Knowledge, Evidence, and Naked Statistics

Many who think that naked statistical evidence alone is inadequate for a trial verdict think that use of probability is the problem, and something other than probability – knowledge, full belief, causal relations – is the solution. I argue that the issue of whether naked statistical evidence is weak can be formulated within the probabilistic idiom, as the question whether likelihoods or only posterior probabilities should be taken into account in our judgment of a case. This question also identifies a major difference between the Process Reliabilist and Probabilistic Tracking views of knowledge and other concepts. Though both are externalist, and probabilistic, epistemic theories, Tracking does and Process Reliabilism does not put conditions on likelihoods. So Tracking implies that a naked statistic is not adequate evidence about an individual, and does not yield knowledge, whereas the Reliabilist thinks it gives justified belief and knowledge. Not only does the Tracking view imply that naked statistical evidence is insufficient for a verdict, but it gives us resources to explain why, in terms of risk and the special conditions of a trial.

1. Introduction

Judges and mock jurors are reluctant to convict a defendant or hold them liable on the basis of naked statistical evidence alone. For a toy illustration (based on Smith 2018), suppose 24 of the 25 television sets that left the premises of your store in the course of a looting were stolen. You know that one of them was not stolen because you have a record of the sale – indeed you have the receipt itself because the customer forgot it – but due to the mayhem you have no memory

of their face, and since they paid cash the receipt doesn't tell you their name. Even if you could round up the 25 people on the street carrying the relevant tv sets, many would say we couldn't convict any of them of theft merely on the evidence that 24 of the 25 people stole the set they carried out the door.[1] Though we are 96% sure of each of them that they stole their set, as far as we know any one of them could be that one who we know paid for his.

A literature has grown around the question why, and indeed whether, this attitude is justified. Most commentators about the law think that naked statistics are insufficient as proof of an individual event and that this needs an explanation because the probabilities involved can be so high, though there are some who think that this reluctance is excessive or even misplaced (Comesaña 2009, 18; Asasi 2019; Papineau 2019). For those who think such evidence is weak there is a tendency to assume that probability is the problem or weakness – as indicating a standard of proof or as a type of evidence, respectively – and that something other than probability, something qualitative rather than quantitative, is the solution. Among philosophers of this persuasion there are those who think that legal verdicts require knowledge, not merely justified belief (Thomson (1986), Littlejohn (2017), Moss (2018)), and those who think that counterfactual conditions like safety (Pritchard 2018) or sensitivity (Enoch, Spectre, and Fisher 2012; Enoch and Spectre 2019) must be fulfilled, for knowledge or for another reason. There are those who think verdicts require justified full belief, not merely justified high degree of belief (Buchak 2014, Littlejohn 2017), and some have thought that verdicts require evidence that is causally connected in an appropriate way to the event that is being judged. (Thomson 1986) Naked statistics, it is argued, don't fulfill these conditions, and that's why they are insufficient for finding a person guilty or liable.

A naked statistic is a frequency probability for a population, a mere correlation between two properties. It is a single base rate, for example, a high frequency of A-doers in a population of B's to which the defendant belongs, and our question is whether it can be sufficient for conviction or holding liable.[2] There is no question here of statistics in general being unsuitable for proof; multiple base rates and relative conditional probabilities connecting three or more variables can be strong evidence for causation, and courts appreciate it as such. There are many kinds of probabilistic evidence, a variety of different ways of using these kinds of evidence to come to beliefs, and a variety of different ways that the probability axioms can be used to define properties salient to our epistemic success and failure.

[1] Courts have explicitly agreed in civil cases where the standard of proof is lower than in criminal cases like theft, making the preference even more striking. See, e.g., *Sargent v. Massachusetts Accident Co*. (1940, dictum); *Smith v. Rapid Transit Inc.* (1945); *Guenther v. Armstrong Rubber Co.* (1969, dictum).

[2] There are other relations the base rate could have to the claim against the defendant. For example, the doer of A could be a member of B, most of whose members the defendant is responsible for. The negative attitude toward the kind of example described in the text is striking because the relation between the base rate and the fact at issue is as simple and direct as it can be.

Naked Statistical Evidence is a single base rate:

The frequency of A's among B's is r.

$P_{FR}(A|B) = r$

It is unfortunate that distinctions among different types of probabilistic claims, and different probabilistically defined properties, have not played a role in recent philosophers' discussion of naked statistical evidence in the law, because the differences between two probabilistic approaches to epistemology – Process Reliabilism and Probabilistic Tracking – provide an apt analysis of what is at issue in questions about the quality of this sort of evidence. The question about single base rates as evidence can be formulated within the probabilistic idiom, and when we do this, I will argue, we get an explanation of the weakness of single base rates as evidence about individuals, and see how great the difference is between Process Reliabilism (PR) and Probabilistic Tracking (PT).

PR and PT are both externalist approaches to epistemological concepts. That is, they imply that the epistemic status of a belief depends in part on facts about the relationship of the belief to the world outside the subject's consciousness that go beyond the belief merely being true. In using probability they differ from externalist theories that appeal to counterfactual relationships, like safety (Sosa 1999, Pritchard 2018) – the belief couldn't easily be false – and sensitivity (Nozick 1981, Zalabardo 2017) – if p weren't true the subject wouldn't believe p – but beyond this they diverge sharply.

The classic Process Reliabilist, Alvin Goldman, thinks the epistemic status of a belief depends on a property of the history of its formation, that it was formed by a process that produces true beliefs on most occasions. (Goldman 2012) A tracking approach (whether probabilistic or counterfactual) has the epistemic status of a belief depend on the subject's dispositions to be responsive to the truth. For Probabilistic Tracking these dispositions are captured, schematically, in the conditions that there be a high probability the subject believes p given that p is true, and a high probability the subject doesn't believe p given that p is false.[3] (Roush 2005)

If there is a problem with beliefs formed on the basis of naked statistical evidence then classic PR certainly seems to be afflicted with it. Consider another toy case like that with the televisions (adapted from Nesson 1979, 1192):

> 25 prisoners are taking exercise time in the prison yard. A video camera records 24 of them attacking the guard, and shows one of the prisoners at a distance, not

[3] A simple conditional probability doesn't have the modality to be dispositional, but the conditions stated here are merely schematic. The Probabilistic Tracking conditions for knowledge are actually universal quantifications over sets of probability functions, as defined in Roush 2005, Ch. 3.

> participating. That prisoner's face is turned away, and the video is too grainy to identify any of the prisoners whose faces are showing. The guard was attacked from behind and too quickly for him to be able to identify his assailants. There is no further evidence, and none of the prisoners is talking.

24 of the 25 prisoners are guilty of assaulting the guard. However, a prosecutor is unlikely to bring charges against any of them, because a competent defense attorney for any given prisoner would point out that there is no evidence at all ruling out the possibility that their client was the one prisoner who we know did not participate in the assault, and a judge would not let the case go to a jury.[4]

If this response to such evidence is right, then the problem for classic PR is that reasoning from the known very high fraction of prisoners who assaulted the guard to a belief of each one of the prisoners that he assaulted the guard, when one has no further information, is a process that produces true beliefs on most occasions. It will produce a true belief in 24 out of the 25 cases it can be applied to in this example, that is, in 96% of cases, so beliefs that result from it satisfy the classic PR requirement for the positive epistemic status of justifiedness. The process of coming to these beliefs could be described more generally as using a known high fraction, say r > .95, of A's among B's to come to believe of a randomly chosen particular instance among the B's that it is an A, when one has no other evidence.[5] The higher that fraction is, the higher the fraction of true beliefs among beliefs formed by that process can be expected to be, and so, on a graded version of the PR view, the more justified a belief formed by that process will be.

If one shares the intuition that the frequency data are not sufficient reason to believe of any prisoner that he assaulted the guard, it will also be puzzling to consider what many think would be sufficient: eyewitness testimony or forensic evidence and circumstantial evidence identifying which prisoner was the abstainer. This is puzzling because if one has the intuition at all it persists even if we stipulate that this individualized evidence makes the probability of guilt only 95%. By the numbers, according to PR, the base rate information makes one's belief in guilt more justified, but epistemically it is felt to be weaker.

PT's analysis of the case is quite different from PR's. A subject using a base rate to decide what to believe about each case will use it to come to a probability for each case and believe that case has the property whenever that probability is greater than, say, .95. Because the base rate is .96, such a person will believe every one of these prisoners is guilty. PT-type knowledge of a prisoner's guilt would require a high probability the subject doesn't believe the prisoner is

[4] That none of the prisoners talk is crucial to this. (See Nesson 1979, 1192-1194.) If prisoner 1 is on trial for assault, and prisoner 2 bears witness that he was the one who did not participate and prisoner 1 did, then that testimony is evidence that prisoner 1 is guilty. It may be poor evidence depending on the credibility of prisoner 2, but that it exists would be enough for the judge to send the case to trial because weighing evidence has to be done by a jury.

[5] One might want to add a condition that the process through which that individual was chosen is known to be random, in the sense that it will have a .96 chance of picking a prisoner who is guilty. See Levi 1977, 10. The point can be made for whatever one thinks are the conditions for direct inference.

guilty given that the prisoner is not guilty. But our subject who is using the base rate would have a 100% chance of believing a prisoner guilty given that he is innocent, and thus would not qualify for the high epistemic status of knowledge.

