher rate of true *conviction* than at any alternative setting of α without incurring a higher rate of false *non*-conviction than any alternative setting of α.

This argument shows that if we relax the assumption that we only care about the rate of true convictions, then we can still derive the basic result. But it still relies on another assumption in 2.2, namely that there are many more cases in which evidence of either type exists than there are cases that we can feasibly prosecute (i.e. that $K \gg D$). The situation becomes considerably more complicated if we *also* relax that assumption: I'll discuss this in section 4.

3.2.3 Varying the mean of a and b

Third: in the present model, I have set the *means* of the distributions of a and of b at the same value i.e. 0.5. My reason for doing so was not empirical but rather to control for that parameter, thereby exposing the effect of variation in the other parameter λ. But that setting is most unrealistic. Any eyewitness who is not completely incompetent is almost certain to have a hit rate exceeding 0.5; and in any case, it is reasonable to suppose that the hit rate of the average eyewitness exceeds the frequency of gate-crashing in the average stadium. However, this idealization does not affect the main result, because it *understates* the true conviction-rate of eyewitness evidence: if $a \sim \beta_{\lambda\mu,\lambda(1-\mu)}$ and c and λ are held fixed, an increase in μ can only increase the value of $E(a|a > c)$.

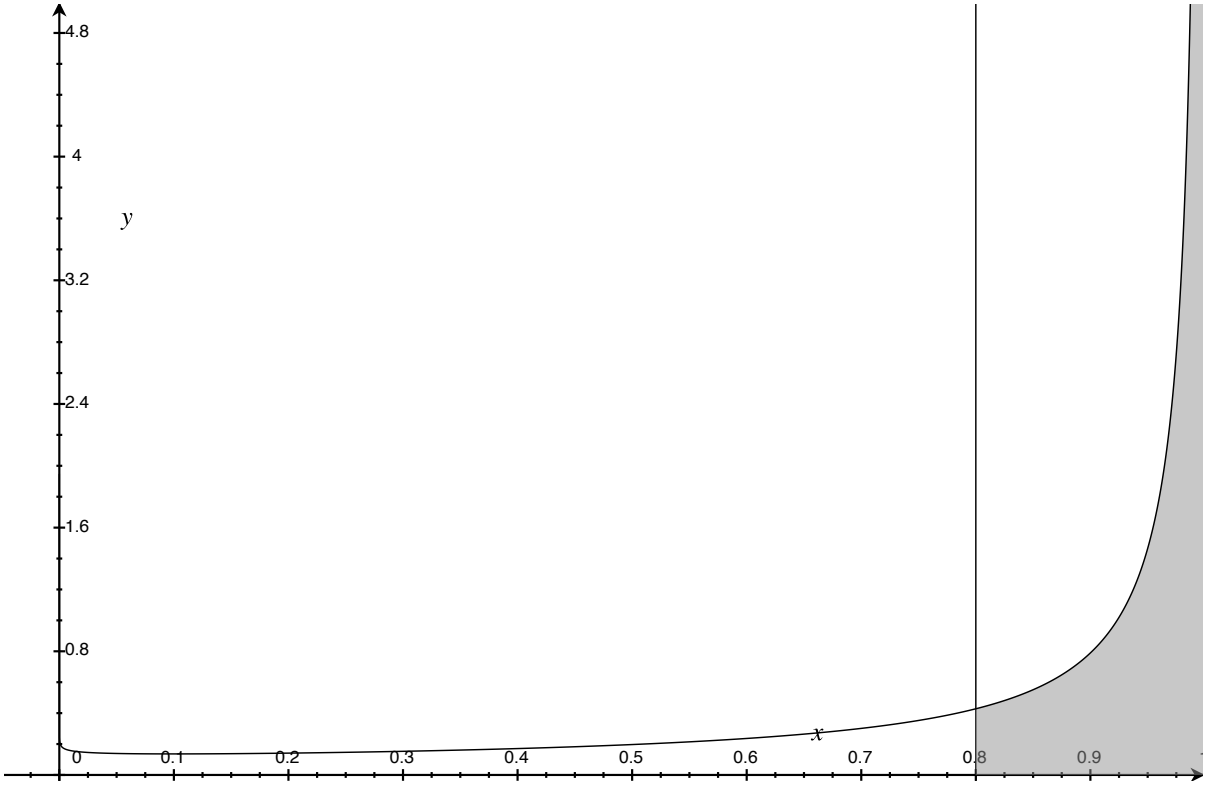

Figure 5

Graphically, imagine the curve in Figure 3 with some of the mass shifted to the right, for instance as in Figure 5, which represents the situation when $\mu = 0.9$ but still $\lambda = 1$. Clearly the expectation of this distribution to the right of $x = 0.8$ (or to the right of $x = c$, for any feasible c) is greater than that in Figure 3.

3.2.4 Varying the type of their distribution

Fourth: we can see from the graphical intuition behind the argument that it is possible to relax the assumption that a and b take the specified beta distributions. All that is required for the argument to go through is that: (i) the distribution of a is skewed to the right in the region $x \geq c$; and (ii) the distribution of b is skewed to the left in the region $x \geq c$. These features don't depend on a and b having the beta distributions that I chose for illustration. They don't even depend on a and b having beta distributions at all. What matters is just that their distributions have the right kind of shape, regardless of their precise algebraic form. It suffices for that, that there is a strong correlation between successive outcomes of the trials from which we generate a, but a weak correlation between successive outcomes of the (counterfactual) trials that determine b. And it suffices for *that* that visual acuity is a trait that we expect to remain stable across successively examined stages of a single person; whereas a tendency to enter a single rodeo without paying entrance is *not* a trait that we expect to remain stable across successively examined stages of *different* persons.

3.3 Generalizations and limitations

The argument in this paper targeted one specific pair of cases in which eye-witness evidence is contrasted with statistical evidence, namely the case involving Alice and Bob, in which the statistical evidence against Bob consists of gate receipts from attendance at a large event. I argued that if we model the optimization problem as in section 2, accuracy considerations alone can motivate a policy that convicts Alice but not Bob, even though the probability of Alice's guilt is the same as the probability of Bob's. Now as well as asking – as I just did – how we

can weaken those assumptions whilst still deriving that result, we can also ask how we can strengthen that result whilst maintaining those assumptions. Can we show that accuracy considerations alone may suffice to motivate preferring other kinds of 'individualized' evidence to other kinds of 'statistical' evidence?

