Adaptive Burdens of Proof

Boris Babic*

California Institute of Technology

Abstract

Disputes involving the use of statistical evidence in civil litigation are often taken to suggest that in legal fact finding the demands of morality can come apart from the requirements of epistemic rationality. In this paper I show that this is not necessarily the case. More constructively, I develop a model of the burden of proof as an adaptive Bayesian hypothesis test in order to better understand how statistical information affects a judge or jury's decision behavior. The implication of the model is that decision makers who apparently ignore statistical evidence are not epistemically irrational – they are risk averse.

1 Introduction

Among the many apparent problems with statistical evidence in legal fact finding is that in both real and hypothetical disputes judges and juries appear to ignore available and relevant base rates in order to reach verdicts they consider to be morally appropriate.[1] This is a judgment that many legal commentators endorse even upon reflection.[2] As a result, the demands of morality seem to be incompatible with the requirements of epistemic rationality. To act rightly as a legal fact finder one may have to believe irrationally. This is, for example, the implication of Nesson (1986)'s argument. I develop a model of the burden of proof which implies that a decision maker may avoid apparently morally inappropriate decisions without ignoring base rates provided she is risk averse.

In addition, the model I propose can explain *both* why so-called taboo or forbidden base rates of the sort discussed by Tetlock et al. (2000) are often inadmissible, even if accurate, *and* why DNA random match profiles are relatively uncontroversial. In this sense, the model is significantly more robust to changes in the nature of the statistical evidence in issue as compared to its alternatives.[3] Indeed, while I am most concerned with civil disputes, the approach is equally effective in the criminal context.[4]

The model is very simple: from the decision maker's perspective, the plaintiff has satisfied her burden of persuasion with respect to an element of the prima facie case if the posterior odds exceed a threshold determined by a ratio of the decision maker's error costs. Developing this carefully will take some work, but that's it. The model is *adaptive* because the error parameters are not determined in advance. The model is risk eliciting because the decision maker's choice *reflects* her underlying

*Postdoctoral Scholar, California Institute of Technology. JD, Harvard Law School; PhD in Philosophy and MS in Statistics, University of Michigan, Ann Arbor. Thanks to Paulo Barrozo, Gordon Belot, Kevin Blackwell, Edward Cheng, I. Glenn Cohen, Daniel Drucker, Rich Gonzalez, Scott Hershovitz, Zoë Johnson-King, James Joyce, Louis Kaplow, Peter Railton, Brian Weatherson, and audiences at Harvard Law School and the University of Michigan. Research for this project was supported by the Social Sciences and Humanities Research Council of Canada.

[1] See e.g., Wells (1992).

[2] See e.g., Tribe (1971), Nesson (1985), Wasserman (1991), and for a more general overview Colyvan et al. (2001) and Schauer (2003).

[3] Competing models are developed in Posner (1999), Kaplow (2014), Cheng (2013), and Cheng & Pardo (2015), among others.

[4] This is because the adaptive model has a flexible threshold that can take any value on the unit interval. That threshold is determined by the agent's tolerance to risk of error. Some values, of course, will be obviously morally inappropriate.

attitudes to risk of error. What that attitude ought to be will be context sensitive and determined in part by the factual circumstances of the relevant dispute. The adaptive model is especially helpful for understanding mass exposure cases, pharmaceutical class actions, and complex business litigation, where statistical evidence is often unavoidable.[5]

This approach makes several empirically verifiable predictions. If I am correct, then we should expect to see a strong correlation between a decision maker's sensitivity to risk of error – which may be elicited by presenting her with a sequence of increasingly risky epistemic prospects, as I suggest in Babic (2017) – and her aversion to statistical evidence in various hypothetical scenarios, such as those presented to the subjects in Wells (1992).[6] Moreover, we may apply the adaptive model for the normative assessment of legal decisions, by attending to the values to risk they elicit and considering their reasonableness. Finally, the model may be used as a tool for predicting the resolution of future disputes (or, more specifically, the admissibility of statistical evidence in such disputes).

The paper proceeds as follows. First, I explain the relevant formal concepts, showing how the likelihood ratio test and Bayesian hypothesis test are both related to the odds-likelihood expression of Bayes' Theorem (§2). Then, I explain what is typically taken to be the problem of statistical evidence, situating it in the context of the treatment of probabilities both in the case law and under the Federal Rules of Evidence (§§3.1). Next, I give a genealogical presentation of the so-called blue bus puzzle (§3.2). As we will see, what we call a paradox is essentially the same problem that Kahneman & Tversky (1972) and Bar-Hillel (1980) used to illustrate the base rate fallacy. The relationship between their presentation of the problem and the life it has taken on in legal scholarship has been significantly under appreciated.

In §4.1, I introduce a trichotomy that the statistician Richard Royall draws for making sense of statistical inference. Royall distinguishes three separate questions we might ask after making a set of observations: (Q1) what should we believe?, (Q2) what does the evidence say?, and (Q3) what should we do? I argue that the reason statistical evidence cases can appear paradoxical is because we have so far modeled burdens of proof as answering Royall's first or second questions, when we should be trying to answer his third question. Indeed, it is (Q3) that even the classic Neyman & Pearson (1933) null hypothesis significance testing procedures are designed to answer. Once the focus is on (Q3) it becomes clear that sensitivity to risk of error will be the key ingredient in constructing a decision procedure for legal choice. As a result, in §§4.2-4.3, I extend a theorem initially developed by DeGroot & Schervish (2012)[7] to show that if our goal is to minimize a linear combination of false positive and false negative error rates, we can do no better than to apply a Bayesian hypothesis test. This is the mathematical justification for using the adaptive model in legal fact finding.

In §5, I put the adaptive model to work. First, I explain how it can handle the usual apparent paradoxes of statistical evidence and compare its performance to Cheng (2013)'s likelihood ratio test (§§5.1-5.3). In §5.4, I evaluate the case law on statistical evidence to show how well the adaptive model predicts the data we have and to suggest how easily it could be used to predict the admissibility of statistical evidence in future disputes (§5.5). Koehler (2002), for example, develops a four-fold taxonomy for when statistical evidence is likely to be admissible. The adaptive model is much more efficient in its forecasting: all we need to do is (a) consider the decision maker's sensitivity to epistemic risk in light of (b) the factual circumstances in issue. It enables us to make very specific predictions when the information available to us would justify such specificity while at the same time making it possible to put some bounds on our estimates when data is sparse.

In §6, I consider several concerns and objections. First, I distinguish Kaplow's welfare-based

[5] See Rosenberg (1984) for helpful examples. See also *In re Agent Orange Prod. Liab. Litig.*, 597 F. Supp. 740, 835-836 (E.D.N.Y. 1984), for a discussion of the inevitability of statistical evidence, and the need for a model of the preponderance standard that accommodates it, in mass exposure litigation.

[6] There is some empirical support of a related relationship in the context of loss aversion and its effect on interpretations of the burden of proof (Ritov & Zamir, 2012).

[7] The theorem seems to have been first developed by Morris DeGroot in the 2d (1986) edition of the text.

interpretation of the rejection threshold from the adaptive model's more general interpretation of the costs of error (§6.1). As we will see, Kaplow (2014)'s approach is a special case of the adaptive model. Second, I explain the difference between a model that elicits the decision maker's attitudes and a choice rule (§6.2). Finally, I clarify how the adaptive model fits within the more general subjective expected utility optimization approach to decision making by situating it in what I call a principal-agent choice environment (§6.3). By way of conclusion, I make some connections between the adaptive model and evidence proportional theories of recovery as applied to, for example, DES manufacturers.[8]

2 Modeling Burdens of Proof

There is a substantial literature in economic and statistical analyses of evidence law on modeling burdens of proof.[9] Or, to be specific, the burden of persuasion.[10] I am interested in how these models handle statistical evidence and in particular the apparent paradoxes generated by sensitivity to base rates. From that perspective, we can roughly divide existing models in two families: welfare-based and more generally economic approaches and accuracy-first approaches. Both families are decision theoretic but they vary across several important dimensions.

2.1 Economic vs. accuracy approaches

First, the welfare approach assumes that the *only* consideration in setting the burden of proof should be its effect on social welfare, where social welfare is a function exclusively of the utilities of the relevant individual decision makers. Kaplow (2011)'s model is paradigmatic.[11] For Kaplow, proportionality, autonomy, or retributive punishment, for example, are not considered unless we have a preference for living in, say, a legal system which imposes punishments that approximately fit the wrong or crime.[12] Other economic approaches are more general and consider costs that, strictly speaking, may not be reducible to their effects on individual utilities.[13] On the accuracy-first approach, correctness of verdicts is the overarching consideration.[14] Since accuracy is the focus of these models, false positive (Type I) and false negative (Type II) error rates tend to play a dominant role in setting the optimal burden of proof. There are no uniform constraints on the costs of each type of error. The costs could be effects on individual utilities, but they need not be.

Second, in most welfare and more generally economic models, behavior is assumed to be endogenous.[15] As a result, as Kaplow (2012) puts it, the optimal threshold is determined by asking "how behavior will change as a function of a change in the evidence threshold?" (378). The relevant perspective, therefore, is said to be ex ante because we are interested in how setting a particular threshold will affect harmful and beneficial behavior. A low evidence threshold deters harmful behavior (like anticompetitive business practices, for example), but it also chills innocuous or beneficial behavior (such as entering into mutually beneficial agreements). In accuracy models, meanwhile, we take behavior as given and find the rule that performs best with respect to some tolerable error rate.

[8] *Sindell v. Abbott Labs.*, 26 Cal. 3d 588 (1980) (developing the notion of market share liability). See Rosenberg (1984) for a general defense of proportional liability.

[9] One of the earlier articles to take a decision theoretic approach to legal fact finding is the now classic Kaplan (1968).

[10] For models that look at the burden of production instead, see e.g., Hay & Spier (1997).

[11] See also Kaplow (2012).

[12] See Kaplow & Shavell (2001), arguing that any non-welfarist approach of assessing gains and losses may violate the Pareto principle, which implies that it could require deeming socially superior outcomes under which all are worse off.

[13] See e.g., Miceli (1990) (considering the value of retribution and proportionality). Miceli builds proportionality into the utility function. Kaplow & Shavell (2001) could do this too, but they only consider it in the discussion following their model, as one among several ways in which their approach could be relaxed. I suspect they do not take this possibility too seriously, though, since their primary aim in Kaplow & Shavell (2006) is to argue against non-consequentialist approaches to the assessment of legal standards.

[14] See e.g., Cheng (2013) and Cheng & Pardo (2015); Cf. Kaplow (1994).

[15] That is, behavior changes in response to changes in the values of the parameters in the burden of proof model. But see Rubinfeld & Sappington (1987) for an economic model of expected social losses from errors in adjudication that takes behavior to be exogenous, focusing instead on the relationship between the litigation effort of defendants and the judge's ultimate assessment of their guilt.

The analysis is said to be ex post or backward looking because the action already took place and our goal is to try and avoid either error and make a correct decision.