A comparison between these two probabilistic approaches is complicated by the fact that what PR measures is whether or how far the subject is justified in his belief in p, whereas the PT criteria stated measure whether the person knows p. This difference will wash out in the end because Goldman takes his view of justified belief to become a view of what knowledge is when it is supplemented with a successful clause for ruling out Gettier cases, and the cases that concern us here are not Gettier cases. So, we will be able to use the criterion for justified belief combined with a condition that the belief be true, to tell us what PR requires for knowledge in these cases.

I will come back in the end to view our problem as one about whether knowledge is present, but the concrete comparison I will begin with is between what classic Process Reliabilism requires for justified belief and what Probabilistic Tracking requires for good evidence. The PT view has definitions of evidence and good evidence in addition to the definition of knowledge, and it is a consequence of these definitions that better evidence makes you more likely to know. Thus the concepts of justified belief and good evidence occupy analogous places in relation to the concept of knowledge in the two systems. Evidence as the PT view defines it also has clear comparators in the Bayesian view of evidential support of hypotheses, which will make possible a comparison between PT and a recent re-formulation of PR in Bayesian terms.

Analysis of the implications of PR and PT for the use of naked statistical evidence in court verdicts will bring into clear focus a major difference between these two epistemological views. The requirements of PT provide a diagnosis and explanation of the weakness of naked statistical evidence for court, while the requirements of PR offer no restriction on, or warning about, the use of this evidence in any context. We will see that the concrete benefits and harms that PT is concerned with are highly salient to trial verdicts, which will show that it provides the superior set of conceptual tools for this context.

2. Process Reliabilism(s)

A non-legal case similar to the ones we have considered was used by Jonathan Adler to argue against Process Reliabilism as a view of what justified belief is. (Adler 2005) Suppose a manufacturer of widgets knows that one in a thousand of the widgets it produces is defective in a particular way due to an imperfection in the manufacturing process. An industrious manager wants to reduce the percentage of defective widgets that reach the market from his plant, so he introduces a type of detector that is to be applied to each widget at the end of its assembly. This detector has an error rate of .003, when used by an ace technician.

Two workers at the factory, Smith and Jones, respond differently to the order from the manager to use one of the detectors. Jones is industrious, and an ace technician. She follows the instruction to use her detector and approves those widgets the detector says are okay, and sets aside widgets the detector says are defective. Smith doesn't use his detector at all and approves every widget that goes by. The manager is furious at Smith and threatens his job, but Smith protests that he's using a method that is both more accurate and more efficient than fussing with the detector. His error rate is one in a thousand (since only one in a thousand widgets is defective). The error rate of the detector is three in a thousand.

Adler thought it was obvious that the manager was right on epistemic grounds to disapprove of Smith's method and prefer Jones' method, and that this case presents a serious problem for the classic PR view of justified belief. On that view, all of Smith's beliefs about the widgets that go by him, which are all to the effect that the widget is not defective, are justified, because 999 of 1,000 beliefs he forms by this method will be true.[6] If we allow for gradations of justifiedness, Smith's beliefs must also be counted by classic PR as more justified than those of Jones, because the rate of true beliefs among all beliefs, i.e., the reliability, of Jones' process is lower, .997 vs. .999.

Adler thought it was clear that PR gives the wrong verdict about Smith, and that Jones' method is superior. This is in part because he thought Jones' method justified her in having a *full* belief rather than merely a *high degree of* belief that widgets that pass the detector test are not defective, and because her method makes those beliefs of hers *knowledge* (when they are true). Smith's method would give him an advantage if they were *betting*, but it does not give him knowledge. (Adler 2005, 447-448) These reactions mirror the reactions noted above that some epistemologists have to the legal cases.

I find this diagnosis of the situation mysterious and unsatisfying. Imagine that instead of widgets this factory made N95 masks for health care workers. A defective mask means that the worker who uses it is not protected against microbes their patients might be breathing out. Whether the mask is defective doesn't merely affect customer satisfaction surveys, and in this version of the case getting it right more often seems the most important matter, even the most important epistemic matter. Seen this way, invocation of intuitions about the presence or absence of justified *full* belief and *knowledge* sounds like incantation. Smith will get it right more often; why does anything else matter?

Although I am as much a fan of knowledge and the concept of knowledge as anyone is, in what follows I will avoid the term "knowledge" and appeals to intuitions about it to make comparisons, in order that my conclusions not depend on these things. Those conclusions will have implications about whether knowledge is present, but only by means of what the definitions of knowledge do with the factors we analyze. I will focus first on comparing the tangible epistemic benefits that each approach to the types of evidence about widgets or masks

[6] Someone who has a standard higher than .999 can change the example to make Smith's rate even higher.

or assault-defendants brings with it, expressed in terms of probability, and argue that the list of relevant benefits of fulfilling the tracking criteria have a crucial value in some contexts.

If one is interested in tangible benefits, how could one possibly think that Jones' reliability of .997 is epistemically better than Smith's reliability of .999? Smith has a higher number – with his method we can expect[7] fewer mistakes per thousand beliefs – but the percentage of true beliefs among total beliefs formed is an underdescription. Given Smith's policy of believing of every widget that goes by that it is OK, there is a kind of error he is guaranteed never to make – believing a widget is defective when it is not – and a kind of error that he is guaranteed never to catch – believing of a widget that it is OK when it is defective. With Jones, it is likely that 3 out of 1,000 beliefs she forms are false, but she is equally likely to make a mistake when given a defective widget as to make a mistake when given a non-defective widget. There is no error she is guaranteed not to make, but there is also no error she is guaranteed not to catch.

The consequences of Smith's and Jones' distinct distributions of types of error are easier to see if we scale up the volume of widgets. Suppose 1,000,000 widgets go through the factory and 1,000 of them are defective, 999,000 OK. Smith would make only 1,000 errors about the widgets whereas Jones would likely make 3,000 errors, but their errors would be different. For the 999,000 widgets that are OK, Smith would form the correct belief that they are OK, but of every one of the 1,000 that are defective, Smith would mistakenly believe that they are OK. Of the 999,000 that are OK, Jones would likely correctly believe of 996,003 them (.997) that they are OK, but incorrectly believe of 2,997 of them (.003) that they are defective. And of the 1,000 that are defective Jones would likely correctly believe of 997 of them that they are defective, and incorrectly believe of 3 of them that they are OK.

Out of a million widgets, Jones would likely only let 3 defective widgets go to market, whereas Smith would allow 1,000 defective widgets through. Out of a million widgets Smith will never prevent a non-defective widget from going to market, whereas Jones will likely prevent 2,997 non-defective widgets from going to market. Thus far one might think there is nothing intrinsically better about Smith's or Jones' method. The widget manufacturer might fire the manager for introducing a procedure that deprives the company of the profits from nearly 3,000 perfectly fine widgets. The N95 mask manufacturer might thank the manager for a procedure that prevented potentially nearly 1,000 health care workers' deaths from being on their conscience. We would need to decide on the relative importance of the two types of

[7] Adler stipulates, and Comesaña emphasizes, that a difference between Smith and Jones comes from the fact that the manufacturer *knows*, for sure, that exactly one out of *every* thousand widgets is defective, but this can't be known because it can't be true. Imagine 10,000 widgets made by this factory. There are 10,000!/(1,000!(9,000!)) different 1,000-member subsets of this set. Even if exactly ten of the 10,000 widgets are defective, it cannot be the case that every one of those 1,000-member subsets has exactly one defective widget. Whether one or more, or less, defective widget(s) come(s) through in, say, the first 1,000 widgets off the assembly line is a probabilistic matter.

errors in order to know which method is better for us. A focus merely on number or proportion of errors will not give us a preference for Smith's or Jones's method.

Depending on which type of error we are more concerned about we may prefer Smith's or Jones' method, but classic PR will take the level of justifiedness of the resulting beliefs to depend only on the fraction of true beliefs over total beliefs that we can expect the two methods to yield. So classic PR is committed to Smith's method giving beliefs that are more justified than those Jones' gives. It allows a belief to be ranked as more justified without any requirement about avoiding the error of believing a widget is OK when it is not. Should justifiedness depend on anything about what type of error one is vulnerable to? Is there anything but intuition to appeal to here? Yes, there is quite a bit more.

If we view the workers' situations through the lens of degrees of belief constrained to be probabilities, then there is a clear sense in which Jones' method is epistemically better, pointed out by Comesaña (2009). Smith has a right to a .999 credence of each widget that it is OK, that comes via direct inference from the known frequency of OK widgets among all widgets. Jones knows this frequency so her prior degree of belief of any widget that it is OK is .999, to which she applies the evidence she receives from the detector. Let:

e = Detector says OK
H = Widget is OK

Then Jones' prior probability of H, P(H), is .999. In the cases where the detector says the widget is OK, Jones' posterior degree of belief in H, the probability of H taking e into account, Pr(H|e), is

$$\Pr(H|e) = \Pr(e|H)P(H)/P(e)$$

$$= (.997)(.999)/(.996006)\,[8]$$

$$= .999997$$

.999997 is greater than .999, so the degree of confidence in the non-defectiveness of a widget that Jones has a right to when the detector says the widget is OK is higher than the degree of confidence Smith has a right to when he makes that judgment automatically.