The answer is that we can probably do so in some but certainly *not* in *all* cases where intuition distinguishes them. What is needed for the argument to go through is that there be variables a and b, corresponding to the posterior probability of guilt given an arbitrary piece of evidence of the individualized and of the statistical type respectively, such that the distribution of a is more heavily weighted towards the right than is the distribution of b, conditional on both exceeding the fixed threshold c above which we convict. More precisely, we need $E(a|a > c) > E(b|b > c)$. It suffices for this that there be a stronger correlation between guilt of all the first n people and guilt of the $(n + 1)$th person to face evidence of type a, than between guilt of the first n and guilt of the $(n + 1)$th person in any group from which the incidence rate b is drawn. And it suffices for *that*, that there be a common explanation, for instance a common causal explanation, of the correctness of what the a-type evidence suggests, but no single explanation, or at best a very incomplete single explanation, of the correctness of what the b-type evidence suggests.

Consider (for instance) extended vetting at airports. Suppose that we have the resources to carry out an extensive search for illegal drugs on some relatively small number D of travellers every day, and that the two available means of selection are either (a) the positive judgment of a security officer or (b) membership of some target group (for instance, the traveller's nationality as given by his or her passport). More specifically, suppose that we can select a person for extended searching either (a) on the word of a security officer, *if* testing has shown that that officer's positive judgments are right at least 1% of the time, or (b) because at least 1% of air passengers from the subject's country are carrying drugs. (These figures are likely to be low because the base rate of drug-smuggling amongst international travellers is presumably very much less than 1%.)

On these assumptions, it is very plausible that the mean accuracy of officers whose accuracy exceeds 1% is much greater than the mean smuggling rate in countries for which it exceeds 1%. The reason is that we do not expect a strong correlation between (say) a positive finding for the first n passengers from e.g. France and a positive finding for the $(n + 1)$th French passenger, given that these individuals are otherwise unrelated, whereas we *do* expect a relatively strong correlation between a positive finding for the first n passengers whom a given security officer suspects, and the finding for the next individual whom that officer suspects. Given these statistical assumptions, there is therefore a purely accuracy-based motivation for preferring to base extended vetting of this sort on the judgment of individual security officers rather than on any sort of ethnic, national or religious profiling.

But it is possible to change the statistical assumptions behind the example so that accuracy considerations point in the opposite direction. Suppose it is known that although relevant rate of smuggling amongst persons from all countries is, say, 0.01%, there are two countries for which the corresponding rate of smuggling is as high as 10%. On this assumption, it is *not* plausible that the mean accuracy of officers whose accuracy-of-suspicion exceeds 1% is much greater than the mean smuggling rate amongst travellers from countries in which it exceeds 1%. The demand for accuracy would therefore justify national profiling in this case. Perhaps in *this* case some people's intuitions go that way too. (My own intuitions about this type of case are too corrupted for any report of them to be useful.)

Still, there certainly are cases in which the accuracy-based argument defended here delivers a counter-intuitive result. Suppose there are very many squadrons of N soldiers in an occupied country. Soldiers will almost never fire on unarmed civilians unless all are explicitly ordered to do so by the commanding officer of the squadron; but the rate of compliance with

such orders is very high. In this case, we should certainly expect a strong correlation between the proportion of the first n soldiers in a squadron that fired on unarmed civilians on any occasion and whether the $(n+1)$th soldier did so on that occasion, because if the former quantity is high then we have a reliable indication that the commanding officer ordered all soldiers in that squadron to fire. In this situation, everything will depend on whether the rate of compliance r with an order of this kind exceeds the threshold for conviction c. If $r > c$ then the argument may recommend conviction of an arbitrary soldier in any group in which the proportion of soldiers that fired exceeds c, because in most such groups the rate of firing will be appreciably higher than c; it may even recommend preferring this sort of evidence to eyewitness evidence. This certainly looks counterintuitive.

But the aim of the argument was never to show that accuracy-based considerations can totally recover out intuitions about the differences between individualized and statistical evidence. It was rather to undermine the common view that accuracy considerations alone *cannot* motivate the distinction between statistical and individualized (for instance, eye-witness) evidence. To that end it is sufficient to show that they can do this in at least one case; and I think the foregoing argument does establish as much for the gate-crashing case with which I began, given the Lockean model. I have not shown, and I don't think is true, that accuracy concerns *never* recommend convicting when intuition acquits.

4 Two types of accuracy

The model assumes that the policy problem arises against a background of a high rate of criminality, certainly far more than can be addressed by the penal system.[28] The decision is therefore not over *whether* to convict more people or fewer, but rather over *which* ones to convict, given that one is convicting as many people as one can. Formally this is reflected in the framing of the problem as that of maximizing $T(\alpha) = \alpha A + (1-\alpha)B$, where α is the proportion of convictions on eye-witness evidence.

But it is possible to think of a different policy problem, namely that of deciding *how many* people within a specified class to convict. Suppose that we have the resources to convict *everyone* against whom we have evidence that suffices to convict, for a given threshold (so that $K \approx D$). And suppose that these people fall into two classes, which may be of different sizes: those against whom we have eye-witness evidence from an eyewitness whose accuracy exceeds c, and those against whom we have statistical evidence that implies a probability of guilt exceeding c. For simplicity, we can frame the resultant policy problem as a choice between three options:

(a) Only convict everyone against whom we have eye-witness evidence.
(b) Only convict everyone against whom we have statistical evidence.
(c) Convict everyone against whom we have *either* type of evidence.

Now the target claim – the one that I am saying is false – is that accuracy considerations alone cannot motivate preferring eye-witness or other types of 'individualized' evidence to statistical evidence. Even if we grant that that claim *is* false on the assumptions of the original policy problem (where $K \gg D$), we might still suspect that it is true against the present background of a three-way choice (where $K \approx D$). More specifically, we might suspect that

[28] This is not in itself unreasonable for many crimes. John Adams, speaking in 1770, said that 'guilt and crimes are so frequent in the world, that all of them cannot be punished; and many times they happen in such a manner, that it is not of much consequence to the public, whether they are punished or not'; and the first part of this surely remains true today.

accuracy considerations cannot motivate a policy that treats these types of evidence asymmetrically: they cannot account for why (a) should be preferable to both (b) *and* (c).