Third, the models may be fixed or variable. Cheng & Pardo (2015) develop a fixed standard of proof that applies uniformly to all cases within its scope. Meanwhile, economic models tend to be flexible and vary from case to case. This is to be expected since different cases will have different effects on subsequent behavior. The fixed standard, however, is not required by any specific element of statistical decision theory. Rather, it is a product of Cheng and Pardo's philosophical commitments – that it would be unfair to shift the burden of proof from case to case – and their political forecast – that a shifting burden would lead to charges of political manipulation and illegitimacy within the legal system.

Fourth, and perhaps most importantly, in accuracy models either prior probabilities tend to be set aside for normative reasons, as in Cheng (2013), or Type I and Type II errors are assumed to be equally bad, which is then used as an argument to set aside prior probabilities, as in Cheng & Pardo (2015). In either case, the result is the same – prior probabilities are deemed irrelevant in many legal decision making contexts. This is especially unfortunate given that the models take accuracy as their primary consideration and ignoring priors can and often does, as we will see below, lead to inaccurate verdicts in both real and hypothetical decisions.

Fifth, and finally, in both economic and statistical models of the burden of proof, the model is put forward as a decision rule.[16] In other words, the model is supposed to be action-guiding: to the extent you are persuaded by the approach, you ought to believe that we should reform the legal system accordingly. Kaplow (2012) asks, for example, "how could the burden of proof be reformulated to attend more explicitly to welfare considerations?"[17] But a model of the burden of proof can be normative without being action guiding. While it is true that we often construct decision models as a guide to judgment and decision making, it is equally true that we often construct models in order to better understand how people behave in a particular domain – in this case, it is the domain of legal decision making. But that does not mean we would endorse the model as a decision rule.[18]

In particular, we may use the model as a framing tool in order to elicit particular norms or values underlying choice behavior. This is in the spirit of Ramsey (1926), Savage (1971) and De Finetti (1937)'s elicitation models for subjective probabilities. That is, the model can help us extract clues that drive behavior of interest to us. But whereas Ramsey, Savage and DeFinetti were interested in eliciting strengths of belief – often holding attitudes to risk constant – I will be interested in eliciting attitudes to risk – and will hold dynamic probabilistic coherence constant. This is the purpose for which I propose the adaptive model and the most significant way in which it differs from both economic and accuracy approaches, as they have been articulated in the literature.

Table (1), below, summarizes the salient dimensions along which welfare and more generally economic approaches may be compared with accuracy approaches.

[16] To be clear, it is not entirely clear where Cheng stands on this point. In Cheng (2013) the model is clearly descriptive because, by his own admission, it constitutes a wrongheaded approach to inference (1267, n. 24). In Cheng & Pardo (2015), he criticizes Kaplow (2012) for proposing a rule that is difficult to apply in practice.

[17] Meanwhile, Cheng (2013) takes himself to vindicate the current preponderance standard because it is implied by the accuracy approach. Demougin & Fluet (2008) reach a similar conclusion on the basis of economic efficiency.

[18] Because, for example, the ideal decision rule may be exceedingly difficult to apply and approximating it can be suboptimal, as suggested by results like Lipsey & Lancaster (1956)'s general theory of second best.

	Welfare models	Accuracy models
Flexibility:	Variable	Fixed
Priors:	Relevant	Irrelevant
Perspective:	Ex ante	Ex post
Normative role:	Decision rule	Decision rule

Table 1: Modeling Burdens of Proof

I will develop a model of legal decision making with a shifting burden of proof, for an accuracy-first theorist, that is sensitive to prior probabilities, as well as the costs and benefits of a legal decision, which may well occur as a result of the decision itself. In other words, I do not assume that behavior is exogenous nor do I focus exclusively on effects that are reducible to social welfare, since I am interested in developing a model that can help us better understand why people decide the way they do. For the evaluative model I seek to develop, therefore, the ex ante/ ex post distinction is a false dichotomy. Instead, I ask, in light of the model, what kind of risk profile would vindicate the decision maker's choice, regardless of what she took the relevant costs to be?[19] By paying attention to that profile, we gain insights into the rationality of her decision. The elicited risk attitude is the important part.

It is important because it can help us understand how legal decision makers reach verdicts (in actual or hypothetical cases) that seem at odds with available base rates without assuming that they ignore them. In other words, in the familiar paradoxical cases of statistical evidence the adaptive model I develop shows that *we can be both moral and epistemically rational provided we are risk averse.* By 'moral' I simply mean, very roughly for now, we can avoid conclusions in statistical evidence cases that most people consider to be inappropriate. Meanwhile, I take epistemic rationality to require probabilistic coherence (i.e., conformity of an agent's subjective degrees of belief to the Kolmogorov axioms) and updating by Bayesian conditioning (which I often refer to as dynamic coherence).

2.2 The burden of proof as a hypothesis test

In this section, I develop a general decision theoretic expression of the burden of proof and explain its relationship to Bayes' Theorem. This section is intended in part as a directed introduction to the formalism I rely on in the course of the argument to follow. The important concepts will be conditional probability, Bayesian updating, prior and posterior odds, the likelihood ratio test and, importantly, the odds-likelihood expression of Bayes' Theorem.

Let $L(\mathbf{X}|H)$ represent the likelihood of seeing evidence $\mathbf{X}$ admitted at trial on the assumption that hypothesis H is true. $\mathbf{X}$ is a vector of random variables $\langle X_1, ..., X_n \rangle$ representing a string of information such as, for example, witness testimony, e-mail correspondence, and a manufacturing record. For our purposes, each X_i is drawn from a discrete binary distribution. The corresponding lowercase vector $\mathbf{x} = \langle x_1, ..., x_n \rangle$ represents the realized values of the random variables. So, for example, we might let $X = 1$ if a witness identifies the defendant company as the manufacturer of an allegedly harmful prescription drug and $X = 0$ otherwise. H is some statement about a contested element of the *prima facie* case. More specifically, H is a statement about the value of an unknown parameter of interest, θ. So, for example, we might have two hypotheses, $H_0 : \theta = 0$, standing for the claim that the defendant company did not manufacture a drug whose origin is in dispute (our 'null' hypothesis), and $H_1 : \theta = 1$, standing for the claim that the defendant company did manufacture the drug (the alternative hypothesis). Our parameter space is then $\Omega = \{0, 1\}$ and $\theta \in \Omega$.

In the expression $L(\mathbf{X}|H)$, the vertical bar simply indicates that the likelihood is parameterized

[19] We can also set aside the feasibility debate. Cheng's objection to Kaplow's model is that it would be too difficult to execute. But again, since the model I will develop is a framing device to help us understand legal choice behavior, feasibility is orthogonal. I will not argue that we should, for example, modify jury instructions in a way that fits the adaptive model.

by the hypothesis H. It is the likelihood of observing $\mathbf{X}$ on the assumption that H is true. The difference between the probability distribution and the likelihood is in the argument of the function. When we talk about likelihood, we are interested in how plausible the data is under some hypothesis as a way of learning something about the plausibility of that hypothesis. For example, suppose we have tossed a coin of unknown bias ten times and it came up heads six times (data). We may want to consider which degree of bias (parameter) would make this result most plausible. As a result, the likelihood is thought of as a function of the parameter. Meanwhile, the distribution function is a function of (often not yet generated) data. For example, we want to know how probable it is that if a fair coin (parameter) is tossed ten times, it will land on six heads (data).[20]

Similarly, let $L(\mathbf{X}|\overline{H})$ stand for the likelihood of seeing the evidence admitted at trial on the assumption that $\overline{H}$ (read 'not H') is true. We will typically assume that our two hypotheses H and $\overline{H}$ partition the parameter space Ω in the context of legal fact finding, so that one or the other must be true.[21] A likelihood ratio test is a test that does not reject our legal 'null' hypothesis H just in case the likelihood ratio exceeds some threshold k. That is, we will not reject H if,

$$\frac{L(\mathbf{X}|H)}{L(\mathbf{X}|\overline{H})} > k \tag{1}$$

For example, we may think that in order to accept the Plaintiff's claim, H should be twice as likely as $\overline{H}$. In other words, $k = 2$. The higher the likelihood ratio the more frequently that $\mathbf{X}$ would be generated when the true state is H rather than $\overline{H}$. For a preview of the cases we will consider, we may have, for example, $\mathbf{x} = \langle x_1, x_2 \rangle$ corresponding to two witnesses each identifying a bus as blue where the bus collided with plaintiff's car and the ownership of the bus is in dispute. It is of course more plausible to think that such evidence is more likely to be generated if the bus were indeed blue than if it were, say, green. But if our data consists instead of each witness identifying a bus, without indicating its color, then the likelihood of such evidence would be insensitive to whether the bus was indeed blue rather than green. Either hypothesis about bus color seems equally likely to generate such testimony.

Now suppose that ninety nine percent of buses in town are green. But two witnesses identify a blue bus. It seems reasonable to consider the frequency with which such testimony would be generated in light of the extreme paucity of blue buses in town. Given these assumptions it seems not implausible to consider, say, witness tampering as an alternative explanation. But our likelihood ratio test, as stated in (1), is not yet sensitive to such 'prior' data.

If we are interested in accurate verdicts in the legal process we need to evaluate the likelihood ratio in light of our prior estimate of the respective probabilities of the parties' claims. Let $P(H)$ and $P(\overline{H})$ represent the prior probabilities of each hypothesis. A test that is not blind to background

[20] We do not call the likelihood a probability because as a function of the parameter it may not sum or integrate to 1. If we appropriately re-scale the likelihood over the parameter space, then it will give us the probability of the hypothesis in interest.

[21] Whether or not to make this assumption is a hard question. On the one hand, a well designed hypothesis test should partition the parameter space (Lehmann & Romano, 2005). To see why, consider two hypotheses: either the moon landing was staged or it was broadcast by giraffes from outer space. Suppose our evidence better supports the first alternative (as it surely would) so that we find the support statistically significant and thereby reject the space giraffe hypothesis. How much does this really tell us about whether the moon landing was actually staged? Not much. I find this sufficiently problematic so I have decided to partition the parameter space. Kaplow (2014) takes the same approach. Cheng (2013), however, decides not to partition the parameter space so as to be more faithful to the way litigation practice usually proceeds – namely, by considering the plaintiff's narrative of the events against the defendant's, which may not be mutually exhaustive. After all, we usually require the defendant to put forward an alternative theory of the case, rather than issuing a blanket denial of the plaintiff's allegations. As a result, however, Cheng is forced to make some *ad hoc* assumptions about the model's applicability – for example, it may be that it only becomes relevant after the plaintiff has survived a motion for summary judgment, after which point it is more likely that her hypothesis is at least somewhat plausible.

information would look roughly like this: we will not reject H if,

$$\frac{P(H)}{P(\overline{H})}\frac{L(\mathbf{X}|H)}{L(\mathbf{X}|\overline{H})} > k \tag{2}$$

What we have done here is discount each likelihood by its respective prior probability. This seems plausible, as a way of interpreting likelihood in light of what we know about the hypotheses to begin with. Our test is now closely related to Bayes' Theorem. The Theorem states that the posterior odds are equal to the prior odds times the likelihood ratio. That is,

$$\frac{P(H|\mathbf{X})}{P(\overline{H}|\mathbf{X})} = \frac{P(H)}{P(\overline{H})}\frac{L(\mathbf{X}|H)}{L(\mathbf{X}|\overline{H})} \tag{3}$$

This is what is sometimes called the *odds-likelihood* expression of Bayes' Theorem. In general, when the posterior odds of an event are $n : m$ the probability of that event is $n/(n+m)$. So if we know that the posterior odds of H are $2 : 1$, for example, we can infer that the posterior probability of H is $2/3$. This is just Bayes' Theorem differently expressed. But this expression is helpful to us for two reasons.