For one who wanted to understand why Jones' method is epistemically better, this may seem thin gruel – a difference in the fourth decimal place – but it corresponds to the fact that out of a million widgets coming down the line, we can expect that Jones will only believe of three of them that they're OK when they are not whereas Smith will make this mistake on 1,000 of the widgets.

Comesaña thinks that though this confirms that classic PR is mistaken to attribute higher justifiedness to Smith, this is not a problem for the general PR approach. It just shows that

[8] $\Pr(e) = \Pr(e|H)P(H) + \Pr(e|{-H})\Pr(-H) = (.997)(.999) + (.003)(.001) = .996006$

classic PR misidentified what reliability consists in. Reliability of a process of coming to believe p should not be measured by the fraction of true beliefs out of all the beliefs that the process you use would yield, he thinks, but by the posterior probability of H given the evidence one has used (which evidence also serves to identify the process used). (Comesaña 2009, 6) On this *High Conditional Probability* (HCP) view, Jones' belief that a given widget is OK, formed by her process of using the detector, is more reliable, hence more justified, than Smith's belief of a given widget that it is OK formed by his process.

However it is a mistake to think that discovery of a higher posterior probability of H for Jones is a defeat for measuring reliability by ratios of true beliefs to all beliefs, and hence a defeat for classic PR. What Comesaña's comparison of posterior probabilities has done is restrict attention to beliefs in H and ignored beliefs in -H. Expected ratios of true belief over all beliefs formed by a process can also be separated into two classes, ratios for all beliefs in H and ratios for all beliefs in -H. If we focus only on beliefs in H, then Jones' ratio of true beliefs among all those beliefs will also be higher than Smith's. Recall what happens when our two workers apply their methods to 1,000,000 widgets coming down the conveyor belt of this factory. Jones can be expected to have a total of 996006 beliefs of some widget that it is OK, and 996003 of those will be true.[9] Thus, Jones' fraction of true beliefs among all H-beliefs is .999997. Smith's fraction of true beliefs among all his H-beliefs is merely .999. If we restrict attention to beliefs in H, as the HCP view did, then classic PR *also* says that Jones' process is more reliable, hence that her H-beliefs are more justified than those of Smith.

3. Apples vs. Apples and Oranges

Adler's comparison of Smith and Jones and Comesaña's treatment of it don't defeat classic PR, and they also don't get to the heart of the difference between the two kinds of evidence. One reason the comparison misfires is that Jones' process contains Smith's as its first step. Restricting attention to beliefs in H, that a widget is OK, the only mistake there could be is a case where H is not true, that is, where the widget is defective. Jones uses the base rate information that Smith uses, which assures us that only one in a thousand H-beliefs will be mistaken, and then applies a filter that removes more mistaken H-beliefs to a point where only 3 out of 996006 (~.000003) H-beliefs will be mistaken. Classic PR, High-Conditional-Probability PR, and common sense say that Jones has more right to confidence in her H-beliefs than Smith does. If we compare an apple to an apple and an orange, we shouldn't be surprised that the second set has more fruit. Likewise no one should be surprised if when we *add* 95% reliable eyewitness and forensic evidence about the prisoners who attacked the guard to a 96% base rate of prisoners who attacked, we feel more confident convicting everyone except the prisoner

[9] Of the 1,000 widgets that are not OK Jones believes of 3 of them (.003) that they are OK. Of the 999000 widgets that are OK Jones believes of 996003 of them (.997) that they are OK.

identified by the eyewitness and forensics as not participating; the posterior probability of guilt for the remaining prisoners in that case is not .96 but .998.[10]

The contrast between using a base rate and using eyewitness testimony or forensic evidence that we were looking for in the legal cases was not between base rate and base rate plus individualized evidence, but between base rate and individualized evidence alone. A similar point applies to the widgets. We've compared Smith's process to Smith's process plus an enhancement (on the H-belief side), but we'd get a better sense of the difference between the two types of evidence if we compared what Smith does to what Jones would be doing if she didn't use the information Smith uses, but only the detector. Call the person who uses only the detector "New Jones". Only using the detector means that instead of having a .999 prior probability for each widget that it is OK, she has ignorance, which we can represent as Pr(H) = .50. With her error rates (in both directions) of .003, New Jones has a posterior probability Pr(H/e) = .997 when, following the detector's output 'OK', she concludes of a widget that it is OK. Smith will be entitled to a higher degree of belief that a widget is OK (.999) than New Jones will, and Smith's truth ratio in the H-beliefs (and all beliefs) will be higher than New Jones' will. Classic PR and HCP-style PR will *both* say Smith is more justified.

However, even this isn't the most apt comparison. The difference between the two kinds of evidence we're comparing is that Smith is using a frequency probability and Jones is using a detector. A fair comparison between the two, that highlighted only this difference, would be one where we gave them the same numbers: Smith uses a base rate of .999, and the detector of a worker I'll call "New New Jones" has an error rate of .001 in both directions – when presented with a defective widget and when presented with an OK widget. Like New Jones, New New Jones does not have the information Smith uses, so she has a prior probability Pr(H) = .50 that a given widget is OK. In this case, Smith's probability of H for a given widget is .999, and taking e as "detector says OK", New New Jones' posterior probability Pr(H|e) is the same, .999. Thus, posterior probability does not distinguish the two types of evidence, and Comesaña's HCP can give no guidance as to a difference between them. Likewise, classic PR will take the H-beliefs formed by Smith and New New Jones to be equally justified because for both of these workers 99.9% of their H-beliefs, and 99.9% of their total beliefs, will be true. Is there really no difference between the two types of evidence?

[10] The calculation of the posterior for H (guilty) would go as with Jones, assuming that the eyewitness is equally likely to err when presented with the abstainer and when presented with a different prisoner: Pr(H|e) = Pr(e|H)Pr(H)/Pr(e) = (.95)(.96)/(.914) = .998. The comparison between eyewitness testimony and base rates is often presented as an either-or, but once one learns the base rate it can be intuitively difficult to ignore it, so one's intuitions about the case might reflect it.

4. Likelihoods

Classic PR takes the epistemic status of a given belief formed by a given process to depend on the frequency (or, more generally, probability) of true beliefs among some larger set of beliefs that could be produced by that process under some appropriate set of conditions. The HCP version of PR takes the epistemic status of a belief attitude towards a proposition to rest wholly on the probability the proposition has given the evidence on which the belief or degree of belief was based. Neither view puts any conditions on the likelihood probabilities, the probability of the evidence given the hypothesis, Pr(e|h), and the probability of the evidence given the negation of the hypothesis, Pr(e|-h), and it is in these that the difference between the two types of evidence we have been inspecting lies. Though Smith and New New Jones have equal probabilities of the hypothesis given the evidence, they have different likelihoods.

Likelihood of h: Pr(e|h)
Likelihood of -h: Pr(e|-h)

To see this we can think of what Smith and New New Jones are doing as the application of diagnostic tests to the widgets. Call a particular widget coming down the line "n". The task of Smith and of New New Jones is to determine whether H(n), n is OK, or -H(n), n is defective. Smith's method of doing this is to consult the frequency of OK widgets in the set n belongs to; [11] if that frequency is ≥ .999 he approves widget n, otherwise not. Call this the "base-rate test". New New Jones applies her detector to n; if the detector says OK she approves widget n, otherwise not.[12]

We can represent the test result, or evidence, each of them uses this way:

e_S(n): Base-rate test says OK

e_J(n): Detector says n is OK

When e_S(n) is true of n, Smith moves n along the assembly line, and otherwise not. When e_J(n) is true of n, New New Jones moves n along, and when e_J(n) is false, New New Jones sets n aside.[13]

[11] This is elliptical for: he consults the frequency of the most specific set that n belongs to for which we have a frequency.

[12] I assume the detector always has an output, which is either "OK" or "defective".

[13] Smith and New New Jones would often have degrees of belief, and/or beliefs, in H and -H. Belief states would be important if we were attributing knowledge or justified belief states. But I am assuming the diagnostic tests here have likelihood probabilities regardless of which kind of unconditional belief attitudes a subject may have about the hypothesis, so I leave the latter in the background for now.

An ideal test would always have a positive test result when the property in question is present, and a negative test result when the property is not present. Thus, the two workers' widget tests would be ideal if they had these likelihoods:

$Pr(e_S(n)|H(n)) = 1$ and $Pr(e_S(n)|\text{-}H(n)) = 0$

$Pr(e_J(n)|H(n)) = 1$ and $Pr(e_J(n)|\text{-}H(n)) = 0$

Smith's test would be ideal if the probability that the base-rate test *says* n is OK given that n is OK were 1, and the probability that the base-rate test *says* n is OK given that n is not OK were zero. New New Jones' test would be ideal if the probability that the detector *says* n is OK given that n is OK were 1 and the probability that the detector *says* n is OK given that n is defective were zero. Then the truth values of the e(n)'s would co-vary with the truth values of H(n). In the case of diagnostic tests, likelihoods are error rates, the rate of true positives – the probability of a positive test result given that the hypothesis that the property is present is true – and the rate of false positives – the probability of a positive test result given that the hypothesis is false.