An uncharitable response would be as follows. Suppose that we are concerned with accuracy *only* to the extent that we are trying to maximize the proportion of convictions that are true convictions. And suppose that we have xK cases in which the eye-witness evidence exceeds the threshold for conviction, and $(1-x)K$ cases in which the statistical evidence exceeds the threshold for conviction, for some x between 0 and 1. Then if we write a and b for random variables stating the accuracy of a randomly chosen item of eye-witness evidence or statistical evidence respectively, the rate of true convictions per conviction under policy (a) is still $E(a|a > c)$, that under policy (b) is still $E(b|b > c)$, and that under policy (c) is:

(1) $xE(a|a > c) + (1-x)E(b|b > c)$

And given what I have already defended at length, namely that $E(a|a > c) > E(b|b > c)$, it follows that the rate of true conviction under policy (a) still exceeds that under either policy (b) *or* policy (c).

What makes this response uncharitable is that a concern 'for accuracy' might involve concern not only with the rate of true conviction, but also with the rate of true *non*-conviction. We do not only want to convict only the guilty; we should also like to acquit (or not to convict) only the innocent. As we saw at 3.2.2, if $K \gg D$ then our having this additional aim makes no difference to the analysis: if many more cases exist than can be prosecuted, any distribution of the D available prosecutions between those in which we have eye-witness and those in which we have statistical evidence will result in the same rate of true non-convictions. (Compare: given a fixed number of Americans, and a fixed number of Americans with brown eyes, the proportion of US Senators who have brown eyes makes almost no difference at all to the proportion of all *other* Americans that have brown eyes.) In the original policy problem, *any* choice of α leaves the rate of true non-convictions completely unaffected. Hence the only accuracy-based grounds on which to prefer one choice of α to another is an improvement in the rate of true convictions.

On the other hand, if $K \approx D$ then the choice between (a), (b) and (c) certainly *does* make a difference to the rate of true non-convictions. (Compare: given a fixed population of 100 US Senators of which a fixed number have brown eyes, the proportion of Democratic Senators who have brown eyes could make a relatively big difference to the proportion of Republican (or non-aligned) Senators who do.) We must therefore ask whether under these conditions, anyone whose concern for accuracy takes this *two*-dimensional form can motivate preferring eye-witness to statistical evidence: specifically, preferring (a) to (b) *or* (c) on grounds of accuracy alone.

The analysis in this section aims to show that such a motivation is available. Doing so involves, first, saying a little more accurately what it means to want to maximize, and to be forced into trading off, two competing types of good (in this case, true convictions and true non-convictions); section 1 does so by means of some basic concepts of microeconomics. The next steps are to show that in this framework we can motivate on accuracy grounds a preference for (a) over (b); and that we can – but only in a slightly weaker sense of 'can' – motivate on such grounds a preference for (a) over (c).

4.1 Trade-off between true conviction and true non-conviction rate

In the model that I now want to develop, there are *two* accuracy-related goods: the rate of true convictions and the rate of true non-convictions, and different measures of accuracy trade these off at different rates.

This situation is analogous to a more familiar one that arises with regards to belief. To say that belief aims at truth is really to say that in the formation of belief we have *two* aims: the maximization of true belief and the minimization of false belief. But as James points out in a famous passage, having those two aims, and only those two aims, is consistent with any degree of emphasis on one over the other.

> Believe truth! Shun error! – these, we see, are two materially different laws; and by choosing between them we may end by coloring differently our whole intellectual life. We may regard the chase for truth as paramount, and the avoidance of error as secondary; or we may, on the other hand, treat the avoidance of error as more imperative, and let truth take its chance.[29]

For instance, W. K. Clifford treats the avoidance of false belief as always taking lexical priority over the attainment of true belief. 'Believe nothing, [Clifford] tells us, keep your mind in suspense forever, rather than by closing it on insufficient evidence incur the awful risk of believing lies.'

> You, on the other hand, may think that the risk of being in error is a very small matter when compared with the blessings of real knowledge, and be ready to be duped many times in your investigation rather than postpone indefinitely the chance of guessing true…[30]

Or you might take some approach in between these two. For instance, you might be prepared to adopt any method that gains you one additional true belief for every two or fewer false beliefs. In any case and as James concludes, *this* decision is not dictated by the aim of maximizing accuracy but is rather something for taste, or 'passion' to settle, any such settlement being consistent with that aim. 'We must remember that these feelings of our duty about either truth or error are in any case only expressions of our passional life.'

In the present case we are concerned, not with the rate at which any individual is willing to trade off belief in true propositions against non-belief in false ones, but rather with the rate at which a jurisdiction trades off true convictions against true non-convictions amongst the K individuals that we could potentially convict, because our evidence against them exceeds the Lockean threshold. But the parallel point remains: an overriding concern with accuracy is consistent with any of very many ways in which we can rank these trade-offs.

We can characterize these ways by means of what economists call *indifference curves*. An indifference curve, in this context, is generated by an *objective function* $F: [0,1]^2 \rightarrow \mathbb{R}$. The two inputs to an indifference function, x and y, will represent respectively a *true conviction rate* and a *true non-conviction rate*. The true non-conviction rate is the proportion of non-convictions that are of innocent people, and the true conviction rate is the proportion of convictions that are of guilty people. The objective function is what the jurisdiction is trying to maximize. For each constant $k \in \mathbb{R}$, $F(x,y) = k$ is an associated indifference curve that represents a set of combinations of true conviction rates and true non-conviction rates between which the jurisdiction is indifferent. Intuitively, we can think of indifference curves as contour lines on a map in which altitude corresponds to overall utility. If any two points (x,y) and (x',y') lie on the same indifference curve then they are at the same 'altitude' i.e. we are indifferent between (i) x true convictions per conviction and y true non-convictions per non-

[29] James 2000 [1896]: 209.
[30] James 2000 [1896]: 209; cf. Clifford 1999 [1887].

conviction, and (ii) x' true convictions per conviction and y' true non-convictions per non-conviction.