First, since the left hand side in (2) is just the posterior odds, factored into priors and a likelihood ratio, it makes very explicit the *pull* that the priors have on the evidence. They hold us back from jumping to conclusions. Second, what we have in (2) is a Bayesian hypothesis test, which consists of three principal components: prior odds, likelihood ratio, and a rejection threshold. This is the general statement of a hypothesis testing procedure as applied to legal burdens of proof. Everyone in the literature agrees that a legal hypothesis test should have something like this form. Where we disagree is on which terms should be fixed, and which should vary, as well as their proper interpretation. The first part of my argument is now easy to state: *all the terms should vary.* What remains to be seen is why they should vary and how we should interpret them. But before we get there let us see why I am interested in modeling burdens of proof – namely, because of the so-called paradox of statistical evidence.

3 Naked Statistics: The Phantom Menace

3.1 Probability and the rules of evidence

Fed. R. Ev. 401 is the starting point for determining the admissibility of evidence in the federal courts. It states that evidence is relevant if it has "any tendency to make a fact more or less probable than it would be without the evidence." The definition of relevance, then, is explicitly probabilistic. More than that, it incorporates each of the concepts discussed above – prior probability, posterior or conditional probability, and comparative likelihood – in order to define evidence explicitly in terms of incremental changes in probability. To see why this is the case, notice that according to the definition, X is relevant if $P(H|X)/P(H) > 1$. And we know from (3) that $P(H|X)/P(H)$ is equal to $L(H)/L(\overline{H})$.[22] The likelihood ratio is also known as the Bayes factor, precisely because it is a measure of incremental change in probability. It is the term that, multiplied by the prior, gives us the posterior.

Indeed, courts now generally recognize, as Judge Posner says, that "since all evidence is probabilistic – there are no metaphysical certainties – evidence should not be excluded merely because its accuracy can be expressed in explicitly probabilistic terms, as in the case of fingerprint and DNA evidence" (Posner, 1999, at 1508). Indeed, in *Branion v. Gramly*, 855 F.2d 1256, 1263-64 (7th Cir.

[22]The comments to the rule make clear that the probabilistic language is intended: "The rule summarizes [relevance] as a 'tendency to make the existence' of the fact to be proved 'more probable or less probable.' Compare Uniform Rule 1(2) which states the crux of relevancy as 'a tendency in reason,' thus perhaps emphasizing unduly the logical process and ignoring the need to draw upon experience or science to validate the general principle upon which relevancy in a particular situation depends." Notes of Advisory Committee on Proposed Rules.

1988) the court notes that "[a]fter all, even eyewitnesses are testifying only to probabilities (though they obscure the methods by which they generate those probabilities) – often rather lower probabilities than statistical work insists on" (internal citations omitted).[23] Even in criminal cases, where the state has to establish guilt beyond a reasonable doubt, courts realize that probabilities are inevitable. In *Victor v. Nebraska*, 511 U.S. 1, 14 (1994), for example, the Supreme Court found that "the beyond a reasonable doubt standard is itself probabilistic."[24]

So, then, how much disagreement could there be about the use of prior probabilities in legal fact finding? A lot, it turns out, mostly revolving around the so-called paradox of 'naked' statistical evidence. The purported paradox arises in connection with statistical evidence of identity in civil litigation, especially in negligence torts. The apparent puzzle presents situations where it seems both appropriate to have a high posterior probability in the defendant's guilt and inappropriate to hold the defendant legally responsible on the basis of the evidence that justifies that probability.[25] In other words, you should believe that the defendant is liable and, at the same time, that it would be morally inappropriate to hold her liable. The task, then, becomes one of attempting to reconcile these apparently conflicting judgments.

One way out of the dilemma is to deny that one's posterior probability should indeed be high. But no one disputes the likelihood – i.e., no one has argued that we should, say, ignore direct witness testimony. As a result, the way to bring down the posterior is by arguing that we ignore the prior. Of course we cannot simply avoid it.[26] Instead, what advocates of this position do, explicitly or otherwise, is set the prior odds to 1 despite evidence of their inequality.[27]

Another common approach to the dilemma is to deny that the posterior probability itself is relevant. The task then becomes one of identifying the appropriate alternative epistemic attitude.[28] This solution is more common in the philosophy literature since, as we saw, the Federal Rules explicitly define evidence in terms of incremental changes in probability. An alternative solution is to suggest that probability in the legal context just means something altogether different from

[23] See also Rosenberg (1984)'s influential analysis ("[T]he entire notion that 'particularistic' evidence differs in some significant qualitative way from statistical evidence must be questioned. The concept of 'particularistic' evidence suggests that there exists a form of proof that can provide direct and actual knowledge of the causal relationship between the defendant's tortious conduct and the plaintiff's injury. 'Particularistic' evidence, however, is in fact no less probabilistic than is the statistical evidence that courts purport to shun .... 'Particularistic' evidence offers nothing more than a basis for conclusions about a perceived balance of probabilities.") (870).

[24] "In a judicial proceeding in which there is a dispute about the facts of some earlier event," the Court found, "the fact finder cannot acquire unassailably accurate knowledge of what happened. Instead, all the fact finder can acquire is a belief of what probably happened." Quoting *In re* Winship, 397 U.S. 358, 370 (1970) (Harlan J. concurring); see also *Turner v. United States*, 396 U.S. 398, 415-17 (holding that although some heroin is produced in the United States, the vast majority is imported and as a result, a jury may infer that heroin possessed in this country is a smuggled drug, even under the beyond a reasonable doubt standard).

[25] The puzzle has been the subject of several waves of literature in law and philosophy. First, in the late 60s early 70s, including the classics Kaplan (1968), and Tribe (1971). Then in the 1980s, including Cohen (1981), Nesson (1985), and Thomson (1986). And more recently, with Colyvan et al. (2001), Schauer (2003), Redmayne (2008), Kaplow (2012), Buchak (2014), Cheng (2013) and Cheng & Pardo (2015). A number of scholars offer a response to this puzzle as part of a broader project on evidence law, including Posner (1999). There are also several helpful literature reviews and clarificatory articles, including Brook (1985), Koehler (2002), and Wright (1988).

[26] Kaplow (2014) makes a similar point: "Some have suggested in particular that Bayesian priors be ignored in applying burdens of proof . . . the suggestion is obscure: how can one insist simultaneously on applying a formula and on ignoring some of its elements? It is as if one was asked to choose the rectangle with the greater area, but in so doing to ignore the length of the rectangles under consideration. What seems to be meant, and is sometimes stated explicitly, is that fact finders should decide as if the ignored components were equal." (798).

[27] See Cheng (2013) at 1267, for example.

[28] See e.g., Thomson (1986) (arguing that the right epistemic attitude is knowledge, which is not necessarily equal to a high probability or even probability 1), Enoch et al. (2012) (arguing that we need modally sensitive beliefs, a philosopher's term of art), Redmayne (2008) (arguing that we need modally safe beliefs, another term of art) and Buchak (2014) (arguing that the right epistemic attitude is a belief, which may not be reducible to any particular probability).

mathematical probability. This is Nesson (1986)'s approach.[29] Fortunately, as I will argue, the more radical proposals are not necessary once we take into account a decision maker's sensitivity to risk of error, which already plays a central role in the construction of statistical hypothesis tests.

3.2 One person's fallacy is another's puzzle

We can identify at least three distinct apparent paradoxes of statistical evidence in the law, philosophy, and psychology literature. They are all variations on a seminal case in tort law, *Smith v. Rapid Transit*, Inc. 317 Mass. 469 (1945).[30] In each case, the intuitive judgment reported by legal philosophers is widely accepted as a fallacy by psychologists. While the cases involve the application of Bayes' Theorem, it is important to understand that they are not about Bayesian inference at all. The background information in each case is provided in the form of a population frequency and any statistician (Bayesian or not) should agree that we should condition on the evidence. So why are they paradoxical? In short, they are not. The disagreement occurs because of a confusion in what is required for a high degree of belief (the Bayesian posterior probability) and what is required to make a legal decision on its basis (a procedure for when to accept/reject a proffered theory of the case, or a part thereof).[31]

The first case was made famous by Daniel Kahneman, Amos Tversky and Maya Bar-Hillel in their studies of biases and heuristics in the context of judgment under uncertainty.[32] It goes as follows.

> **Problem 1.** Two bus companies operate in a given town, the Blue Bus Company and the Green Bus Company. Blue Bus Co operates only blue buses and Green Bus Co operates only green buses. Blue Bus Co owns 80 percent of all the buses in town and Green Bus Co owns the other 20 percent of buses. A bus is involved in a hit and run accident late at night. A witness later identifies the bus to be green. The court finds that under similar visibility conditions the witness is able to correctly identify the color of the bus about 80 percent of the time.

Suppose we introduce the witness's testimony in a civil dispute between the victim and the Green Bus Company. In cases like this, it is typically the plaintiff who bears the burden of persuasion on every element required to establish a *prima facie* case, which is said to be satisfied if the preponderance of the evidence favors the plaintiff's theory. So is this enough for the plaintiff to establish a *prima facie* case? Given this version of the problem, and a preponderance standard, the Bayesian answer is no. Let $P(G)$ and $P(B)$ represent the prior probability that a Green Bus Co or Blue Bus Co bus hit the victim, respectively, and let $L(g|G)$ and $L(g|B)$ represent the likelihood that the witness identifies a green bus given that a Green or Blue bus hit the victim, respectively. Since *posterior odds* = *prior odds* × *likelihood ratio* (3), we have,

$$\frac{P(G|g)}{P(B|g)} = \frac{P(G)}{P(B)} \times \frac{L(g|G)}{L(g|B)} = \frac{1}{4} \times \frac{.8}{.2} = 1 \tag{4}$$

Therefore, the posterior probability that a Green Bus Co bus hit the victim given that the witness identified a green bus is 1/2. The evidence is not preponderant.

In a large number of experiments, however, Kahneman & Tversky (1972) and Bar-Hillel (1980), among others, report finding that the average value for $P(G|g)$ among their subjects is approx-

[29] This approach is not persuasive especially because mathematical probability enters through expert testimony into most moderately complex disputes, and almost invariably in damages assessments. It would be very unusual to have the expert's use of the concept explicitly distinguished from legal uses of the term.