True Positive Result:	e and h
False Positive Result:	e and -h
True Positive Rate:	Pr(e|h)
False positive Rate:	Pr(e|-h)

Empirical tests are not ideal. They typically do have false negatives (here $\text{-}e_S(n)$ and H(n), or $\text{-}e_J(n)$ and H(n)) and false positives (here $e_S(n)$ and -H(n), or $e_J(n)$ and -H(n)). What we have been told about Smith and Jones allows us to fill in the actual values of the likelihoods for Smith's and New New Jones' methods. For Smith the likelihoods are:

$Pr(e_S(n)|H(n)) = 1$ and $Pr(e_S(n)|\text{-}H(n)) = 1$

The probability that Smith's test says widget n is OK is 1 whether widget n is OK or defective, which is a way of saying that we can see by his procedure that he will conclude widget n is OK, for all n.

This is not because the base rate is given independently of inspection of the individual widgets; that is true in this case, but not in all relevant cases, as we will see shortly.[14, 15] The reason these likelihoods are both 1 is that the way the base rate is being appealed to by Smith does not allow for the possibility that the base rate of n's group is some other value than .999. It is not that it's not possible for the base rate to *be* some value other than .999, but in the procedure Smith is using to come to his conclusions it is not possible to discover that the base rate is some other value. The test asks whether the base rate of n's population is ≥ .999, and the phrase "base rate of n's population" is being used de re, for the actual value, not de dicto, for whatever the value happens to be. Representing the procedure as a test just makes vivid that limitation of the possibility space of Smith's reasoning. When presented with a new widget Smith does not say, Gee, I wonder what its base rate is. Let me check. Imagining him saying that gives up the game.

The "test result" Smith uses is going to be positive regardless of whether n is OK or defective. Though when his method is applied to this manufacturer's population of widgets, Smith will only be wrong in 1 of 1,000 cases, his false positive *rate* – the probability he classifies a defective widget as OK – is 100%.

About the detectors Adler stipulated that "the probability of an error in any evaluation is .003" (Adler 2005, 447), which means that the false positive and false negative rates are both .003. For New New Jones this number is .001, so the likelihoods for New New Jones are:

$Pr(e_J(n)|H(n)) = .999$ and $Pr(e_J(n)|\text{-}H(n)) = .001$

New New Jones' rate of false negatives $Pr(\text{-}e_J(n)|H(n))$ is .001, so the rate of true positives $Pr(e_J(n)|H(n))$ is .999, slightly less than Smith's 100%. The rate of false positives $Pr(e_J(n)|\text{-}H(n))$ is also .001, dramatically less than Smith's 100%.

The likelihoods formalize something noted above in words, that our two workers have different error profiles, and this qualitative difference persists even when we give them the same numbers: New New Jones is equally likely to make either kind of error, whereas Smith is guaranteed never to make a false negative error and never to catch a false positive error.

[14] Given the way the base rate was determined in this case, via something about the manufacturing method, one might think that the frequency of defective widgets is independent of the defectiveness or not of the individual widgets produced, but that is wrong. Those individual widgets and the base rate have a common cause in the manufacturing method. In the next case we will see, the base rate and the individuals' properties are directly interdependent. So, the claim about the likelihood here is not that the base rate of A's among B's is independent of whether the individual B's each have the property A. Obviously, one more or less B with property A raises or lowers the fraction of A's among B's.

[15] The explanation of the value 1 here is also not that we have a psychological state of certainty about the base rate. The likelihoods are determined by the probability function that precedes learning new evidence. Since $e_{PA}(n)$ is not a necessary truth, $e_{PA}(n)$ wouldn't have probability 1 until we learned it.

Formulating their errors in terms of likelihoods also brings out that Smith's success depends more on the population to which he applies his method than (New) (New)[16] Jones' success does. The error rates of (New) (New) Jones' detector only apply to widgets made by this manufacturer (because the detectors are tuned to the defect(s) created by their process), but her fractions of successes and failures are likely to be the same, or nearly the same, regardless of the proportion of defective widgets in the subpopulation of this manufacturer's widgets that she actually applies the detector to. In comparing Smith and Jones above, we effectively assumed that the subset of widgets they actually apply their methods to has one defective widget for every thousand, but in reality some subsets have a different proportion of defective widgets than the whole population has. Our manufacturer produces tens of millions of widgets, so even if a proportion of 1 in 1,000 overall are defective, some 1,000-member samples are unrepresentative. Out of tens of millions of widgets we could cobble together a subpopulation of 10,000 widgets of which 5,000 are defective. If New New Jones applies her detector to these she is likely to have 5 false positives and 5 false negatives, 1 in 1,000 for each type of error, similar to what she is likely to have for any other sample.

Not so for Smith. If he gets a representative sample of 10,000, in which exactly 10 widgets are defective, then he will be wrong in 10/10,000, that is 1/1,000 of the cases he "tests". However, if he gets an unrepresentative sample of 10,000 where 5,000 are defective, then he will be wrong in 5,000 cases – he never disqualifies a widget as defective – so he will be wrong in 50% of the cases. Smith's actual rate of success in judging widgets depends sensitively on whether he's being fed a representative sample of the population whose base rate he is using. Because he is bound to misclassify 100% of the defective widgets that come his way, the fraction of actual errors among his verdicts is guaranteed to be greater the more unrepresentative (in the defective direction) the sample of widgets he is fed. The larger the sample he is fed, the less likely it is to be unrepresentative, but the smaller the sample the more likely it is to be unrepresentative.

Early on I said that Smith's and Jones' different error profiles don't seem to show that one method is intrinsically better than the other. Which is better depends on how much you care about false positives compared with false negatives. Sending a defective can opener to market (false positive) will cause customer dissatisfaction, but defective can openers are usually useless rather than dangerous, so preventing one more false positive won't be worth having 100 more false negatives, where 100 perfectly fine can openers don't go to market. If it's N95 masks, though, a false positive could be a cause of death, and preventing that might be worth forfeiting the profit from 100 perfectly good masks. However, expressing the workers' profiles in terms of likelihoods exposes something further. Even if Smith and New New Jones do not differ in the relative disutility they attach to false positive and false negative errors, and even if

[16] For economy I am using "(New) (New) Jones" to refer to all of the Jones characters.

the expected utility of using Smith's and New New Jones' methods for judging widgets is the same, their risk profiles are different.

Suppose that the payoff for correctly classifying one thousand widgets is 10 and of incorrectly classifying (all) one thousand widgets is -10. Then the expected utility of using Smith's method on 1,000 of the manufacturer's widgets is 9.98, and the expected utility of using New New Jones' method is also 9.98.[17] But consider the worst-case scenario in which the sample of 1,000 widgets Smith and New New Jones are actually presented with is wholly unrepresentative of the widget population as a whole; in the worst case, 100% of the widgets in this set are defective. In this scenario New New Jones' likely payoff is still 9.98, because she will likely misclassify 1 in 1,000 of these defective widgets. By contrast in this scenario Smith's payoff is -10, because he's guaranteed to misclassify 100% of these widgets. Smith and New New Jones have the same expected utility, but fare dramatically differently in this worst-case scenario. Even if you regard a single false positive assessment as equally costly as a false negative assessment, if you are risk averse about false positives, then you should go with New New Jones' method.

In the probabilistic tracking view of evidence (Roush, 2005, Ch. 5), the quality of e as evidence for h depends on the two likelihoods that give us the error rates of diagnostic tests:

Pr(e|h)

Pr(e|-h)

The first, the probability of evidence e given that the hypothesis h is true, should be high, and the second, the probability of evidence e given that the hypothesis h is false, should be low. The proposition e whose truth or falsity you want to use as an indicator of h's truth or falsity should have a high true positive rate for h and a low false-positive rate for h. Falsely concluding h and not concluding h when h is true are the two kinds of errors you could make. The best evidence for h minimizes the rates of both of those types of errors.

Applying these points to our prison-assault example above, our task as a trier of fact faced with prisoner defendants in a courtroom is to decide whether a given prisoner, n, is guilty of assaulting the guard, or not guilty. If we're like Smith rather than the Joneses we use the base rate to decide this, and the only way it makes sense for us to do this will give the same verdict for every prisoner, because the evidence is the same for every prisoner. To explain this we can imagine a base-rate test that uses the probability .95, say, as the threshold beyond which we will take it that there is no reasonable doubt that the prisoner is guilty. The test for each n asks whether the base rate of assaulters in n's population is >.95 or not. If the base rate is >.95, then the test returns the result "GUILTY". If that base rate is ≤ .95 then the test returns the result "NOT GUILTY".

[17] Smith: (.999)(10) + (.001(-10) = 9.98. New New Jones: (.999)[(.999)(10) + (.001)(-10)] + .001[(.999)(10) + (.001)(-10)] = 9.98

Let
$e_{PA}(n)$: The base-rate test says n is GUILTY.
H(n): Prisoner n assaulted the guard

The evidence we will be using to decide whether n is guilty is thus $e_{PA}(n)$ or $-e_{PA}(n)$, whichever result our test gives for n. But as with Smith the result of every one of these tests can be seen ahead of time. Though it is possible for the base rate to *be* different from .96, it is not possible for the trier of fact to *find that out* in the decision problem we have given him. Because he has no evidence at all besides the base rate, there is nothing on which to base a finding that the base rate is different from .96, so the test can't possibly come out negative. Thus $Pr(e_{PA}(n)) = 1$ for all n, from which it follows that the probability of n's getting a GUILTY test result given that n is not guilty, is also 1. I.e., $Pr(e_{PA}(n)|\text{-}H(n)) = 1$.[18] Drawing a conclusion about a prisoner on the basis of the base-rate test, one is bound to find him guilty even if he is innocent. The false positive rate for guilty verdicts on the basis of a base-rate test is 100%.