For instance, suppose the objective function is $F(x,y) = 2x + y$, so that the indifference curves take the form $2x + y = k$ This means that we are, at any point, just willing to accept a reduction of Δx in the true conviction rate if doing so can get us an increase of at least $\Delta y = 2\Delta x$ in the proportion of innocent people who are acquitted. More generally, if at any point (x_0, y_0) the indifference curve passing through that point is $F(x,y) = k$, then we are willing to trade correct convictions for correct acquittals at the same rate as (the sign-reversal of) the slope of $F(x,y) = k$ at that point: that is, at the rate of

$$-\frac{dy}{dx} = \left[\frac{\partial F/\partial x}{\partial F/\partial y}\right]_{x_0,y_0}$$

correct acquittals per correct conviction. This is the *Marginal Rate of Substitution* (MRS) of true acquittals and true convictions at that point.

Indifference curves have two qualitative properties of interest. First: like contour lines, no two indifference curves ever cross. Conceptually, the reason is that indifference curves are effectively the equivalence classes associated with the equivalence relation of indifference on the unit square; and it follows from the definition of an equivalence class that any two such classes are disjoint. Second, indifference curves generally slope downwards: more precisely, if $x' > x$ and $y' > y$ then there is no indifference curve on which (x,y) and (x',y') both lie. This reflects the fact that any form of concern for accuracy must prefer a 'Pareto improvement' in the rates of true conviction and non-conviction. If policy S' generates higher rates of true conviction *and* of true non-convictions than policy S, then S' *must* be preferred to S.

But these constraints leave open a very wide variety of patterns of preference over combinations of rates of true convictions and rates of true acquittal. Figures 6-9 represent four such combinations. (In case they are hard to see, the curves in Figure 6 are horizontal lines, and those in Figure 7 are vertical lines.) In all cases, the best possible situation is in the north-east corner, at $(1,1)$. That point corresponds to the case where all non-convictions are of innocent and all convictions of guilty suspects. The worst possible situation is the south-west corner $(0,0)$, corresponding to a regime that convicts only the innocent and acquits only the guilty. And travelling north-east from *any* point gets us to a better point, one that is higher up the mountain of utility. But the rate, at which movement in these or any other directions of travel increases altitude, depends on the location of the contour lines, which are very different in all four cases.

Let us consider these examples in a little more detail.

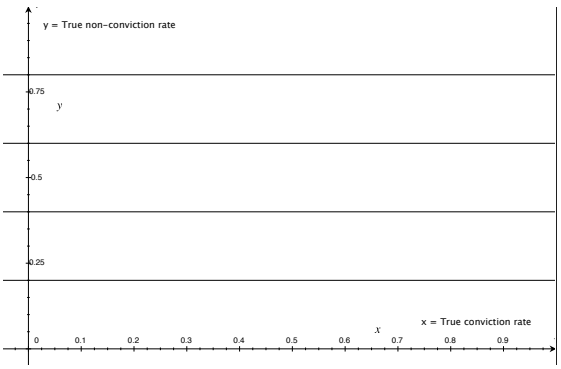

Figure 6: quasi-Jamesian preferences

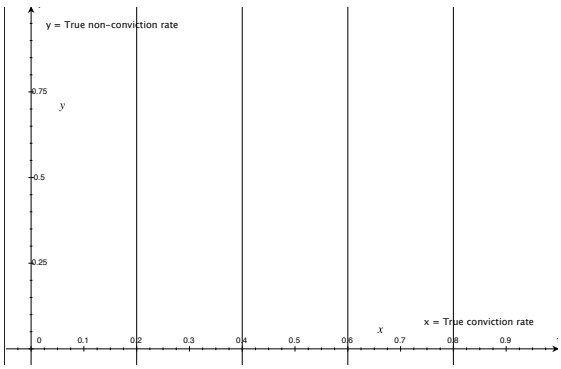

Figure 7: Clifford-type preferences

In Figure 6, all we care about is the rate of true non-conviction: in other words, it is paramount that we convict the guilty. If belief corresponds to conviction and non-belief to acquittal, then this corresponds to James's suggestion that you may 'be ready to be duped many times in your investigation rather than postpone indefinitely the chance of guessing true' – hence 'quasi-Jamesian'. So, the indifference curves are horizontal. At any given rate of true non-convictions, we don't care about the rate of true convictions: we can climb this utility mountain by, and only by, increasing the proportion of innocent people amongst those that we do not convict, for instance by convicting everyone against whom we have the slightest evidence.

In Figure 7, all we care about is the rate of true conviction – we are willing to let indefinitely many guilty suspects go free rather than convict one more innocent one, for instance by never convicting unless the evidence is overwhelming. This corresponds to the imperative that James attributes to Clifford: 'Believe nothing... keep your mind in suspense forever, rather than by closing it on insufficient evidence incur the awful risk of believing lies.' Hence, the indifference curves are all horizontal.

Here are two more realistic examples.

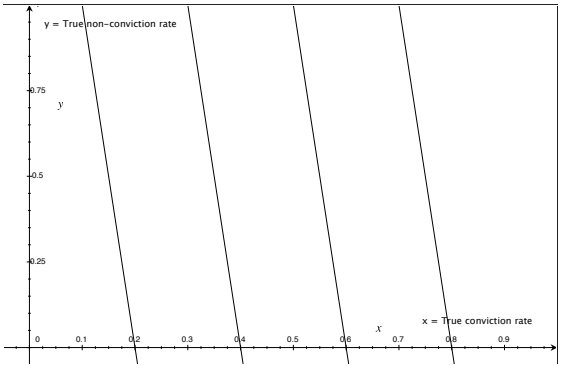

Figure 8: 'Blackstone' preferences

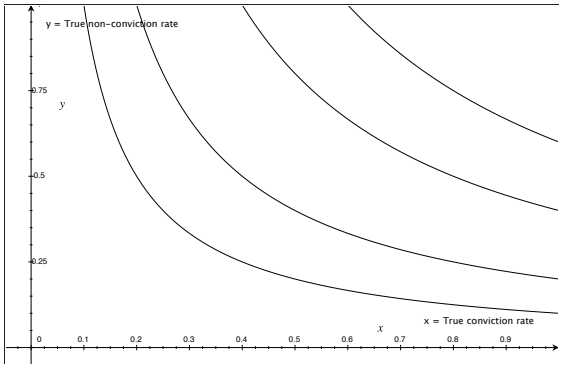

Figure 9: Declining marginal utility

In Figure 8 the slope of the indifference curves is -10. On the (implausible) assumption that about as many people are convicted as not, this represents the minimum degree of caution mandated by the 'Blackstone formula': 'Better that ten guilty persons escape than that one innocent suffer'. If we are convicting 100 people, out of which 20 are guilty, and acquitting 100 people, out of which 10 are guilty, then we are at the point $(0.2,0.9)$; and we are indifferent between this situation and that in which we are convicting 100 people, of whom 19 are guilty, and acquitting 100 people, of whom *none* are guilty, which corresponds to the point $(0.19,1)$.