[30] I omit the details of the actual case, since the literature has grown around fictionalized variations of it, which I consider in detail below. It is really not clear how much of the actual *Smith* case hung on the admissibility of base rates.

[31] The only problem in this vicinity is the well-known reference class problem, but that is not what the cases are about.

[32] Kahneman & Tversky (1972), Bar-Hillel (1980). See generally Kahneman & Tversky (1982).

imately .8, closely tracking the witness's credibility and ignoring the underlying frequency of green buses in the town's bus market.[33] Given that posterior, we should indeed find for the plaintiff despite the above analysis. But that would be a classic case of what Kahneman and Tversky called the base rate fallacy. Indeed, the problem was originally introduced to illustrate the fallacy.

What makes this judgment bias interesting to lawyers and philosophers however, is that even upon reflection people stick with their inaccurate estimate and corresponding decision. But more than that, many scholars themselves believe that the widely observed decision is actually correct, at least in legal contexts – and that in cases analogous to the above, as we will see, our judgment about whether or not the burden of proof has been met should not correspond to the posterior probability. To see why some scholars draw this conclusion, consider the following.

> **Problem 2**. The Blue Bus Co owns 80 percent of the buses in town, all of which are blue, and Green Bus Co owns 20 percent, all of which are green. The witness testifies that a bus hit the victim (this fact is not disputed) but cannot remember its color.

This is the problem as articulated in Thomson (1986) and Nesson (1985)'s seminal papers and as a result this is the version introduced in the legal (rather than behavioral economics/ psychology) literature. On the basis of this evidence, should the plaintiff recover? Well, the only difference between Problem 1 and Problem 2 is that the witness testimony is now assumed to be undisputed and what the witness says is simply that she saw a bus.

Presumably, such evidence is not any more probable if the bus had been green than if it were blue. As a result $L(b|G) = L(b|B)$, which has the effect of making the likelihood ratio equal to 1. But it is still the case that $P(G|b)/P(B|b) \propto 1/4$ which means that $P(G|b) = 1/5$. But now suppose we switch the case around, so that rather than bringing a lawsuit against the Green Bus Co the plaintiff brought a lawsuit against Blue Bus Co.[34] Since $P(G|b) + P(B|b) = 1$, $P(B|b) = 4/5$. The Bayesian answer is now, 'yes, the plaintiff should recover against Blue Bus Co.'

And this is the apparent paradox. While it is true that the posterior probability for the claim that a Blue Bus Co bus caused the injury is .8 – well above any reasonable interpretation of 'preponderance' – it seems morally inappropriate to hold Blue Bus Co liable on this basis.[35] The challenge, then, is to explain why statistical evidence cannot underwrite a verdict against the defendant in cases like this.

But the only difference between Problem 1 and Problem 2 is that the likelihoods are equal (i.e., the ratio is 1) in the latter case. As a result, the posterior odds in Problem 2 are equal to the prior odds, which means that our priors carry all the weight, rather than only some of it, as in Problem 1. This is what drives our strong moral intuitions in Problem 2. But it is not clear why this ought to be morally relevant, and it is quite clear that it is not epistemically relevant. Whether the priors carry some, all, or none of the weight should not make a difference to our assessment. Before we move on, let us consider one more common formulation.

> **Problem 3**. The Blue Bus Co owns 80 percent of blue buses in town, and the Green Bus Co owns 20 percent of blue buses in town. The witness testifies that a blue bus hit the victim (this fact is not disputed).

This is the presentation of the problem as given in Tribe (1971)'s seminal article and it is the version of the problem taken up in the more recent literature, by Cheng (2013) and Buchak (2014), for example. If this is presented in court should the plaintiff recover? The Bayesian answer is similar to the answer in Problem 2, except the likelihood now corresponds to the witness testimony of a blue bus, rather than the witness testimony of a bus. But the relevant base rate is now the proportion of

[33] For experimental evidence on legal hypotheticals in particular, see Wells (1992).
[34] This is structurally identical to Cohen (1981)'s Paradox of the Gatecrasher.
[35] I am assuming here that Blue Bus Co does not introduce any evidence in rebuttal.

blue buses owned by Blue Bus Company, which is again 4/5, and the likelihoods are again presumably equal, since it is not more likely that the witness would identify a blue bus if that blue bus belonged to the Blue Bus Company than if the same blue bus belonged to the Green Bus Company. So we have, again, a posterior probability of 4/5, which we will now denote by $P(B|bl.)$. As in Problem 2, most people feel uncomfortable about concluding that the plaintiff could win, or even make out a *prima facie* case, on the basis of the statistical evidence. Structurally, however, the three problems are identical. In each case, likelihood is likelihood is likelihood. And the uniformly correct solution is to be found by applying Bayes' Theorem.

One worry we may have is that we would not want a judge or jury to come into a case with a prior bias of who is more likely to win, since notions of fairness or impartiality in the legal system require that we consider the competing accounts on an equal footing, so to speak. One response to this would be to deny that this is indeed what we should do. If a plaintiff comes to court with an absurd claim, we should not feign credulity for the sake of impartiality.

For example, it seems plausible that when a plaintiff alleges being bitten by defendant's house cat, or having developed autism as a result of the defendant manufacturer's flu vaccine, we should indeed approach the claim with some initial skepticism.[36] But if the worry is procedural – i.e., that a jury should not rely in their decision making on evidence not in the record – we can simply assume that the fact finder does indeed begin with a uniform prior, and that the base rates are then admitted into evidence.[37] So consider the following problem, which is again mathematically indistinguishable from problems 1-3.

> **Problem 4**. Same as problem 3 except (a) the decision maker (judge or jury) approaches the case with a uniform prior over the two hypotheses and (b) the base rates are entered into evidence by the plaintiff at trial, followed by the witness testimony.

Anyone who shares the intuition that statistical evidence cannot properly underwrite a legal judgment should still share that intuition in Problem 4. After all, the disagreement is not supposed to be about procedure. The apparent problem with statistical evidence is not just that decision makers inappropriately rely on it when it is not in the factual record. It is that even if it were in the factual record, it would not be morally appropriate to rely on it.

In problem 4 we have to update twice. In the first update, the equal priors cancel out, leaving a likelihood ratio of 4/1 which means that the posterior probability that a Blue Bus Company bus injured the plaintiff is 4/5. Now we introduce the eyewitness testimony, which means that the likelihood ratio becomes the new prior odds and the problem becomes identical to that in Problem 3. If the new likelihoods are equal, then they cancel out as well, which means that the new posterior is equal to the new prior odds – namely, $4:1$ – and the probability that a Blue Bus Company bus hit the victim is again 4/5. This is all consistent with starting the case out with the parties in equipoise, as Cheng (2013) puts it. Since Bayesian conditioning is commutative we would get the same result if we reversed the order of the updates.

Therefore, what should have been a case of mistaken reasoning (ignoring base rates) came to form the basis for a family of apparent puzzles about evidence. In the sections that follow, we will see how the adaptive model can help us make sense of cases where statistical evidence seems inappropriate as well as their apparent counter examples (like DNA random match profiles).

[36] Kaplow (2012) says, for example, "it seems unlikely that [legal decision makers] would ignore, for example, whether a characterization of events proffered by a party was a priori quite unlikely or bizarre versus entirely ordinary human behavior." See also Diamond & Vidmar (2001) (describing cases where legal decision makers ordinarily incorporate prior information).

[37] One might go further and argue that it is not even possible to make a decision without relying on information outside the record. A judge or jury cannot evaluate admitted evidence from the perspective of a true blank slate, as it were.

4 Royall's Three Questions

In a classic monograph on statistical inference, Richard Royall distinguishes three related questions about evidence: (Q1) What should I believe?; (Q2) What does this observation tell me about the competing hypotheses?; and (Q3) What should I do after making an observation? (Royall, 1997). The problem, as we will see, with the literature on statistical evidence is that most commentators assume the legal system is in the business of answering (Q1) or (Q2) when instead the evidentiary process is characterized by a decision procedure for dealing with (Q3).[38]

The first question may be answered with a Bayesian posterior probability. There is no exception to this. If all you're interested in is what you should believe – in epistemic heaven, so to speak – then the optimal approach is the Bayesian posterior. This is because in epistemic heaven the only thing you ought to care about is the accuracy of the beliefs you hold. And in a dynamic context, where you will revise your beliefs in response to new evidence, it seems sufficiently plausible that the only thing you ought to care about is the accuracy of the beliefs you end up with. Provided this is true, then for a large class of very plausible measures of probabilistic accuracy (including, basically, every measure used in the forecasting literature),[39] updating by Bayesian conditioning on one's priors maximizes the expected accuracy of the posterior probabilities (Greaves & Wallace, 2006). Bayesian conditioning, therefore, is optimal from the perspective of (expected) accuracy and will in this sense always give the best answer to (Q1). But legal fact finding does not take place in epistemic heaven. We engage in legal fact finding in order to figure out what happened, so that we can grant relief where it is appropriate. This is the hallmark of the evidentiary process – it is not just fact finding in the abstract. It is fact finding as the basis for a subsequent practical decision.

How should we answer (Q2)? I mentioned above that the likelihood ratio is reckless in its responsiveness to evidence. But if all we are interested in is evaluating the relative strength of the evidence, full stop, then the likelihood ratio's recklessness is a virtue. Suppose that $H : \overline{H}$ is $2 : 1$. Then what the evidence says is that the null hypothesis is twice as plausible as the alternative.[40] This approach – i.e., the approach to answering (Q2) – easily lends itself to legal application. It seems reasonable to suggest that in civil litigation we should decide in the plaintiff's favor if the likelihood of the data under her hypothesis is greater than the likelihood of the data under the defendant's hypothesis. In other words, find in favor of the plaintiff if,

$$\frac{L(\mathbf{X}|H)}{L(\mathbf{X}|\overline{H})} > 1 \tag{5}$$

This is effectively the proposal Cheng (2013) puts forth.[41] It is equivalent to (1) with $k = 1$. But in a legal context we are not interested in merely evaluating the relative likelihood of the evidence. The likelihood ratio is a measure of the degree to which the evidence increases the posterior probability that the plaintiff is right. But it would be a mistake to make decisions on the basis of facts about incremental increases in evidence rather that on total evidence.

Suppose again the plaintiff alleges, implausibly, that she was bitten by the defendant's house

[38] For example, Thomson (1986), Buchak (2014), and Enoch et al. (2012) assume that the burden of proof is defined by some fixed epistemic standard – such as a modally robust belief (Q1). If that standard is met, liability is appropriate. Meanwhile, Cheng (2013) assumes that what matters instead is the weight of the evidence only (Q2). If the plausibility of the plaintiff's claim relative to the defendant's claim is high enough, liability is appropriate.