[18] One might wonder how it could be that the probability of a GUILTY outcome in the base-rate test, $e_{PA}(n)$, given the hypothesis that n is innocent, -H(n), is 1, since this implies that $e_{PA}(n)$ cannot raise the probability of H(n); surely the base rate did bring up the probability of H(n) from wherever it was before to .96. And how could $Pr(e_{PA}(n)|\text{-}H(n))$ be 1 when the base rate and whether a given prisoner assaulted the guard are not independent matters?
$e_{PA}(n)$ is not a statement of the base rate, but is extensionally equivalent to the statement that the base rate is ≥ .95. The statement of the base rate is $Pr_{FR}(H|P) = .96$, for H those who assaulted the guard and P those who are prisoners in the yard, and that does raise the probability of guilt for every prisoner. Since the base rate is a frequency probability and individuals don't have frequencies, a posterior probability of guilt for n is arrived at by a direct inference. However we can simplify here by ignoring the fact that the base rate is a statement of probability, and assuming the conditions for direct inference are met (the individual prisoner is chosen randomly and there is no other information about him except the base rate). If so, we can view the inference as a simple conditionalization that reflects the fact that one more or less prisoner participating in the assault changes the base rate, and vice versa. Let
p: 24 of 25 of the prisoners assaulted the guard
H(n): n assaulted the guard
and assume as background that at least one of these prisoners assaulted the guard, that no one other than them did so, and that whether a given prisoner assaults the guard is independent of whether any others do. For the prior probability that prisoner n assaulted the guard, let
$Pr(H(n)) = r$ for all n
Then,
$Pr(p) = (25)(1 - r)\, r^{24}$ and
$Pr(p|H(n)) = (24)(1\text{-}r)\, r^{23}$
So, by Bayes theorem,
$Pr(H(n)|p) = [r(24)(1\text{-}r)r^{23}] / [(25)(1\text{-}r)\, r^{24}] = r(24)/r(25) = 24/25$
Every prisoner gets the same posterior probability of guilt, which does not depend on the prior probability. So, for any prior, r, below 24/25, the base rate raises the probability. Also, the likelihood of -H with the base rate evidence, $Pr(p|\text{-}H(n))$ does not equal 1, but r^{24}, which makes the base rate qualify as very good evidence indeed, on the PT account of evidence. However, as with e_{PA}, if we are to use this to decide whether to find n guilty we will have a diagnostic test that can only give one outcome, because it is not possible for the subject to find out that the base rate is different from what he's actually measured it to be; this is due to the condition on direct inference that said he must not have other information relevant to n than the actual base rate.

The false positives in this kind of example will be cases where innocent people are convicted of crimes, so it is no surprise that our judicial system is wary of conclusions based on base-rate tests. If we rest our case only on an e for which Pr(e|-h) = 1, then we are guaranteed to convict any innocents there might be. David Enoch and Talia Fisher (2015, 571) argue that concern about convicting the innocent does not warrant regarding single base rates as insufficient for conviction or holding liable because 1) in an imperfect system, which they all are, we are more or less guaranteed to convict some innocents unless we follow a rule of convicting no one, and 2) in any given case we are not guaranteed to convict an innocent even if there is one, for the innocent exception may happen not to be on trial. (They might have successfully escaped.)

But this is a distortion of the decisions we face about what rules to give to our institutions. Of course we know that procedures are never perfect, and if we are to convict any guilty people then some innocents will likely be convicted too, but there is a difference between having a system that we can be sure will yield errors of some sort or other, and an innocent who happens to be or not to be on trial, and consciously embracing a rule that you know is guaranteed to convict any innocent party that might be on trial.[19] An analogy has sometimes been made to constructing a highway: for a long enough highway we can been pretty sure there will be at least one death of a construction worker, but we don't for that reason refrain from building highways. (Shaviro 1989, 536) But the proper analogy would be to building a highway with a piece of equipment that you know for a fact will explode and kill the third person who uses it. Building a highway is not a crime, and that equipment may happen not to get used a third time, but if it does then it will be negligent homicide.

The value of the likelihood Pr(e|-h) also goes to explain why we resist single base rates as sufficient evidence even when they yield the same high posterior probability for the hypothesis as eyewitness testimony does. If Pr(e|-h) = 1 then one is guaranteed to have e even given that the guilt hypothesis about prisoner n is false. To describe eyewitness testimony with the same value as evidence we would have to be imagining an eyewitness who will testify that n is guilty *no matter what*. When we talk about eyewitnesses of imperfect reliability, even poor reliability, this is not what we have in mind. If we do imagine this, then it is also obvious that the testimony of that eyewitness is no kind of evidence. In a recent defense of the idea that there is no principled difference between a single base rate and "individualized" evidence, David Papineau (2019, 8) compares only posterior probabilities for the two types of evidence, and so

[19] Enoch and Fisher's claims are the following. 1) For all x if x is a system of justice made and used by imperfect human beings, then it is overwhelmingly probable that if there exists a y such that y is guilty and y is convicted by this system, then there exists a z such that z is innocent and z is convicted by this system. 2) For all x, if x is innocent it is possible he is not on trial, so it is possible that he is not convicted. My claim is that for all x if x is a system of justice that regards base rates as sufficient for conviction, then for all y, if y is put on trial in a case for which the only evidence is a single base rate, and y is innocent, and the prosecutor is competent, then y will be convicted.

fails to take the error rates into account, and in particular the fact that the typical unreliable eyewitness does not have Pr(e|-h) = 1 as the outcome of a base-rate test does.[20]

I have stressed the role of false positive error in trials, but this may seem to apply only to criminal trials, where the high proof standard *beyond reasonable doubt* expresses our resistance to convicting the innocent. It has been argued that the typical standard of evidence in civil trials, *preponderance of evidence*, often interpreted as requiring only a posterior probability greater than .50, expresses a valuation of a false positive (erroneously holding liable) as no worse than a false negative (erroneously clearing of liability), making it all the more puzzling why courts do not regard a single base rate as sufficient in a civil trial. (Schauer 2003, 90) However, the problem of false positive error that we have seen does not depend on a subject's regarding false positive error as worse than, or having lower utility than, false negative error, and the posterior probability makes no difference to it. Taking a single base rate as evidence gives one a 100% false error *rate*, Pr(e|-h) = 1, which is compatible with attaching equal disutility to false negatives and false positives, and with a high or low posterior probability for the hypothesis. Utilities, likelihoods, and posterior probabilities are all independent of each other, so all of the points above about error risk profile and vulnerability to unrepresentative samples apply to civil trials just as they do to criminal trials.

5. Likelihood Ratios and Posterior Probabilities

On the Bayesian view, e evidentially supports h if and only if Pr(h|e)/P(h) > 1 – that is, h has a higher probability given e than it does on its own – which is true if and only if the Likelihood Ratio (LR), Pr(e|h)/Pr(e|-h), is greater than 1, which it cannot be if Pr(e|-h) = 1.

Prior Probability of Hypothesis: Pr(h)
Posterior Probability of Hypothesis: Pr(h|e)
e *evidentially supports* h: Pr(h|e) > Pr(h), or Pr(h|e)/Pr(h) > 1
Bayes Theorem: Pr(h|e) = Pr(e|h)Pr(h)/Pr(e)
Likelihood Ratio (*LR*): Pr(e|h)/Pr(e|-h)

[20] This also addresses a line of thought that says our resistance to base rates alone is merely due to the fact that in the statement of a base rate the possibility of error is made explicit and quantified, whereas the potential for error with eyewitnesses is only implicit. (Shaviro 1989) There is a real difference in the false positive error rates, and sensing it doesn't require higher math.

This makes sense intuitively because LR > 1 says that the evidence in question is more likely when the hypothesis is true than when it is false. A higher LR takes a given prior probability of h to a higher posterior probability for h; how far the LR is above 1 is one way that a Bayesian can measure the degree of evidential support a piece of evidence provides to a hypothesis, and it is the definition that is consistent with the Probabilistic Tracking view of evidence.[21] Judged by their LR's, New New Jones, New Jones, and Jones, all have dramatically better evidence than Smith does:

Smith:	$P(e_S(a)\|Ha) = 1$ and $P(e_S(a)\|\text{-}Ha) = 1$	LR = 1
New New Jones:	$Pr(e_J(a)\|Ha) = .999$ and $Pr(e_J(a)\|\text{-}Ha) = .001$	LR = 999
New Jones:	$Pr(e_J(a)\|Ha) = .997$ and $Pr(e_J(a)\|\text{-}Ha) = .003$	LR ≅ 332
Jones:	$Pr(e_J(a)\|Ha) = .999$ and $Pr(e_J(a)\|\text{-}Ha) = .001$	LR ≅ 332

But on any Bayesian view of how to measure degree of evidential support, (New (New)) Jones' evidence is better than that of Smith, because all such measures agree that for e to positively support h at all requires $Pr(h|e)/Pr(h) > 1$, which is true if and only if LR > 1, and some evidential support is better than none.