Finally, in Figure 9 there is no one rate of trade-off common to all points. Rather, the figure depicts a jurisdiction that is (a) willing to see more guilty people go free when the regime is unduly harsh, and (b) willing to see more innocent people convicted when the regime is unduly lenient. An extreme instance of (a) would be a situation where the offence is rare, and yet the courts simply convict everyone that the police suspect i.e. in which the rate of true conviction stands at the base rate B (which is close to 0) and the rate of true non-conviction at 1. Relative to that point, it might seem an improvement to adopt a more lenient attitude that generates a true conviction rate of $B + 0.1$ and a true non-conviction rate of, say, 0.8. An extreme instance of (b) would be a situation where only absolute certainty was sufficient for conviction for an offence that is in fact prevalent throughout society. In this case the rate of true convictions would be close to 1 and that of true non-convictions close to $1 - B$, where B is the base rate of the offence (which is close to 1). From this starting point, it might be considered an improvement for the courts to start convicting on relatively weak evidence, leading e.g. to a fall in the rate of true convictions to 0.8 and a rise in the rate of true non-

convictions to $1.1 - B$. More generally the attitude, of caring more about increasing the rate of true convictions (true acquittals) when that rate is lower, implies indifference curves that are bowed towards the origin – as e.g. in Figure 9, where $F(x, y) = xy$.

For present purposes, what matters more than the specific features of the sets of indifference curves depicted in Figures 6-9, is that they are all consistent with an exclusive concern for accuracy. Having as our sole motivation the maximization of accuracy in findings does not yet tell us what to maximize, since accuracy itself encompasses two goods (corresponding to James's goods of believing the true and not believing the false). Just as an exclusive preoccupation with the accumulation of precious metals is consistent with a preference for gold over platinum, but also with the reverse, so too an exclusive preoccupation with accuracy is consistent with any set of indifference curves lying 'between' the James-type and the Clifford-type as described here.

4.2 (a) vs (b)

Suppose that of the N suspects against whom we have evidence of either type and at any level of probability, a proportion α face eyewitness evidence and a proportion $(1 - \alpha)$ face evidence of the statistical type. And suppose in line with our earlier model that the distributions of the accuracies a and b for these two types of evidence are beta, with a common mean and with the reliability of the eye-witness evidence more 'dispersed' than that of the statistical evidence. That is, we have: $a \sim \beta_{\lambda\mu,\lambda(1-\mu)}$ and $b \sim \beta_{\lambda'\mu,\lambda'(1-\mu)}$, with $\lambda' > \lambda$. As before, we can suppose for the sake of illustration that $\lambda = 1$, $\lambda' = 4$ and $\mu = 0.5$.

How does policy (a) fare on the two measures of accuracy that now concern us? We already know that amongst the suspects who are convicted, the proportion of convictions that are true convictions is given by:

$$(2)\ x_a = E(a|a > c) = \frac{\int_c^1 x\beta_{\lambda\mu,\lambda(1-\mu)}(x)dx}{\int_c^1 \beta_{\lambda\mu,\lambda(1-\mu)}(x)dx}$$

What about the proportion of non-convictions that are true non-convictions? Well, the people that we *don't* convict under policy (a) fall into two classes:

- Class 1: *all* those against whom we have statistical evidence;
- Class 2: all those against whom we have eye-witness evidence at a level that does not exceed the threshold.

To work out the rate of true non-conviction under policy (a), we need to calculate, for $i = 1,2$, the size s_i of class i and the rate of offending within it. We can then calculate the total number of non-convictees and the total number of offenders amongst them.

The size of Class 1 is clearly $s_1 = N(1 - \alpha)$ and the rate r_1 of offending within it is just the mean of the distribution for b, that is, $r_1 = \mu$. The size of Class 2 is given by:

$$(3)\ s_2 = \alpha N \int_0^c \beta_{\lambda\mu,\lambda(1-\mu)}(x)dx$$

And the rate of offending in class 2 is just the mean accuracy of eyewitnesses whose accuracy does not exceed the threshold, that is:

$$(4)\ r_2 = \frac{\int_0^c x\beta_{\lambda\mu,\lambda(1-\mu)}(x)dx}{\int_0^c \beta_{\lambda\mu,\lambda(1-\mu)}(x)dx}$$

The total rate of offence amongst those that policy (a) does not convict is therefore:

$$(5)\ \frac{\sum_{i=1}^{2} r_i s_i}{\sum_{i=1}^{2} s_i} = \frac{(1-\alpha)\mu+\alpha\int_0^c x\beta_{\lambda\mu,\lambda(1-\mu)}(x)dx}{(1-\alpha)+\alpha\int_0^c \beta_{\lambda\mu,\lambda(1-\mu)}(x)dx}$$

So the rate of true non-convictions per non-conviction achieved by policy (a) is given by:

$$(6)\ y_a = \frac{(1-\alpha)(1-\mu)+\alpha\int_0^c (1-x)\beta_{\lambda\mu,\lambda(1-\mu)}(x)dx}{(1-\alpha)+\alpha\int_0^c \beta_{\lambda\mu,\lambda(1-\mu)}(x)dx}$$

By exactly parallel reasoning, the rate of true convictions amongst those that policy (b) convicts is:

$$(7)\ x_b = E(b|b>c) = \frac{\int_c^1 x\beta_{\lambda'\mu,\lambda'(1-\mu)}(x)dx}{\int_c^1 \beta_{\lambda'\mu,\lambda'(1-\mu)}(x)dx}$$

And the rate of true non-convictions on this policy is given by:

$$(8)\ y_b = \frac{\alpha(1-\mu)+(1-\alpha)\int_0^c (1-x)\beta_{\lambda'\mu,\lambda'(1-\mu)}(x)dx}{\alpha+(1-\alpha)\int_0^c \beta_{\lambda'\mu,\lambda'(1-\mu)}(x)dx}$$