[39] The measures I am referring to are the so-called strictly proper scoring rules. See Winkler & Murphy (1968), Savage (1971), Schervish (1989), and Lindley (1982) for classic discussions and Gneiting & Raftery (2007) for a contemporary overview. In studies of probabilistic forecasting, Brier (1950)'s quadratic score is usually applied as a measure of accuracy. See e.g., Merkle et al. (2016) (applying the Brier score to evaluate geopolitical probability judgments).

[40] See e.g., Hacking (1965) and Sober (2008) for a likelihoodist approach to evaluating evidence.

[41] Strictly speaking, Cheng's model takes this form because he assumes that the prior odds are equal to 1. I evaluate this assumption in detail in §5.1.

cat, but the defendant chooses to respond by arguing, instead, that the bite mark was caused by a third party's goldfish. Since the denominator of this likelihood ratio will be virtually zero, we are guaranteed to satisfy (5) no matter how improbable the house cat theory is. But that does not mean we should accept the goldfish theory. Doing so would be an instance of what Spanos (2013) calls the fallacy of rejection (misinterpreting evidence against a hypothesis as evidence for the alternative). This fallacy arises in the example because, by offering a specific theory in rebuttal, the defendant has chosen to respond in a way that results in a non-partitioned parameter space. In a legal context, we want to evaluate the evidence in light of what we know about the world and the underlying factual circumstances so as to make our best guess about the most probable sequence of events leading to the complaint.

As a result, in the context of legal decision making we need an answer to Royall's (Q3). Now it may seem like (Q3) has very little to do with statistical inference. But it is really (Q3) that Neyman & Pearson (1933)'s classic null hypothesis significance testing approach seeks to answer. In hypothesis testing, as in legal decision making, we need a justifiable procedure for when to reject one or the other of the competing claims. What will be especially important to us here, though, is not so much the particular procedure used in NHST (the uniformly most powerful level α test) but rather the method by which such a procedure is constructed and the normative assumptions presupposed in its development.

4.1 Hypothesis testing and epistemic risk

To construct a hypothesis test, we start by identifying a null hypothesis H_0 and an alternative hypothesis H_1, which are statements about $\theta \in \Omega$. For example, $\theta = 0$ and $\theta \neq 0$. To make this concrete, these might be statements about, say, the correlation between blood pressure and sugar consumption, appropriately defined. A hypothesis testing procedure is a rule that specifies the sample points for which H_0 is accepted. This is our acceptance region S_0. And it specifies a set of sample points for which H_0 is rejected. This is our rejection region S_1. These two partition the sample space so that $\mathbf{x} \in S_0 \cup S_1$. The rejection region is usually defined through a test statistic $W(\mathbf{X})$, which is a function of the data. For example, we may use the observed correlation, and agree that we will reject a hypothesis of no effect (in the sugar example, correlation 0) if the correlation observed in the sample is, say, $\rho > |.25|$.

The important question is the following: on what basis should we select such a testing procedure? This is what we need to answer (Q3). In identifying a hypothesis test, there are two important things to worry about. Notice that if we set the rejection threshold really high, say .75 (recall that the range of the correlation coefficient is between -1 and 1), we will almost certainly not reject the null hypothesis. So it is extremely improbable that we will conclude that sugar affects blood pressure when in fact it does not. This is a very conservative procedure. At the same time, however, by setting the threshold so high we are taking a different kind of risk. Namely, the risk of failing to appreciate an effect between sugar and blood pressure that in fact exists, though perhaps not to such a strong degree. To increase the probability that we detect an effect when indeed it is there we should bring the threshold down. But as we do this, of course, we also increase the probability of rejecting our null hypothesis in response to sampling noise, which would be very probable if we set it to, say, 0.0001.

The first kind of error is a Type I or false positive error. We will express its probability as,

$$P(\mathbf{X} \in S_1 | H_0)$$

This is to be read as the probability that our observed data fall into the rejection region, on the assumption that the null hypothesis is in fact true. The second kind of error is known as a Type II

or false negative error, whose probability is,

$$P(\mathbf{X} \in S_0|H_1)$$

which is to be read as the probability that our observed data fall outside the rejection region when in fact the null hypothesis should be rejected (i.e., the alternative is true). An ideal test would eliminate the probability of error altogether. In practice this is impossible. As a result, we have to select a test by choosing a tolerable level of both Type I and Type II error rates. This is what is important to us here: the fact that a hypothesis test is selected by considering our tolerance to risk or error – or, *epistemic risk*. Equivalently, this implies that every test reflects an implicit trade off between the different types of error rates.

So, then, what would be a reasonable level of epistemic risk to assume in the legal context? Surely this depends on the circumstances: what is at stake for the parties who will be bound by this decision? And how may we expect the decision to affect future conduct? Under Neyman & Pearson (1933)'s NHST approach, however, such considerations are generally not relevant. Instead, we identify some tolerable Type I error probability α and then look for the test, among all level α tests, that minimizes the probability of Type II error. Since α rarely varies from case to case, the test is not context sensitive. Moreover, false positives are uniformly privileged. But this is not the only way to identify a hypothesis test. In the next section, I will identify a more flexible selection procedure and apply it to legal decision making.

4.2 Minimizing a linear combination of error rates

To balance the relevant consequences in identifying a decision procedure we need to pay attention to the relative costs of the different types of error, including forgone benefits that would have accrued if we rendered an accurate verdict. For example, we might agree with the plaintiff that a product injuring them was defective when in fact it was not (perhaps the plaintiff sustained injuries through misuse) – a Type I or false positive error. Or, we might reject plaintiff's allegation when in fact it is true – a Type II or false negative error. We have significant room for judgment in designing evidential procedure so as to balance the two types of error rates. For example, a system that awards damages to every plaintiff who makes a colorable case by surviving a motion to dismiss would effectively eliminate type II errors. On the other extreme, we have the standard in criminal cases – proof beyond a reasonable doubt – which reflects significantly more concern for Type I errors. The preponderance standard is typically taken to be somewhere between these two extremes.

Let δ refer to the evidential procedure in our venue. Then the Type I and Type II error rates are functions of δ and we can refer to them as $\alpha(\delta)$ and $\beta(\delta)$, respectively. In other words,

$$\alpha(\delta) = P(Reject\ H_0|H_0) = P(\mathbf{X} \in S_1|H_0)$$
$$\beta(\delta) = P(Accept\ H_0|H_1) = P(\mathbf{X} \in S_0|H_1)$$

Suppose δ is the evidential procedure that finds for every Plaintiff surviving a motion to dismiss. Then $\alpha(\delta) = 0$. Meanwhile, where δ is, let's say, an extremely strict version of the beyond a reasonable doubt standard, $\beta(\delta) \approx 0$. Notice, however, that in the first case where $\alpha(\delta) = 0$, $\beta(\delta)$ will be high. Its precise value depends on the underlying distribution of actually harmful acts among acts that are alleged to be harmful but if we assume for the sake of this example that approximately half of the defendants have committed the act they are accused of committing then $\beta(\delta) \approx .5$.[42] Meanwhile, under the same assumption, in the case where $\beta(\delta) \approx 0$, $\alpha(\delta) \approx .5$.

It seems reasonable to suppose, then, that in order to identify a decision procedure and, in turn, answer Royall's (Q3), we should strike some balance between $\alpha(\delta)$ and $\beta(\delta)$. My argument here is very minimalist. I do not intend to argue for a particular way of striking that balance. Rather, I simply

[42] Cf. Laudan & Allen (2008) (estimating the frequency of false exoneration in criminal trials).

suggest that in order to identify an appropriate procedure we should consider what that balance ought to be. Or, to put this in evaluative terms, *we can assess legal decision making by reference to whether the balance that it reflects about the relative costs of error of either sort is appropriate from a moral perspective.*[43]

Here is a very general proposal. We have two types of error rates, $\alpha(\delta)$ and $\beta(\delta)$, and we necessarily have some costs associated with them, let us call these a and b, respectively. It seems sensible that regardless of our attitude to risk of error of either kind, in the legal context we should seek to minimize a *linear* combination of *weighted* error rates. Why linear? Because it is effectively the least restrictive mixture of two quantities. An affine combination is a linear combination that requires the weights to sum to 1 and a convex combination is an affine combination with non negative weights. But a linear combination of error rates puts no restrictions on the weights. Since I want to develop a broadly applicable model of legal decision making, the fewer assumptions we make the better. Therefore, we should identify the evidentiary procedure δ among the set of all available procedures Δ which satisfies,

$$\min_{\delta \in \Delta} a\alpha(\delta) + b\beta(\delta) \tag{6}$$

This is a very general expression since we have not yet specified a value for any of these parameters and there are no restrictions on the weights. Its generality is a strength. Our attitude to risk of error in either direction may be affected by a number of factors. We have already seen one such concern above – that is, we may think that the unfairness to an innocent person who is wrongly convicted in a criminal case is worse than the unfairness to the victim (or perhaps society) of having a guilty person falsely acquitted. This is a predominantly backward looking or ex post consideration. It may be reducible to its effect on individual utilities (because, for example, we disprefer living in a society that puts innocent people in jail at a greater rate than we disprefer a society that acquits guilty people) but it may not be. We may believe instead, as Tribe (1971) suggests, that there is a particular injustice to the autonomy of an individual by falsely punishing her on the basis of her membership in a class.

For Tribe (1971) and Wasserman (1991), this is an injustice that goes beyond what can be captured in the social welfare function. But that is not a problem for the linear combination approach, because the view I defend is a more general version of Kaplow (2014)'s economic model. It enables us to capture Tribe and Wasserman's concerns because it implies that the reason we may think false positives are so bad is not so much because of the social consequences that the decision may produce (ex ante) but rather because of the severe injustice that we accrue by violating an individual's autonomy in punishing her on the basis of class membership (ex post). If we constrain the model I propose by adding the assumption that the only considerations permissible in determining the values of a and b are considerations that affect the social welfare function, then the approach will be equivalent to Kaplow's. In other words, Kaplow's model is a special case of the adaptive model with the social welfare condition on the support of a and b.

It is also possible to maintain that none of this is relevant to the fact finding process in legal trials, in which case $a = b = 1$. This is effectively what Cheng (2013) and Cheng & Pardo (2015) propose. But as lawyers like to say, inaction is an act, so it is worth keeping in mind that refusing to evaluate the relative normative importance of a and b by setting them equal to each other is itself a choice reflecting a value judgment about the permissible attitudes to risk of error. In any case, my main point here is that whatever approach you prefer to setting the parameter values, the basic idea – that we should minimize a weighted linear combination of error rates – is very plausible. If this claim is right, then it provides a very strong justification for the model I defend – namely, an adaptive Bayesian likelihood ratio test. We will now prove this.

[43]The most clear example of a decision procedure generated by considering the error ratio is in the criminal context where the Blackstone dictum suggests that a Type I error is ten times as bad as a Type II error.