Comesaña's High Conditional Probability version of Process Reliabilism was able to deliver the verdict that Jones' belief that a given widget is okay is more justified than Smith's by means of the fact that Jones' posterior probability that the widget is OK given that the detector says it is OK, .999997 (using a prior probability of .999), is higher than the probability that the widget is OK that Smith gets from the base rate, .999. He notes that Smith's LR is 1[22] and Jones' LR is roughly 332, and he allows that the Likelihood Ratio would be an alternative way of defining reliability in a reliabilist view of justified belief, but considers that definition to be something you would use only in the unfortunate case where the proposition h "just doesn't have an unconditional probability". (Comesaña 2009, 16)

I am not interested in defending high LR as a criterion by calling it "reliability", but Comesana's preference for high posterior over high LR as a way of measuring how justified a belief is faces a problem. The LR's comparative assessment of the quality of Smith's and Jones' evidence for the hypothesis that a widget is OK matches (in direction though not in degree) the HCP's assessment of their comparative justifiedness in believing that hypothesis. However the two ways of assessing diverge in their comparison of Smith and New New Jones. The LR assessment has New New Jones, analogously to Jones though better, with dramatically good evidence for the hypothesis that a widget is OK – LR = 999 – and it has Smith's test giving him no evidential

[21] There are many others. See Christensen (1999) and Fitelson (1999) for discussion. For arguments in favor of the LR measure see Good (1983) and Good (1985), Roush (2005, 163-165), and Zalabardo (2009).

[22] This is why the conditionalization that Comesaña represents Smith as making has the same prior and posterior probabilities for the hypothesis that widget n is OK, namely .999, although this is obscured in the presentation. (Comesaña 2009, 11)

support at all for that hypothesis – LR = 1. But, as noted above, the HCP view has Smith and New New Jones coming out as equally justified because they have the same posterior probability for the hypothesis: .999.

The two ways of assessing don't agree in all cases, and they disagree in particular in the case most apt for evaluating the views on their treatment of naked statistical evidence. When a base rate stands head-to-head against evidence with a high LR, where both yield a high posterior probability, a belief formed via the base rate is just as good, says the HCP view. The classic PR view has the same verdict, since both Smith and New New Jones can expect to have 999 out of 1000 of their beliefs that a widget is OK be true.[23] On this analysis, Comesaña's revision of PR doesn't change anything with regard to naked statistical evidence, but he is not troubled by this consequence, for he "… think[s] that the mistrust [in pure statistical evidence] is misplaced, …" (Comesaña 2009, 18, fn. 26) Whatever Comesaña's preferences, the choice between the LR evaluation and the HCP (or classic PR) evaluation of Smith and New New Jones is where the question about the quality of naked statistical evidence, and its ability to justify belief, resides.

There is another contrast between high posterior probability and high LR that matters pragmatically in some contexts, including a trial. Obviously, a high LR does not imply a high posterior probability, and a high posterior probability does not imply a high LR. However, the consequences of having only the one or only the other of these conditions are not symmetrical. If we have a high posterior for the hypothesis and a low LR,

Pr(h|e) high

Pr(e|h)/P(e|-h) = 1 or little more than one

then it follows algebraically that the prior probability of the hypothesis is high. That is, the high posterior for h is not *based* on e but on the opinions about h that you came in with. Because the prior probability of the hypothesis, Pr(h), does not figure in the LR, and because the evidence e does not figure in Pr(h), the LR provides a pure measure of the extent to which your posterior probability for h is due to the evidence e; for example, if the LR is 1 then the posterior probability owes nothing to e and everything to the prior probability.

Though a high LR does not by itself determine the posterior probability, from a given value of the LR we can calculate the posterior probability of h from its prior probability.[24] For example, suppose we have an ignorance prior for h, Pr(h) = .5. If so then a LR of only 19 is required to give us a posterior of .95. With a LR of 171 we can go from a prior of .1 to a posterior of .95. If you don't have any background knowledge, or don't know what unconditional prior probability you have a right to for h, and suspect it could be quite low, striving to amass more evidence

[23] This is assuming Smith is presented with roughly representative samples.

[24] Pr(h|e) = [LR(Pr(h)/Pr(-h)]/[1 + LR(Pr(h)/Pr(-h)], which is Pr(h|e) = odds(h|e)/(1 + odds(h|e)), with odds(h|e) = LR(odds(h)), and odds(h) = Pr(h)/(1- Pr(h)).

each of whose LR is greater than 1[25] is the only way to get to a high posterior probability. This is instructive for the context of court trials, where the mandate is that the verdict depend only on evidence presented at trial, and not on jurors' or judges' prior assumptions about the case. The LR for the total evidence in the trial would be a way to measure how far this has been achieved.

One's preference between the requirement of a high posterior and the requirement of a high likelihood ratio may depend on what question one is asking. If the question is which criterion marks better whether one's belief in h is *justified or not*, the answer may require a high posterior probability since it is plausibly a necessary condition for a right to believe h that the posterior probability of h be high. However, if the question is what is the best criterion for deciding whether a belief is based on good evidence, the answer can't be h having a high posterior probability, because the evidence one has conditionalized on to get to that may be wholly ineffectual by having LR = 1, in which case the high posterior probability is due to a high prior probability. But the question I began with is what, if anything, is epistemically wrong with convicting a prisoner or holding someone liable on the basis of a very high single base rate, and the answer of the classic and HCP Process Reliabilists must be that nothing is wrong – the belief is justified – while the answer of the Probabilistic Tracking view is that a base-rate test does not provide good evidence for a conclusion about a particular member of that population, because it has a false positive rate of 100%.

6. Knowledge

Several theorists mentioned at the beginning have argued that what is deficient about naked statistical evidence for trial verdicts is that it can't give us knowledge – even if it gives a justified belief or a legitimate degree of belief – and knowledge is needed for trial verdicts. (Thomson 1986, Littlejohn 2017, Moss 2018) But whether it can give us knowledge or not depends on what we mean by knowledge; we've seen above that that there is at least one perfectly respectable theory of what knowledge is that says an inference from a single base rate can give it to us – classic Process Reliabilism – and another that says it cannot – Probabilistic Tracking. Making an inference from a base rate of A's among B's that is > r to a belief that an individual who is a B is an A is a reliable process, as long as one has no defeating information and r is at or above the threshold for justified belief, because this process can be expected to yield a true belief on greater than r% of occasions of its use. If the belief so formed is true, then on the PR view it is knowledge. Contrarily, on the PT view, a subject who is using the base rate in this way would have a 100% chance of believing a prisoner guilty given that he is innocent, and thus his belief would not count as knowledge even if it was true.

[25] The LR of a conjunction of pieces of independent evidence is the product of their LR's. So, if the LR for each of e_1, e_2, …, e_n is > 1, then the LR of the total evidence rises exponentially with n. For example, if the LR of each piece of evidence is 2, then the LR of n pieces together is 2^n.

Thus, one could re-describe the problem about naked statistical evidence in trials as the question of what is required for knowledge. However that would require thinking that knowledge is required for a trial verdict, and we haven't seen here any reason to think that. My explanation for what is deficient about naked statistical evidence rests entirely on its quality as evidence, whose crucial role in trial verdicts doesn't need an argument. Usually knowledge is taken to require full belief,[26] but no full belief was required in our analysis of the question a juror faces about whether to render a guilty verdict. The juror may have a high degree of belief, .96, say, that the prisoner before them is guilty, and no full belief one way or the other. The question about whether this degree of belief is enough to find the defendant guilty, on the Probabilistic Tracking view, depends on the likelihoods using the evidence (diagnostic test result) presented at trial that got them that degree of belief. Two people may have the same high degree of belief in a hypothesis and very different likelihoods, because of differences in the quality of their evidence.

However there are a variety of naysayers of knowledge, and the concept of knowledge, who should not feel any comfort from my conclusion that knowledge is not the primary site of this problem, and should see some lessons. Like many who think that the concept of knowledge holds the key to the puzzle of naked statistical evidence, Papineau (2019) thinks that the reason people have the intuition that we should not convict or hold liable on such evidence alone is that a naked statistic (a single base rate) does not give a causal path to a claim about an individual, and so does not give knowledge. But he thinks the concept of knowledge is a stone-age tool that does much ill, and we should not let lack of it obscure expected utility considerations.

Echoing Enoch and Fisher, Papineau points out that since we are not perfect, unless we are willing to let all the guilty go free our system will convict some innocents. Put any disutility we like on that outcome relative to the outcomes of convicting and not convicting guilty people, and not convicting innocents, and there will be a p such that following a rule that says "Convict when the probability of guilt is greater than p (even if this is so by means of a single base rate)" will maximize expected utility. Call this a "High-p Rule". If there are cases where you don't convict though you have that probability of guilt, that is, where you don't follow this rule, it seems you will not be maximizing expected utility overall.

Our case of the prisoners provides an example. Suppose our High-p Rule takes the probability of guilt that is sufficient for conviction as .95. Suppose that the utility of convicting a guilty person is 1, the utility of letting a guilty person go free is -1, the utility of not convicting an innocent person is 0, and the utility of convicting an innocent is -20. Since for any of our prisoners the probability of guilt is .96, following the rule will have us convict all of them. The expected utility of following the rule in this case is:

[26] An exception is Moss (2018).