We already know that for the assumed values of $c, \lambda, \lambda', \mu$ we have $x_a \approx 0.93$ and $x_b \approx 0.87$, so that for any value of α policy (a) has a better true *conviction* rate than does policy (b). But things are different when it comes to true *non*-conviction rates. By inspection of (12) and (14) and a little algebra, when α is close to its extreme values the approximate values of those rates are as follows:

	$\boldsymbol{\alpha \approx 0}$	$\boldsymbol{\alpha \approx 1}$
$\boldsymbol{y_a}$	$1-\mu$	$1-E(a\|a\leq c)$
$\boldsymbol{y_b}$	$1-E(b\|b\leq c)$	$1-\mu$

Table 1

We can see intuitively from Table 1 and Figures 3 and 4 that if α is close to 1 then $y_a > y_b$; but if α is close to zero then $y_b > y_a$.[31]

It follows that if α is close to 1 then policy (a) has a better true *non*-conviction rate than (b), as well as a better *true* conviction rate. In those circumstances, policy (a) is a Pareto improvement on policy (b): no matter the rate at which we are prepared to trade off true convictions for true non-convictions, we should always prefer a policy of only convicting on eye-witness evidence to a policy of only convicting on statistical evidence.

But if α is close to zero then things are not so clear. In that situation, although it is still true that policy (a) has a better *true* conviction rate than policy (b), policy (b) has a better true non-conviction rate. That makes intuitive sense: if almost none of your evidence is eye-witness evidence then a policy of convicting only on eye-witness evidence is going to let many more

[31] Here is a formal argument for the case that $\alpha \approx 0$. Note that $\mu = kE(b|b\leq c) + (1-k)E(b|b>c)$, where $k = \int_0^c \beta_{\lambda'\mu,\lambda'(1-\mu)}(x)dx$. Since $0<k<1$ and $E(b|b\leq c) < E(b|b>c)$, it follows that $E(b|b\leq c) < \mu$. Therefore $1-\mu < 1-E(b|b\leq c)$.

genuine offenders escape than a policy of convicting on other, more readily available sorts of evidence. So in that case, neither policy is a Pareto improvement on the other.

But even in the case that α is close to zero (though still strictly positive), it remains *possible* for accuracy considerations to motivate a preference for (a) over (b). It all depends on the rate at which one is willing to trade off these things. Suppose that one takes the 'Blackstone' attitude that it is very much better to let 10 (or 100, or 1000) guilty men go free than falsely to convict one. That is, suppose that the *way* in which one cares about accuracy rates the true conviction rate as being very much more important than the true non-conviction rate at almost any level that these two rates might take. Then one might well care more about the fact that policy (a) has a better true conviction rate than about the fact that policy (b) has a better false conviction rate.

The kind of concern for accuracy that might bring this situation about is as illustrated in Table 2 and Figures 10 and 11. Table 2 specifies the numerical values of the true conviction and non-conviction rates on the two extreme assumptions for the values of α, calculated on the usual assumptions that $c = 0.8, \lambda = 1, \lambda' = 4, \mu = 0.5$.

	$\alpha \approx 0$	$\alpha \approx 1$
x_a	0.93	0.93
y_a	0.5	0.68
x_b	0.87	0.87
y_b	0.54	0.5

Table 2

Figure 10 plots the points (x_a, y_a) and (x_b, y_b) on the assumption $\alpha \approx 0$ against some notional indifference curves that would express a relatively high concern for the rate of true conviction relative to the rate of true non-conviction; in this case, I have chosen members of the family $x^{0.9}y^{0.1} = k$ for varying k. Figure 11 plots the points (x_a, y_a) and (x_b, y_b) on the assumption $\alpha \approx 1$ against the same indifference curves.

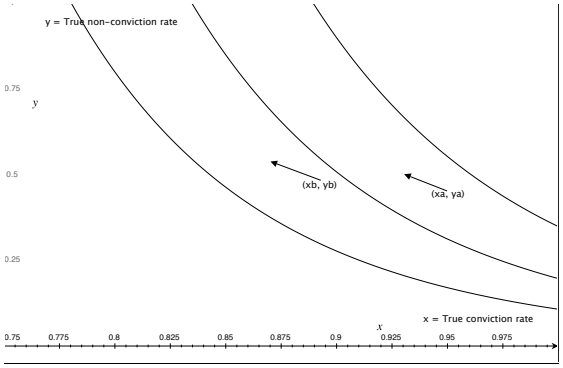

Figure 10: $\alpha \approx 0$

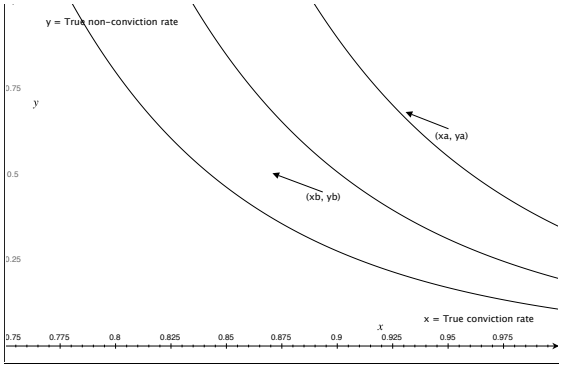

Figure 11: $\alpha \approx 1$

It is clear from these figures that in *both* cases the point (x_a, y_a) is preferred to – because it lies 'higher up the utility mountain than' – the point (x_b, y_b).[32] That is, even if practically all of one's evidence is statistical, anyone with this kind of concern for accuracy would *still* prefer the combination of true conviction rates and true non-conviction rates arising from a policy that convicts only on eyewitness evidence, to the combination of those rates arising from a policy that convicts only on statistical evidence. In consequence of this, it can also be shown

[32] This would be true in Figure 11 for any background indifference curves, because in that case (x_a, y_a) lies to the north-east of (x_b, y_b) i.e. the former is a 'Pareto improvement' on the latter.

that anyone with this kind of concern for accuracy would retain these preferences for *any* value of α.[33]

We may conclude that on this new model, whilst it is true that a concern for accuracy may not *mandate* a preference for only-eyewitness over only-statistical evidence, it certainly *can* motivate such a preference. More precisely: the indifference curves in Figures 10 and 11 illustrate a way of caring about accuracy, and *only* about accuracy, that would rationalize just such an invidious attitude towards these two kinds of evidence.