4.3 Epistemic risk and the adaptive model

DeGroot & Schervish (2012) show that if our goal is to minimize (6) then the optimal test will be in the form of a risk-weighted likelihood ratio. I will assume, as I mentioned above, that our data is drawn from a binary distribution – for example, $X = 1$ if the defendant company owns the bus that caused the accident and $X = 0$ otherwise. Since we want to find the test δ that minimizes $a\alpha(\delta) + b\beta(\delta)$, which is equal to $\sum_{\mathbf{x}\in S_1} af(\mathbf{x}|H) + \sum_{\mathbf{x}\in S_0} bf(\mathbf{x}|\overline{H})$, where $f(\cdot|H)$ is the probability distribution of the data under H, then, by rearranging this expression, we have to choose a critical region that minimizes,

$$b + \sum_{\mathbf{x}\in S_1} [af(\mathbf{x}|H) - bf(\mathbf{x}|\overline{H})] \tag{7}$$

In other words, we want the region that includes every point x for which $af(\mathbf{x}|H) - bf(\mathbf{x}|\overline{H}) < 0$ because every such point will decrease the overall sum. Therefore, the test δ^* that minimizes the sum in (7) will reject the null hypothesis when $af(\mathbf{x}|H) > bf(\mathbf{x}|\overline{H})$. As a result, we will reject the null whenever the probability of the data under it, weighted by the cost of falsely rejecting it, is less than the probability of the data under the alternative, weighted by the cost of falsely accepting it. Rearranging and expressing the statistic as a function of the parameter, we get a weighted likelihood ratio test. That is, accept the plaintiff's claim H if,

$$\frac{L(\mathbf{X}|H)}{L(\mathbf{X}|\overline{H})} \geq \frac{b}{a} \tag{8}$$

Notice that if we let $k = b/a$ then (8) is equal to (1). What we get from DeGroot & Schervish (2012), however, is an interpretation of k in terms of the risk of error – i.e., the parameters a and b corresponding to Type I and Type II error costs, respectively – and a proof for the claim that this is the test we need to use if our goal is to minimize a linear combination of error rates, as I argued it should be.

I explained above that a likelihood ratio on its own is far too sensitive to the evidence and in particular to evidential noise. We are looking for an answer to (Q3) whereas (1), as we saw, gives us an answer to (Q2). But we can extend the proof to get what we need. First, multiply both sides by $P(H)/P(\overline{H})$, to get,

$$\frac{P(H)}{P(\overline{H})}\frac{L(\mathbf{X}|H)}{L(\mathbf{X}|\overline{H})} \geq \frac{P(H)}{P(\overline{H})}\frac{b}{a} \tag{9}$$

Since there is no restriction on the cost parameters a and b, let $b^* = bP(H)$ and $a^* = aP(\overline{H})$. Then the following test is likewise risk optimal.

$$\frac{P(H)}{P(\overline{H})}\frac{L(\mathbf{X}|H)}{L(\mathbf{X}|\overline{H})} \geq \frac{b^*}{a^*} \tag{10}$$

The left hand side should be familiar now – by (3) it is equal to the Bayesian posterior odds $P(H|\mathbf{X})/P(\overline{H}|\mathbf{X})$. Strictly speaking b^* and a^* are now a different pair of constants, but since there is no restriction on their range I will drop the asterisk, below.[44]

We now have an answer to Royall's (Q3) and a recipe for constructing a legal standard of proof: decide in favor of the plaintiff if and only if the risk-weighted posterior probability of her hypothesis is greater than the risk-weighted posterior probability of the competing hypothesis. That is, our optimal

[44] This change in the value of the risk parameters in going from a pure likelihood ratio to a prior weighted likelihood ratio does imply that if a Bayesian and a likelihoodist are to reach the same verdict in contested statistical evidence cases, such as problems 2-4, the Bayesian must be more sensitive, so to speak, to risk of error. Spelling this out in detail would take us too far off field, however, as the core argument does not hang on this remark.

test δ^* takes the following form: Accept plaintiff's claim if,

$$P(H|\mathbf{X})/P(\overline{H}|\mathbf{X}) > b/a \tag{11}$$

This is identical to our statement of the adaptive model in (2), except now the rejection threshold is finally defined in terms of a ratio of error costs and a statistical optimality proof is given to justify the approach.[45] Moreover, the test imposes a probability threshold on legal decision making, in the sense that we decide for the plaintiff if the posterior probability of her claim, $P(H|\mathbf{X})$, exceeds $b/(a+b)$. But that threshold is conditional on the decision maker's tolerance for risk of error – that is, the relative magnitudes of a and b. What we have added, therefore, is some substance to the parameters of our Bayesian hypothesis test. The model is adaptive because it has a shifting rejection threshold. And that threshold shifts in response to the decision maker's sensitivity to risk of error, or epistemic risk. Each of the three terms in the model – prior, likelihood, error rate – are variable. Let us now compare this to Cheng (2013) and Cheng & Pardo (2015)'s alternative accuracy-first model.

5 Risk Adaptive Burdens of Proof

To keep things as simple as possible, let p_1/p_2 denote the prior odds for H and $\overline{H}$, respectively, and let L_1/L_2 denote their respective likelihood ratio. Our Bayesian hypothesis test then enjoins us to accept the plaintiff's claim if $(p_1/p_2)(L_1/L_2) > b/a$, where a and b are the weights of the Type I and Type II error rates, respectively. This is just a simplified expression of (11).

5.1 The restrictive approach

The apparent puzzle in problems 2-4 is that since $p_1/p_2 = 4$ and $L_1/L_2 = 1$ we are committed to the conclusion that we should find in favor of the plaintiff provided that $a = b = 1$. This is the restriction. To say that we assume a and b are equal, given this model, is equivalent to saying that we should decide in favor of the party with the comparatively higher posterior probability. In other words, this is now the familiar threshold approach, where we decide for the plaintiff if the probability of her theory of the case exceeds .5. That is, decide for the plaintiff if $(p_1)(L_1) > (p_2)(L_2)$.

But why should we set a and b equal to each other? Cheng argues that in civil litigation at least, it is plausible to assume that the cost of false positives is equal to the cost of false negatives. Posner (1999) makes the same assumption. The idea here is simply that $a = b = 1$ is the mathematical equivalent of assuming that the colloquial expressions 'preponderant' and 'more likely than not' are synonymous. But this apparently reasonable assumption is what gave rise to the apparent puzzles of statistical evidence in §3.2. Therefore, to avoid implausible verdicts in problems 2-4 while keeping $a = b = 1$, Cheng stipulates that we artificially set the prior odds to $p_1 = p_2 = 1$ as well. "In civil trials," Cheng says, "the prior probabilities as a normative matter should arguably be equal" (1267).

This is extremely important and it is a position he is forced into. After setting $a = b = 1$, in order to capture what he takes to be a central property of the preponderance standard, Cheng has to either concede that even extremely strong statistical evidence could be insufficient for legal liability or, to counterbalance that move, he can set $p_1/p_2 = 1$ as well. As Posner (1999) puts it: "If the prior odds are assumed to be 1 to 1, on the theory that the jury begins hearing the evidence . . . without any notion of who has the better case, then the posterior odds are equal to the likelihood ratio" (1508). That is exactly correct. As I highlight below, Cheng's approach is not so much a solution of the apparent paradox as it is a mathematical restatement of it.

For Cheng (2013) and Cheng & Pardo (2015), therefore, both the prior odds p_1/p_2 and the rejection threshold b/a are fixed at 1. Both assumptions lead to an unduly restrictive model of decision making. First, let us take up the assumption that $a = b = 1$. We are not told what the

[45] Whether we use a strict or non-strict inequality does not matter, since a and b are unrestricted constants.

normative reasons are that require such specificity in the treatment of the cost parameters. Indeed, such specificity seems implausible. Consider mass exposure cases, like asbestos litigation. One kind of error we could make is to hold a manufacturer liable in a world where asbestos is harmless. This imposes a direct cost on the manufacturer. Moreover it imposes indirect costs on other manufacturers by setting a precedent for holding them wrongly liable in subsequent disputes. The other kind of error is failing to hold the manufacturer liable when in fact asbestos caused the plaintiff's illness. This imposes a direct cost on the plaintiff by making it impossible for her to recover the expenses associated with her illness. Similarly, it imposes indirect costs by setting a precedent against recovery from asbestos manufacturers. As a result, manufacturers continue to produce the harmful substance, leading to debilitating illness and premature death across many generations. This analysis of course holds for mass exposure cases in general, not just asbestos.[46] My argument does not rely on convincing the reader that the latter cost is necessarily greater than the former cost (though it probably is). My argument merely relies on denying that we ought to stipulate in advance that these costs are exactly equal, whatever they happen to be. That is, it strikes me as presumptuous to suppose that regardless of the case and its factual circumstances, the two kinds of errors are necessarily equally important. But this is what Cheng's model of fact finding in the civil context commits us to.

Second, at this point the only thing left to vary in the model is the likelihood ratio. That is, decide for the plaintiff if $L_1 > L_2$. Since $L_1 = L_2$ in the problems we have considered, Cheng is able to deliver the intuitive result – neither side would have preponderant evidence. But the test is no longer a Bayesian hypothesis test. As Cheng says in one of the footnotes to the above quoted text, "setting the prior odds to 1 for normative reasons necessarily means that the expression no longer equals [the posterior odds] in the strict mathematical sense" (Cheng, 2013, 1268, n. 26). As a result, he is now stuck with all the implausible verdicts that a simple likelihood ratio would generate. For example, if Blue Bus Co. owned $> 99\%$ of buses in town we could still not hold it liable because that would not affect the ratio L_1/L_2. The cure is worse than the disease. Cheng, I suspect, recognizes that the likelihood ratio on its own is not a robust estimator. This is well-known, and it leads to the following extremely important footnote: "Setting the prior odds at 1:1 may be wrongheaded as a matter of inference . . . but that does not mean that courts do not do it" (Cheng, 2013, 1267, n. 24).

This is the trade off Cheng is forced to make – and that I would prefer to avoid. In order to force a result that is consistent with the statistical evidence intuitions, he has to build into his model of legal fact finding what, by his own admission, is a wrongheaded approach to inference – a model that requires us to commit the base rate fallacy – and impute that approach to judges and juries. In other words, Cheng does not resolve the apparent conflict between the demands of epistemic rationality and our moral obligations to the defendant. Rather, he stipulates that in cases like problems 2-4, our normative commitments supersede the requirements of epistemic rationality. Our moral commitments enjoin us to be epistemically irrational.