(p)(1) + (1-p)(-38) = (.96)(1) + (.04)(-20) = .16

Since we have no other evidence, not following the rule in this case means convicting no one. The expected utility of that is:

(p)(-1) + (1-p)(0) = (.96)(-1) + (.04)(0) = -.96

The expected utility of following the High-p Rule is higher than that of not following it. What else is there? asks Papineau.

There will always be possible errors, and attaching a great disutility to convicting innocents, which pushes up the value of p in our High p-Rule, is inscribing a safeguard against convicting innocents into our practice. Why is this not *sufficient* expression of concern for innocents? To answer this we must distinguish between two probability values involving the properties of innocence and conviction. One is the probability that someone is convicted, C, *and* innocent, -G, and the other the probability that a person is convicted *given that* they are innocent. The first is the probability of a false positive *result*, Pr(C.-G), and the second is the *rate* of false positives, Pr(C|-G), whose importance I have been stressing.

Probability of a False Positive/False Conviction: Pr(C.-G)
False Positive/False Conviction Rate: Pr(C|-G)

By definition, Pr(C.-G) = Pr(-G)Pr(C|-G). Using the single base rate to decide whether a prisoner is guilty in our example will give us both Pr(-G) = .04 and Pr(C|-G) = 1. Together these yield a very low probability of false convictions: Pr(C.-G) = .04. This, I suggest, is the source of some people's intuition that a High-p Rule is adequate for trials.

However, the fact that Pr(C|-G) = 1 means that our only insurance of this low probability of false convictions is the base rate together with our assumption that that is the base rate of our sample, not just the base rate of a population from which our sample was taken. Recall that if the sample is unrepresentative, then the fact that Pr(C|-G) = 1 means that the fraction of convicted innocents in our actual application of the rule could be anything from 0 to 1. In a court case our "sample" is the defendant or set of defendants on trial. If we have all 25 prisoners on trial then the base rate of the population used is also the base rate of the sample,[27] but if we have only one of the prisoners on trial, then all bets are off. If we use a single base rate for conviction, our actual proportion of false convictions is determined by whether the sample on trial is representative of the population whose base rate that was, and by nothing else. What is wrong with this picture is that the sample of people put on trial as

[27] Note that even in this best case where we have on trial the whole population the base rate comes from, the fact that Pr(C|-G) = 1 means that if we use the High-p Rule we are guaranteed to convict one innocent.

defendants in a criminal case is chosen by the prosecution. Thus, in accepting as a rule conviction on the basis of a single base rate we would be entrusting protection of innocent people entirely to the prosecution.

A natural objection to my focus on the false conviction rate is that it endorses risk aversion, and risk aversion is irrational. But it is obvious that risk aversion is irrational only if one is assuming that expected utility maximization is the only dimension of rational decision-making. If one does assume this, then two options with equal expected utility but different risk profiles are equally choiceworthy, so one should be indifferent between them. Experts today are rather less dismissive of the rationality of risk aversion than they once were,[28] and concern about a 100% false conviction rate might be better understood as loss aversion, since defendants stand only to lose the status quo or not, rather than to possibly win a prize. (Kahneman and Twersky 1979) But it is indisputable that human beings actually have a tendency to risk aversion towards possibilities with low probability but high negative consequence, which is understandable and often exhibits itself in public policy. So at the very least we may have identified a descriptively explanatory account of most people's reaction to naked statistical evidence in court verdicts.

However, risk aversion or loss aversion is not actually required to see the importance of the false conviction rate; its importance in court verdicts is clear, as explained just above, from the commitment of the justice system to protect the innocent from conviction via the procedures within the courtroom, and not only through the decisions of prosecutors and plaintiffs about who to bring to trial. Moreover, the expected utility calculation above for the High-p Rule was incomplete. In expected utility calculations about rules, such as the rule that a base rate is sufficient for conviction, we must take into account the probable consequences of adoption of the rules and not just the consequences of its use in a given case. In the calculation above we have taken account of the negative utility when a High-p Rule is applied to a given case where we do not have more specific evidence. But the adoption of this rule would lead to *more* such cases – cases with guaranteed false convictions – because it would leave prosecutors and plaintiffs with no incentive to look for more specific evidence in any case where they already had a base rate. If we accept that convicting innocents is of some disutility, then this factor would be a further weight against a High-p Rule, in the terms of expected utility alone.

Enemies and doubters of the importance of knowledge have created some unfortunate misimpressions in the discussion of naked statistics in court verdicts. So, Papineau says:

> I don't think the *rejection of statistical evidence* is a good thing. I recognize there is a widespread intuitive feeling that purely statistical convictions are improper. But I say that we should ignore this feeling and *allow the courts to use statistical evidence*. … [F]ew have been prepared to reject the *ban on purely statistical evidence*. (Papineau 2019, my emphasis)

[28] See, for example Hintze et al. (2015), and see Buchak (2013) for a risk-weighted expected utility theory defended on the basis of the general motivations behind classic expected utility theory.

Naked statistical evidence, defined above, is a single base rate, and the claim at issue is whether this can be sufficient for conviction or holding liable. To deny this is not to put a ban on purely statistical evidence, for its being insufficient does not imply that it is inadmissible. It is also not at all to reject statistical evidence, or disallow the courts to use statistical evidence; the vast majority of statistical evidence takes a more complex form than a single base rate, and can be sufficient for many types of ruling and verdict, even some requiring proof of causation, a fact that courts appreciate. Nothing I have said supports an objection to these practices.

Enoch and Fisher (2015) share Papineau's skepticism about letting the concept of knowledge guide the courts. They pose a thought experiment in which we choose which kind of legal system we want our children to live under, one that uses statistical evidence or one that uses only "individualized" evidence. They associate statistical evidence with accuracy and reliability and individualized evidence with knowledge, and ask how we could ever prefer a badge of epistemic "respectability" like knowledge over accuracy in trial verdicts. One could only prefer it if one had a "knowledge fetish". (Enoch and Fisher 2015, 579)

The choice presented here is deeply misleading. Again we are presented with a choice between a system that uses statistical evidence and a system that uses none, when the choice that gnaws at us is between being permitted or not to base a verdict on a single base rate alone. Beyond this, an externalist about knowledge will find the presented choice puzzling because for us it is inconceivable that knowledge would not itself *include* insurance of accuracy. Unlike purely internalist views of knowledge for which the required relation of a belief to the world is merely that that particular belief be true, externalist definitions of knowledge all have requirements of robustness in the relationship between the belief and the world; they have relationships that reliably yield accuracy over all or most cases built into them.

The classic Process Reliabilist and HCP Process Reliabilist require that the process used yield a high probability that the belief in p it leads to is true, or in other words that the process limit the probability of a false belief in p, which is one kind of accuracy one might be interested in. If a true belief with this property is what one thinks knowledge is, then one won't think that naked statistical evidence fails to deliver knowledge. One will take a single base rate as sufficient for conviction, but this is not to prefer accuracy over knowledge or vice versa. It is to prefer a certain epistemic property because it robustly, or reliably, insures the kind of accuracy one desires, and some people call true beliefs with that property "knowledge".

In addition to the desire to limit the probability of a belief that is false, one's concern with accuracy may include error rates. In this case one will prefer more accuracy than PR-type knowledge insures, and in so doing one will be preferring Probabilistic-Tracking type evidence. If a person with these preferences cares about the concept of knowledge, then their taste will be for the PT type of knowledge. This is not "to be willing to pay a price in accuracy in order to secure some epistemic respectability of the legal system". (Enoch and Fisher 2015, 579) It is to prefer knowledge, under a particular definition of this property, *because of* two particular kinds of accuracy that it insures, minimizing both false-positive and false-negative error rates. For an

externalist at least, the question Enoch and Fisher present us with is a false choice; for us, concern about knowledge *is* concern about accuracy.

Those at most risk of fetishism about knowledge are knowledge-first epistemologists who take the concept of knowledge as primitive. (Williamson 2001, Littlejohn, 2017). On their view it cannot be broken down into component properties, so we cannot draw out the consequences of such properties. If knowledge is a primitive, then one will have intuitions about it and its relations to other things, but one definitely won't have the kind of explanation given here of the epistemic weakness of single base rates in terms of the glaring error they make us vulnerable to; knowledge will have no sub-parts to derive consequences from.

7. Conclusion

Applying Process Reliabilism and Probabilistic Tracking to the problem of naked statistical evidence in trials brings into sharp focus a key difference between these externalist views of knowledge and related concepts. Comesana's High Conditional Probability version of PR shares with classic PR a focus on posterior probabilities and no conditions on likelihoods, and as a consequence both views have the implication that there is no principled difference between single base rates and other kinds of evidence about individuals, and no difference in the knowledge or justifiedness status of belief states or conclusions formed on their basis. A belief based on a high enough base rate is justified, and if true is knowledge. On the Probabilistic Tracking view of knowledge, forming a belief that a widget is OK or a prisoner guilty of assault on the basis of a single base rate cannot give one knowledge because the single base rate insures that the probability you will believe a widget is OK or a prisoner guilty given that they are not is 100%, when for knowledge this rate should have been less than a small number, say, 5%. Other kinds of claims – eyewitness testimony, forensics – typically don't have this 100% error rate, and in cases where we think they do we don't think we should base beliefs solely on them.