4.3 (a) vs (c)

Similar points can be made, this time more briefly, for the comparison of policy (a) with policy (c), on which we convict people on either kind of evidence. From the perspective of accuracy, what matters is (i) the rate of true convictions and (ii) the rate of true non-convictions under the two policies.

We already know what these rates are for policy (a), for a given proportion α of cases in which we have eye-witness evidence. These are the quantities x_a and y_a as described in equations (8) and (12) for arbitrary α and as specified in Tables 1 and 2 for extreme values of that parameter. What about policy (c)?

The rate of true conviction under policy (c) is given by the rate of offence amongst all those against whom evidence of either sort supports a probability of guilt exceeding c. This quantity is the weighted sum of the rates of offence amongst those who face eye-witness and those who face statistical evidence, where the weights are given by the proportion of suspects that face evidence of either type. So we need to look at the size, and the rate of offending, in two classes:

- Class 1: those who face eye-witness evidence that implies a probability of guilt that exceeds c;
- Class 2: those who face statistical evidence that implies a probability of guilt exceeding c.

The size and rate of offence in Class 1 is given by:

$$(9)\quad s_1 = N\alpha \int_c^1 \beta_{\lambda\mu,\lambda(1-\mu)}(x)dx$$

$$(10)\quad r_1 = \frac{\int_1^c x\beta_{\lambda\mu,\lambda(1-\mu)}(x)dx}{\int_1^c \beta_{\lambda\mu,\lambda(1-\mu)}(x)dx} = x_a$$

Similarly, the size and rate of offence in Class 2 are given by:

$$(11)\quad s_2 = N(1-\alpha) \int_c^1 \beta_{\lambda'\mu,\lambda'(1-\mu)}(x)dx$$

$$(12)\quad r_2 = \frac{\int_c^1 x\beta_{\lambda'\mu,\lambda'(1-\mu)}(x)dx}{\int_c^1 \beta_{\lambda'\mu,\lambda'(1-\mu)}(x)dx} = x_b$$

It follows that the rate of true convictions under policy (c) is:

[33] To show this, it suffices to show that y_a is increasing in α and y_b is decreasing in α. Differentiation of (12) implies that $\partial y_a/\partial\alpha = \int_0^c(\mu - x)\beta_{\lambda\mu,\lambda(1-\mu)}(x)dx/v^2$, where $v = \alpha + (1-\alpha)\int_0^c \beta_{\lambda'\mu,\lambda'(1-\mu)}(x)dx$. The denominator of this fraction is always positive if $0 < \alpha < 1$; the numerator is positive if and only if $\mu > \int_0^c x\beta_{\lambda\mu,\lambda(1-\mu)}(x)dx/\int_0^c \beta_{\lambda\mu,\lambda(1-\mu)}(x)dx = E(a|a<c)$. Since $\mu = E(a)$ this condition clearly holds, so y_a is increasing. Similar reasoning shows that y_b is *decreasing* if and only if $\mu > \int_0^c x\beta_{\lambda'\mu,\lambda'(1-\mu)}(x)dx/\int_0^c \beta_{\lambda'\mu,\lambda'(1-\mu)}(x)dx = E(b|b<c)$, and again this is obviously true.

(13) $x_c = \frac{s_1 r_1 + s_2 r_2}{s_1 + s_2} = kx_a + (1-k)x_b$, where

(14) $k = \frac{\alpha \int_c^1 \beta_{\lambda\mu,\lambda(1-\mu)}(x)dx}{\alpha \int_c^1 \beta_{\lambda\mu,\lambda(1-\mu)}(x)dx + (1-\alpha)\int_c^1 \beta_{\lambda'\mu,\lambda'(1-\mu)}(x)dx}$.

Since $0 < \alpha < 1 \rightarrow 0 < k < 1$, it follows from (19), (20) and $x_a > x_b$ that we *always* have $x_a > x_c$. This is intuitively plausible: since relying on eye-witness evidence creates a better true conviction rate than relying on statistical evidence alone, one would expect that a policy of convicting given either type of evidence will dilute the true conviction rate relative to the policy of convicting only on eye-witness evidence. More precisely, the true conviction rate under policy (c) must lie somewhere between the higher true conviction rate on policy (a) and the lower true conviction rate under policy (b).

What about the rate of true *non*-convictions under policy (c)? By parallel reasoning to that regarding y_a and y_b, we can calculate this quantity as:

(15) $y_c = \frac{\alpha \int_0^c (1-x)\beta_{\lambda\mu,\lambda(1-\mu)}(x)dx + (1-\alpha)\int_0^c (1-x)\beta_{\lambda'\mu,\lambda'(1-\mu)}(x)dx}{\alpha \int_0^c \beta_{\lambda\mu,\lambda(1-\mu)}(x)dx + (1-\alpha)\int_0^c \beta_{\lambda'\mu,\lambda'(1-\mu)}(x)dx}$

Putting all of this together, and simplifying (21), we can tabulate our results for extreme values of α as follows:

	$\boldsymbol{\alpha \approx 0}$	$\boldsymbol{\alpha \approx 1}$
$\boldsymbol{x_a}$	$E(a\|a > c)$	$E(a\|a > c)$
$\boldsymbol{y_a}$	$1 - \mu$	$1 - E(a\|a \leq c)$
$\boldsymbol{x_c}$	$x_b + \Delta_1$	$x_a - \Delta_2$
$\boldsymbol{y_c}$	$1 - E(b\|b \leq c)$	$1 - E(a\|a \leq c)$

Table 3

In Table 3, Δ_1 and Δ_2 are both small positive quantities. That both are positive follows from the fact that (19) makes x_c increasing in α; moreover, the latter entails that $x_a > y_a$ for any value of α.[34]

It follows from Table 3 that even if $y_c > y_a$ – which certainly will happen if α is close to zero – it is still possible for there to be an accuracy-based motivation for preferring policy (a) to policy (c). In English: even if convicting on both types of evidence (above the threshold) generates a smaller rate of true non-conviction than convicting only on eye-witness testimony, which is what we would expect if most available evidence is statistical, one *can* still justify, *on grounds of accuracy alone*, an exclusive focus on the latter type of evidence. The argument is the same as at section 4.2: if the *way* in which one cares about accuracy rates the true conviction rate as much more important than the true non-conviction rate at almost any level for either, then one might well care more about the fact that policy (a) has a better true conviction rate than about the fact that policy (c) has a better true non-conviction rate, and so not be willing to trade off the former for the latter given that there is *some* eye-witness evidence. Indifference curves of the type illustrate in Figures 10 and 11 represent just such a way. I conclude that accuracy by itself does not mandate a preference for policy (c) over policy (b).