Cheng & Pardo (2015), drawing on Wald (1945), argue that we should ignore prior probabilities, not merely as a normative matter, but because that is the decision rule that minimizes the maximum loss due to error. This is an improvement in the sense that the assumption that we ignore prior probabilities is given a decision theoretic foundation. But a likelihood ratio test will minimize maximum error loss (i.e., is minimax optimal) *only if* we assume that the costs of Type I and Type II error rates are necessarily equal. This is because a minimax optimal test asks us to consider the severity of our error, weighted by its probability, if the plaintiff is correct, against the severity of our error, weighted by its probability, if the defendant is correct. This test reduces to comparative likelihood only if the weights of those errors are equal, because it is only under the assumption of equality that the worst case outcome is *either* a false negative decision *or* a false positive decision. By changing the weights asymmetrically, we can change the worst case outcome, in which case the optimal test

[46] See e.g., Rosenberg (1984) ("Even a single instance of product defect, carelessness, or risk-taking may increase for thousands or even millions of people of one or more generations the danger of contracting cancer or some other insidious disease.").

will require us to consider the likelihood of the plaintiff's hypothesis against some multiple of the likelihood of the defendant's hypothesis. Such a test, of course, would not be coextensive with Cheng & Pardo (2015)'s comparative likelihood approach.

So while their revised model gives an argument for the assumption that $p_1 = p_2 = 1$, it does not defend the assumption that $a = b = 1$. Another way of putting this is to say that on their revised model, if $a \neq b$ then either $p_1 \neq p_2$ or they cannot vindicate the familiar judgments in statistical evidence cases. Their initial model ignores the priors only because the costs of both error rates are assumed to be equal. The revised model gives an argument for ignoring priors provided you agree that the costs of error rates are indeed equal.

More generally, the minimax approach is really a special case of the linear combination of error rates model that I defend here. In particular, it is the linear combination of $a\alpha(\delta) + b\beta(\delta)$ with $a = b = 1$. So the adaptive model generalizes Cheng and Pardo's minimax loss model in allowing a and b to take on different values. And as we will see below, it generalizes Kaplow's approach in being more liberal about what sorts of considerations can affect those values.

5.2 The adaptive alternative

Unlike Cheng and Pardo, I let everything in the model vary – the priors, the relative costs of error and, of course, the likelihood. This approach helps us to understand why most people are hesitant to find against the defendant in problems 2-4 without assuming that the reasoning process of judges and juries is epistemically defective or wrongheaded from a truth-seeking or inferential perspective. It also has another important advantage – namely, it accommodates just as well apparent counter examples to the inadmissibility of statistical evidence. In doing so, however, it exposes the inevitable encroachment of value judgments – in particular, the relative sensitivity to epistemic risk – on legal fact finding. This is extremely important for our understanding of the preponderance standard. The implication is that *there is no one size fits all threshold even when the burden of proof is defined as preponderance of the evidence.* Rather, we have a Bayesian hypothesis significance test whose parameter values are determined by the factual circumstances of the case. Hence, the *adaptive burden of proof.* This is consistent with the empirical evidence on judges' understanding of the preponderance standard. McCauliff (1982), for example, reports a study of 175 judges where a significant number took 'preponderance' to mean anything between .5 and .8 probability. While the median was .5, 63 judges understood it to require a probability greater than .6, and six judges responded greater than .9.[47]

5.3 Epistemic risk and the phantom menace

Consider Problem 3, as that is the most popular statement of the puzzle (everything here generalizes to the other descriptions). Applying the adaptive model to Problem 3, we get the following expression: $4 > b/a$ or, more helpfully for us, $b < 4a$. Since the posterior probability is greater than .5 if you share the intuition that a civil judgment is inappropriate here, then you must be especially concerned about false negative errors – failing to find the bus company liable when in fact its bus injured the victim. But we can do better than that – since the posterior probability is equal to .8 we can put a precise bound on your relative concern for Type II error.

If you share the judgment that in Problem 3 statistical evidence is inappropriate, then you are denying that $b < 4a$ which in turn implies you must think that $b \geq 4a$. And that is why you do not want to let yourself be pushed by the priors to find the company liable – they are just not strong enough given your particular degree of sensitivity to error. Now consider even more extreme examples. For instance, if the prior odds had been $7 : 1$ then the implication would be that for someone who

[47] Interestingly, the distribution was so left skewed that zero judges gave an answer less than .5. This is exactly what the adaptive model predicts. If we think of $a/b = 1$ as the epistemically risk neutral position in legal decision making and $a/b > 1$ as risk avoidant, then $a/b < 1$ would be a risk seeking attitude. A judge who believes that preponderance implies a threshold of less than .5 would then be a risk seeking decision maker which would be very odd in this context.

still believes they should not be used $b \geq 7a$. This trade-off is starting to look irrational. In other words, we can understand what seem to be commonly held judgments about statistical evidence by evaluating the decision maker *as if* she were implicitly setting the weights to be less than or equal to the reciprocal of the prior odds. That is, $b/a \leq 1/(p_1/p_2)$ or, equivalently, $bp_1 \leq ap_2$. The latter expression makes explicit what we are modeling the decision maker as doing – namely, discounting the prior probability that the plaintiff's hypothesis H is true by the weight we put on false negative (Type II) errors and comparing that to the probability that the alternative hypothesis $\overline{H}$, discounted by the cost of a false positive (Type I) error, is true.

But the point of the model is not merely to accommodate just about any judgment. Rather, because the agent's decision reflects a particular normative attitude – their degree of sensitivity to error – we can use the reasonableness of the implied attitude to assess the quality of the fact finder's decision. In other words, the adaptive model sharpens the normative considerations at stake. At $b \geq 4a$, this may still be a reasonable attitude to risk. At $b \geq 7a$, it is less clearly reasonable. At $b \geq 1,000,000a$ it is definitely irrational.

Here is the especially nice part. DNA random match profiles are often highlighted as a counter example to the normative irrelevance of base rates as evidence of identity, since virtually everyone agrees that DNA evidence, despite being inherently statistical, should be used in legal fact finding (Zabell, 2005). On Cheng's approach, it is really not clear how we can make room in our model for such exceptions to the rule, since he requires us to to set the prior odds to 1 in advance. Since this value is fixed, there is no longer any room for a base rate, even when everyone agrees it is a good one. But what happens on the adaptive model? Well, DNA evidence is usually indeed quite extreme. If the defendant is identified by DNA the prior odds will be at least $1,000,000 : 1$. If you *still* think that this is not enough for a verdict then what this says about your attitude to epistemic risk is that in fact $b \geq 1,000,000a$ which, again, is clearly irrational in legal decision making. While I argued above that it is inappropriate to assume that false positives are *exactly* as bad as false negatives, it is equally obvious that whatever their relative cost, it cannot be that false negatives are a million times worse than false positives. We can assume *that* much.

So the adaptive model not only captures what are taken to be the hallmark problem cases (i.e., problems 2-4) but it also captures what are taken to be the hallmark exceptions to the usual diagnosis (e.g., DNA evidence).[48] Cheng does not think, and neither do I, that our models should form the basis for reforming the legal system. He wants a model that captures the way courts currently approach problems like this. As do I. Where we disagree is on the conclusion to draw from problems like 2-4 because of the discrepancy in how we parameterize our models. Cheng is forced to assume that judges and jurors are epistemically irrational. But, he suggests, such epistemic irrationality may be mandated by the nature of the legal system. Meanwhile, I conclude that legal decision makers are very risk sensitive. So perhaps that leaves us with competing trade offs. But the tie breaker in my benefit, I think, lies in the adaptive model's ability to accommodate countervailing judgments (such as in the case of DNA profiles).

5.4 The adaptive model in action

It is usually supposed in the literature on statistical evidence that the case law is compatible with popular intuitions in problems 2-4: namely, even high posterior probabilities are inappropriate evidence in support of identity or more generally causation when they depend exclusively or almost exclusively on base rates. This is simply not true. Sometimes statistical evidence is excluded but very often it is not. Whether or not statistical evidence is permissible varies from context to context. And the adaptive model helps us understand (and predict) when such evidence would be admitted.

[48] See e.g., *United States v. Bonds*, 12 F.3d 540, 551-68 (6th Cir. 1993) (allowing overtly probabilistic evidence concerning DNA profiles to be submitted to the jury).

Koehler (2002), for example, suggests that courts are more likely to view base rates as relevant when they arise in cases he describes as having a statistical structure. The idea is that some people think intuitively about probability in terms of repeated sampling, and this is more appropriate in some contexts than others. When it comes to evidence of identity in torts or crimes, courts are likely to find it especially inappropriate to think about the defendant or the trial as a randomly selected point from a random sample of similar defendants or trials. The thought, mirroring Tribe (1971) and Wasserman (1991)'s arguments, is that we owe it to the defendant to adjudicate her case as an autonomous individual. Indeed, Tetlock et al. (2000) suggest that people think of some base rates as morally forbidden.

Some cases fit this profile very well. In *State v. Claflin*, 690 P.2d 1186, 1190 (Wash. Ct. App. 1984), the court found that testimony that 43% of child molestations were committed by father-figures, in a case where the defendant was a father-figure, was "extremely prejudicial and should not have been admitted." This is a classic case of what Tetlock et. al. have in mind as a taboo or forbidden base rate – i.e., the proportion of child molesters who are also father figures. And the result is predictable under the adaptive model because it is precisely under circumstances like this – i.e., circumstances of morally circumspect or taboo base rates – that we should be especially worried about the cost of falsely convicting a defendant on the basis of their membership in an otherwise innocuous class (i.e., the class of father figures).

So suppose $a = 10b$. This seems perfectly reasonable in a case where someone might go to jail because they belong to a class consisting of father figures. Our rule, then, is to convict only if $(p_1/p_2)(L_1/L_2) > 10$. If we assume that $L_1/L_2 = 1$, then we will convict only if $p_1 > 10p_2$. In other words, the base rate would have to be $p_1 > .91$ (i.e. 10/11ths) – more than twice the base rate that the court rejected in the actual case – if the evidence is not to be excluded as unduly prejudicial (or, as the case may be, on other grounds).

Meanwhile, in the context of Title VII disparate impact claims, where the *prima facie* case requires the plaintiff to produce evidence in support of the claim that a facially neutral practice has produced a pattern of discrimination, courts have held that statistical evidence alone may be sufficient. In *Bridgeport Guardians, Inc. v. City of Bridgeport*, 933 F.2d 1140, 114647 (2d Cir. 1991), for example, the court found that "[t]his showing may be made through statistical evidence revealing a disparity so great that it cannot reasonably be attributed to chance."[49] Even more directly, the EEOC guidelines on employee selection state that "adverse impact may be inferred where, assuming not too small a sample, the members of a minority group are selected at a rate that is less than four-fifths of the rate at which the majority group is selected."[50] The EEOC guidelines effectively identify what the parameter values should be in the adaptive model: namely, $a = 4$ and $b = 1$.[51]

Another area where courts consistently admit statistical evidence is, as I mentioned above, forensic base rates in the form of DNA or fingerprint profiles. At least in the case of DNA, such

[49] See also *Hazelwood School District v. United States*, 433 U.S. 299, 307-08 (1977) ("gross statistical disparities ... may in a proper case constitute prima facie proof of a pattern or practice of discrimination"); *Castaneda v. Partida*, 430 U.S. 482, 496-97 (1977) (using analysis of variance (ANOVA) to make an inference about the underlying practice); *Bazemore v. Friday*, 478 U.S. 385, 400-01 (U.S. 1986) (noting that an inference based on linear regression may satisfy the preponderance standard); *Smith v. Liberty Mut. Ins. Co.*, 569 F.2d 325, 329 (5th Cir. 1978) ("This Court has always recognized the strong probative value of statistics in proving race discrimination cases.").