It is thus possible to describe the issue of whether naked statistical evidence is weak in terms of knowledge, or one's preference in concepts of knowledge. However this is not necessary, and my explanation of the weakness of this kind of evidence for trials resides in the conditions for good evidence, which in turn resides in the likelihoods for the test producing the evidence for one's conclusion. Conditions on these minimize the rates of false positive errors and false negative errors. Pragmatic factors relevant to the content of a given decision determine how much relative disutility these errors have for a person, but it is a rare person who is completely and always immune to a concern about false positive error.

Defective can openers tend to be annoying rather than dangerous, but the can-opener manufacturer may soon be out of business if all of the openers she puts on the market are

defective. She is concerned with false positive results, the probability that an opener is defective given that she put it on the market, and this can be addressed directly by insuring that the posterior probability that an opener is OK given that her staff approved it is high. The higher this probability, the lower the probability of a can opener that is defective and approved. Raising the posterior probability that an opener is OK is a job that a single base rate can do; if the base rate is correct for the sample then the higher it is the lower the number and fraction of results that are approved and defective.

Thus it is not that high posterior probabilities and high base rates do nothing to prevent false positive results, but they do nothing to constrain the false positive rate, Pr(e|-h). A high posterior probability for h is compatible with any value for the likelihoods, thus with any measure of the quality of evidence e. If one is using a base-rate test to decide whether h, the test will have Pr(e|-h) = 1, and if the sample one is judging is a subset of the population whose base rate one is using, the number and fraction of false positive outcomes one can expect is highly sensitive to the representativeness of this sample. In the worst-case unrepresentative sample, h is false for every member of the sample, so one will be mistaken in 100% of one's conclusions. Expected utility calculations about false positives concern the expected results, not rates, so they also use unconditional or posterior probabilities and do not take likelihoods into account. Thus, as we saw, two people can assign the same disutilities as each other to false positives and false negatives, and have the same posterior probabilities, like Smith and New New Jones, yet have completely different risk profiles for false positives.

The application of these points to evidence in trials is clear. The fact that 24 of the 25 prisoners assaulted the guard makes the probability that a given prisoner assaulted the guard .96, for every prisoner. A High-p Rule of .95 justifies us in finding every member of this group guilty. However to use this evidence to decide on a verdict is to use a procedure with Pr(e|-h) = 1. If the sample on trial is all 25 prisoners, then the sample one draws conclusions about is the same as the population whose base rate is used in the base-rate test, so one does not risk a great number of false convictions. Yet one is guaranteed to convict one innocent person. A system that sanctions such a decision rule permits not just false convictions but guaranteed false convictions.

That was the best case scenario. For the worst case scenario consider an adaptation of L. Jonathan Cohen's famous Gatecrasher case. (Cohen 1977, 74) There are 100 people attending a concert but the automatic counter says only 40 had tickets. Those who had and those who didn't are now indistinguishable because those who had tickets had them taken by the automated turnstile on the way in. (Those who didn't have tickets jumped the turnstile.) Suppose the organizers managed to grab 45 of the 100 attendees as all exited the concert hall, and sued these 45 for the admission money. Suppose the standard of proof for civil trials – more likely than not – can be expressed as requiring that the posterior probability that the defendant is guilty be greater than 50%.

For each of these 45 defendants the probability they didn't have a ticket is .6, so by a High-p Rule of >.5 every one of them should be found liable. However, the probability that this base-rate test says a person is guilty given that they're innocent is 1. Now, it could be that the 45 that have been picked up are all members of the set of 60 who had no ticket. However, for all we know as many as 40 of those 45 people did have tickets. Thus if the system is such that the plaintiff can get a judgment using this High-p Rule, then a trier of fact risks having 89% of the judgments in the case be false. (If only 40 or less of the concert-goers had been rounded up, then there would be a risk of 100% erroneous judgments, and as far as we know anything from 0% errors to 100% errors are equally likely outcomes.) It is not only that licensing a procedure with such high rates of potential error should give us pause, but that, using this rule, what determines whether an innocent defendant gets a judgment against them is pure happenstance surrounding events outside the trial, the events that determined whether they would be defendants. In a criminal case, this means that whether an innocent person will be found guilty in the trial rests on decisions made by the prosecution.

Cohen thought that problems like that posed by the Gatecrasher case showed that mathematical probability is ill-suited for analysis of legal evidential standards. I think that probability is well-suited to explain the problem with relying on base rates alone for a verdict, if we understand and distinguish between posterior probabilities and likelihoods.

References

Adler, Jonathan (2005) "Reliabilist Justification (or Knowledge) as a Good Truth Ratio", *Pacific Philosophical Quarterly* 86: 445-458.

Asasi, Kamyar (2019), "In Defense of Statistical Evidence: From Epistemology to Courtrooms", MA Thesis, Central European University.

Buchak, Lara (2013). *Risk and Rationality*. Oxford: Oxford University Press.

---------- (2014), "Belief, Credence, and Norms", *Philosophical Studies* 169: 285-311. DOI 10.1007/s11098-013-0182-y

Christensen, David (1999), "Measuring Confirmation", *The Journal of Philosophy* 96 (9): 437-461.

Cohen, L. Jonathan (1977). *The Probable and the Provable*. Oxford: Clarendon Press.

Comesaña, Juan (2009), "What Lottery Problem for Reliabilism?" *Pacific Philosophical Quarterly* 90 (2009) 1–20.

Enoch, David, Levi Spectre and Talia Fisher (2012), "Statistical Evidence, Sensitivity, and the Legal Value of Knowledge", *Philosophy & Public Affairs* 40 (3): 197-224.

Enoch, David and Talia Fischer (2015), "Sense and 'Sensitivity': Epistemic and Instrumental Approaches to Statistical Evidence", *Stanford Law Review* 67 (557): 557-611.

Enoch, David and Levi Spectre (2019), "Sensivity, Safety, and the Law: A Reply to Pardo", *Legal Theory* 25: 178–199.

Fitelson, Branden (1999), "The Plurality of Bayesian Measures of Confirmation and the Problem of Measure Sensitivity", *Philosophy of Science* 66: S362-S378.

Goldman, Alvin (2012), "What is Justified Belief?" in Ernest Sosa, Jaegwon Kim, Jeremy Fantl, and Matthew McGrath eds., *Epistemology: An Anthology*. Second Edition. Malden, MA: Blackwell Publishing Ltd, 333-347.

Good, IJ (1983). *Good Thinking*: *The Foundations of Probability and its Applications*. Minneapolis: University of Minnesota Press.

---------- (1985), "Weight of Evidence: A Brief Survey", in JM Bernardo et al. eds. *Bayesian Statistics 2*. North Holland: Elsevier Science, 249-270.

Guenther v. Armstrong Rubber Co., 406 F.2d I3I5, I3I8 (3d Cir. 1969) (dictum).

Hintze, Arend and Randal S. Olson, Christoph Adami and Ralph Hertwig (2015), "Risk sensitivity as an evolutionary adaptation", *Scientific Reports* 5, Article number: 8242.

Kahneman, D. & Tversky, A. (1979). "Prospect Theory: An Analysis of Decision under Risk", *Econometrica* 47 (4): 263–291.

Levi, Isaac (1977), "Direct Inference", *The Journal of Philosophy* 74 (1): 5-29.

Littlejohn, Clayton (2017), "Truth, knowledge, and the standard of proof in criminal law", *Synthese* DOI 10.1007/s11229-017-1608-4

Moss, Sarah (2018), *Probabilistic Knowledge*. Oxford: Oxford University Press.

Nozick, Robert (1981). *Philosophical Explanations*. Cambridge, MA: Belknap Harvard Press.

Papineau, David (2019), "The Disvalue of Knowledge", *Synthese* https://doi.org/10.1007/s11229-019-02405-4

Pritchard, Duncan (2018), "Legal risk, legal evidence and the arithmetic of criminal justice", *Jurisprudence*, 9(1): 108-119, DOI: 10.1080/20403313.2017.1352323

Roush, Sherrilyn (2005). *Tracking Truth: Knowledge, Evidence, and Science*. Oxford: Oxford University Press.

Sargent vs. Massachusetts Accident Co., 307 Mass. 246, 250, 29 N.E.2d 825, 827 (I940) (dictum).

Schauer, Fredrerick (2003). *Profiles, Probabilities, and Stereotypes*. Cambridge, MA: Belknap Harvard Press.

Shaviro, Daniel (1989), "Statistical-Probability Evidence and the Appearance of Justice", *Harvard Law Review* 103 (2): 530-554.

Smith, Martin (2018), "When does evidence suffice for conviction?" *Mind* 127 (508): 1193-1218. doi:10.1093/mind/fzx026

Smith vs. Rapid Transit, 3I7 Mass. 469, 58 N.E.2d 754 (1945).

Sosa, Ernest (1999), "How to Defeat Opposition to Moore", *Philosophical Perspectives* 13, Epistemology, 141-153.

Thomson, Judith Jarvis (1986), "Liability and Individualized Evidence", *Law and Contemporary Problems* 49: 199-219.

Williamson, Timothy (2002). *Knowledge and its Limits*. Oxford: Oxford University Press. DOI:10.1093/019925656X.003.0010

Zalabardo, José (2009), "An argument for the likelihood-ratio measure of confirmation"

Analysis 69: 630-635.

---------- (2017), "Safety, sensitivity and differential support", *Synthese* https://doi.org/10.1007/s11229-017-1645-z