[34] Since $x_a > x_b$, it is obvious from (19) that x_c is increasing in k, and from (20) that k is increasing in α. So x_c is increasing in α.

5. Conclusion

Our initial cases A and B seemed alike in respect of the accuracy of the evidence presented in each case; that is because those particular cases *are* alike in that respect. I have argued, contrary to many legal and philosophical commentators – perhaps also contrary to intuition – that it does not follow that accuracy considerations alone are powerless to motivate a distinction between the types of evidence that these cases involve. In the model of section 2 accuracy considerations will, and in the model of section 4 they can, motivate by themselves an invidious attitude towards those types of evidence.

The position of the 'accuracy-fetishist' who takes this attitude is therefore like that of a rule consequentialist for whom a *just* rule may have many instances that are *unjust* considered by themselves.

> Nor is every single act of justice, considered apart, more conducive to private interest than to public; and it is easily conceived how a man may impoverish himself by a single instance of integrity, and have reason to wish that, with regard to that single act, the laws of justice were for a moment suspended in the universe. But however single acts of justice may be contrary either to public or private interest, it is certain that the whole plan or scheme is highly conductive, or indeed absolutely requisite, both to the support of society, and the well-being of every individual.[35]

Convicting Alice but not Bob amounts to treating differently cases that are alike in point of the accuracy of the evidence that we have in those two cases, and therefore *seems* to evince a concern for something other than accuracy. But it does not: rather, it illustrates the fact, for which this paper has been an extended argument, that an *exclusive* concern for accuracy can motivate rules for the treatment of evidence whose individual instances, when considered in isolation, seem incompatible with it.

REFERENCES

Allen, R. J. and M. S. Pardo. 2007. The problematic value of mathematical models of evidence. *Journal of Legal Studies* 36: 107-43.

Bar-Hillel, M. 1980. The base-rate fallacy in probabilistic judgments. *Acta Psychologica* 44: 211-33.

Bentham, J. 1978 [1827]. *Rationale of Judicial Evidence*. New York: Garland.

Blome-Tillman, M. 2015. Sensitivity, causality and statistical evidence in courts of law. *Thought* 4: 102-112.

———. 2017. 'More likely than not': knowledge first and the role of statistical evidence in courts of law. In A. Carter, E. Gordon and B. Jarvis (ed.), *Knowledge First - Approaches in Epistemology and Mind*. Oxford: OUP: 278-29

Brilmayer, L. 1986. Second-order evidence and Bayesian logic. *Boston University Law Review* 66: 673-91.

Brook, J. 1985. The use of statistical evidence of identification in civil litigation: well-worn hypotheticals, real cases, and controversy. *St Louis University Law Journal* 29: 293-352.

Clifford, W. K. 1999 [1877]. The ethics of belief. In his *Ethics of Belief and Other Essays*. T. Madigan, ed. Amherst, MA: Prometheus: 70-96.

Cohen, L. J. 1977. *The Probable and the Provable*. Oxford: Clarendon Press.

———. 1986. *The Dialogue of Reason: An Analysis of Analytic Philosophy*. Oxford: Clarendon Press.

[35] Hume 1978 [1740] III.ii.2.

Duff, R. A. 1998. Dangerousness and citizenship. In Ashworth, A. and M. Wasik (ed.), *Fundamentals of Sentencing Theory*. Oxford: Clarendon Press: 141-63.

Enoch, D., L. Specter and T. Fisher. 2012. Statistical evidence, sensitivity, and the legal value of knowledge. *Philosophy and Public Affairs* 40: 197-224.

Hume, D. 1978 [1740]. *Treatise of Human Nature*. Ed. with an analytical index by L. A. Selby-Bigge. Oxford: Clarendon Press.

James, W. 2000 [1896]. The will to believe. In his *Pragmatism and other essays*. G. Gunn, ed. London: Penguin: 198-21.

Jackson, J. and S. Doran. 2010. Evidence. In Patterson, D. (ed.), *Blackwell Companion to Philosophy of Law and Legal Theory*. Oxford: Blackwell: 177-87.

Koehler, J. N. and D. N. Shaviro. 1990. Veridical verdicts: increasing verdict accuracy through the use of overtly probabilistic evidence and methods. *Cornell Law Review* 75: 247-279.

Neidermeier, K. E., N. L. Kerr and L. A. Messé. 1991. Jurors' use of naked statistical evidence: exploring the basis and implications of the Wells effect. *Journal of Personality and Social Psychology* 76: 533-42.

Nesson, C. 1979. Reasonable doubt and permissive inferences: the value of complexity. *Harvard Law Review* 92: 1187-92.

———. 1985. The evidence or the event? On judicial proof and the acceptability of verdicts. *Harvard Law Review* 98: 1357-92.

Pritchard, D. Forthcoming. Legal risk, legal evidence and the arithmetic of criminal justice. *Jurisprudence* (special issue on *Law and Virtues*, ed. A. Amaya and C. Michelon).

Redmayne, M. 2008. Exploring the proof paradoxes. *Legal Theory* 14: 281-309.

Thomson, J. J. 1986. Liability and individualized evidence. *Law and Contemporary Problems* 49: 199-219.

Tribe, H. L. 1971. Trial by mathematics: precision and ritual in the legal process. *Harvard Law Review* 84: 1329-93.

Wasserman, D. 1991. The morality of statistical proof and the risk of mistaken liability. *Cardozo Law Review* 13: 935-76.

Wells, G. L. 1992. Naked statistical evidence of liability: is subjective probability enough? *Journal of Personality and Social Psychology* 62: 739-52.

———, A. Memon and S. D. Penrod. 2006. Eyewitness evidence: improving its probative value. *Psychological Science in the Public Interest* 7: 45-75.