[50] EEOC Uniform Guidelines on Employee Selection Procedures, 29 C.F.R. §1607.4D.

[51] But perhaps you are suspicious that what courts have in mind here is statistical evidence put forth precisely in support of the causal element. To be sure that is indeed the case, consider *Watson v. Fort Worth Bank & Trust*, 487 U.S. 977, 995 (1988), where the court notes that to establish a prima facie case, statistical disparities "must be sufficiently substantial that they raise ... an inference of causation." See also *E.E.O.C. v. Joint Apprenticeship Comm. of Joint Indus. Bd. of Elec. Indus.*, 186 F.3d 110, 117 (2d Cir. 1999) ("a plaintiff may establish a prima facie case of disparate impact discrimination by proffering statistical evidence which reveals a disparity substantial enough to raise an inference of causation."). It is pretty clear from *Watson* and its progeny that statistical evidence *may* support a finding of causation, which is what opponents of statistical evidence often categorically deny. See e.g., Wright (1988) (arguing incorrectly that statistical evidence could not be evidence of causation).

evidence is almost uniformly held to be admissible.[52] Koehler (2002)'s explanation of this is that the evidence is offered to rebut a chance hypothesis (i.e., getting a DNA match by chance would be extraordinarily unlikely). I suspect, rather, that courts' comfort with DNA evidence has more to do with its extremely high probability than with the fact that the alternative explanation would be chance. Indeed, there exists an alternative chance explanation in every dispute, legal or otherwise. Fortunately we now have a better diagnosis. DNA evidence is usually deemed admissible because the prior probability of H is so high that the discrepancy between a and b, as we saw, would have to be patently unreasonable to cancel out the prior odds. The adaptive model predicts the admissibility of DNA evidence by simply putting some obvious bounds on the rationality of different attitudes to risk. For example, by assuming that $1,000,000b > a$.[53]

Perhaps the strongest support of my hypothesis that what really matters is the relationship between the posterior probability and the ratio of error rates may be found in *Kaminsky v. Hertz Corp.*, 288 N.W.2d 426 (Mich. Ct. App. 1980). In that case, the plaintiffs sustained personal injuries when their car was struck by a large piece of ice that fell from the top of a yellow truck bearing the distinctive Hertz logo. The plaintiffs put forward evidence showing that Hertz owned 90% of Hertz labeled yellow trucks. Now this might remind you of *Smith* and its stylized versions, as presented in problems 1-4. If it does, you're not incorrect – it is because *Hertz* is almost exactly like *Smith*. In *Hertz*, however, the appellate court ruled that the 90% base rate was not only relevant evidence for the plaintiff, but that it established a rebuttable presumption of ownership sufficient to preclude summary judgment for the defendant. This case is rarely mentioned in the literature spawned by *Smith*, even though it suggests the exact opposite conclusion than what many philosophers and legal scholars want to draw on the basis of *Smith*.

One might suspect that I must be cherry-picking in highlighting *Hertz*, but it is far from an outlier. For example, in *Kramer v. Weedhopper of Utah, Inc.*, 490 N.E.2d 104 (Il. App. Ct. 1986) (quoted in Koehler (2002)), the plaintiff was injured by a bolt from Weedhopper's model aircraft kit. It was shown in court that Weedhopper received its bolts from two companies – 90% from Lawrence and 10% from Hughes. On this basis, an Illinois appellate court reversed a trial court's summary judgment in favor of Lawrence, arguing that "[t]his evidence, while circumstantial, permits the inference that the . . . [bolts] supplied to Kramer were purchased from Lawrence" (105-108).

From Cheng's perspective, there is no way to capture cases like *Hertz* and *Kramer*. If preponderance requires setting the prior odds to 1 then what happened here? On the adaptive model, not only can we accommodate *Hertz* and *Kramer* but we can explain the discrepancy between *Hertz*/*Kramer*, on the one hand, and stylized versions of *Smith*, on the other. Suppose that $a = 4b$, a plausible trade-off between the competing costs. On this conjecture, a decision maker would indeed reject the Plaintiff's claim in all stylized versions of *Smith* while accepting a prima facie case in *Hertz* and *Kramer*. Indeed, there is a real case closer to stylized versions of *Smith* than *Smith* itself, namely *Guenther v. Armstrong Rubber Company*, 406 F.2d 1315 (3d Cir. 1969), where the court ruled in favor of the defendant's motion for summary judgment despite evidence that the defendant manufactured 75-80% of tires sold at the Sears store where the plaintiff purchased her defective tires. With $a = 4b$, this is exactly what we would expect. We can understand both pairs of judgments which seem, initially, to be completely at odds, by simply considering what kind of attitude to risk might be reflected by the decision makers' choices in these situations. And there is a perfectly acceptable

[52] Nat'l Research Council, Committee on Forensic Data Technology: An Update, The Evaluation of Forensic DNA Evidence 185 (1996). See also Zabell (2005) for a helpful overview, including a discussion of the difference between the probative value of DNA evidence, on the one hand, and fingerprints, on the other.

[53] I avoid taking a stand here on an underlying theory of practical rationality, but any theory that has as its consequence that $a \geq 1,000,000b$ should be treated as suspect.

attitude that accommodates both pairs of cases, namely, $a = 4b$.[54]

The last area I want to highlight is mass exposure cases, where courts have embraced base rates in part due to necessity – that is, because direct evidence is often unavailable. In *In re Agent Orange Prod. Liab. Litig.*, 597 F. Supp. 740, 835-836 (E.D.N.Y. 1984), Judge Weinstein provides a very sophisticated discussion of statistical evidence and its relationship to the preponderance standard. The issue in that case was whether plaintiff Vietnam war veterans could use market share data as evidence of likelihood that a particular chemical manufacturer produced the deadly Agent Orange herbicide (used pervasively by the U.S. military as part of its herbicidal warfare program during the Vietnam War) that caused their injuries. The court distinguishes two versions of the preponderance rule. A strong all-or-nothing version, under which statistical evidence alone is not sufficient for identity, and a weak version, which "would allow a verdict solely on statistical evidence" (835).[55] Judge Weinstein then explains that while there would "appear to be little harm in retaining the requirement for 'particularistic' evidence of causation in sporadic accident cases" where "such evidence is almost always available," in mass exposure cases, "where the chance that there would be particularistic evidence is in most cases quite small, [and] the consequence of retaining the requirement might be to allow defendants who, it is virtually certain, have injured thousands of people and caused billions of dollars in damages, to escape liability" the 'weak' version of the preponderance rule "appears to be the preferable standard to apply." What Judge Weinstein calls the weak standard has been applied in a number of mass exposure cases including, most notably, in the formulation of the market share liability doctrine for DES manufacturers.[56]

Judge Weinstein's discussion is extremely important to my argument. It is not simply that the discussion is compatible with the adaptive model I propose. Rather, he articulates the very concerns that prompted me to develop such a model. Whether or not statistical evidence is appropriate, Judge Weinstein suggests, depends on the underlying factual circumstances, including what is at stake and whether alternative methods of proof are available.

We can think about the adaptive model I develop as a generalized version of Judge Weinstein's approach from *In re Agent Orange*. What he does is, first, distinguish two evidential interpretations of the preponderance rule – a strong rule and a weak rule – and, second, argue that which of the two applies depends on the costs of error in the case at hand. I generalize this by having the parameters a and b transform the rule into a continuum, from the strongest to the weakest, where the relative strength is determined by the factual circumstances of each case.

6 Concerns and Objections

In this section I consider some potential concerns and objections. First, I explain the difference between Kaplow's welfare based notion of optimality and my accuracy based notion of optimality, as it is important not to confuse the two approaches. Second, I articulate the difference between an elicitation model of the burden of proof and a decision rule for legal fact finding. And third, I use a principal-agent framework to argue that the adaptive model, properly understood, is compatible with the general subjective expected utility framework of Savage (1954) or von Neumann & Morgenstern (1944).

6.1 Social welfare and epistemic risk

Like the adaptive model, Kaplow (2014)'s approach is similarly flexible in that his decision procedure enjoins a judge or juror to compare the ratio of posterior probabilities to a ratio of losses to

[54] Compare, for example, Buchak (2014)'s diagnosis. Buchak argues that the conclusion to draw from *Smith* is that legal judgments require a *belief*, which is an altogether different doxastic attitude from a posterior probability, and indeed that there is no probabilistic threshold above which we can say the posterior constitutes a belief. But such a diagnosis, like Cheng (2013)'s, cannot make sense of cases like *Hertz* and *Kramer* together with those like *Smith* and *Guenther*.

[55] The all-or-nothing version is not an inherent component of the preponderance rule and has not been thought of as such for decades. See C. McCormick, McCormick's Handbook of the Law of Evidence §31 at 118 (1935).

[56] See *Sindell v. Abbott Labs.*, 26 Cal. 3d 588 (1980) (developing the notion of market share liability).

gains. However, the only considerations permitted in Kaplow's model are those that could affect the individual utilities and in turn the social welfare function. This is made explicit in Kaplow (2011). Now suppose you believe, as many legal scholars do, that falsely punishing someone on the basis of their membership in a reference class alone constitutes a moral wrong that cannot be reduced to its impact on individual utilities.[57] This is a cost that cannot enter into Kaplow's decision model. He is explicit about this because any weighting that is not reflected in the individual utilities implies a non consequentialist normative objective that could lead to outcomes which are in conflict with the Pareto Principle (Kaplow & Shavell, 2001). For opponents of social welfare, however, this begs the question.[58] Their main point is that the Pareto Principle and more generally the social welfare approach fail to capture salient moral considerations. This disagreement is part of a broader debate about the moral foundations of legal institutions (Kaplow & Shavell, 2006).

Fortunately, the adaptive model enables us to sidestep this debate. I want to capture how people actually make decisions and undoubtedly some people do so by taking into account considerations irreducible to welfare. For example, Diamond & Vidmar (2001) describe videotaped jury deliberations in negligence disputes containing frequent references to plaintiffs insurance coverage and attorney fee arrangements, as part of a broader concern for whether the plaintiff is made whole or treated fairly. In particular, I want to understand how if at all legal decision makers – including those whose substantive normative views differ from Kaplow and Shavell's, such as some of the subjects described in Diamond & Vidmar (2001) – could take high posterior probabilities to be insufficient for liability (as in problems 2-4) without being epistemically irrational. The adaptive model shows that provided you agree we should minimize a linear combination of error rates, high posterior probabilities could be insufficient if the decision maker is correspondingly risk averse. Therefore, I offer a well epistemically motivated template that helps us to understand legal choice behavior regardless of the decision maker's underlying normative commitments.

6.2 Elicitation models and decision rules

Kaplow proposes the optimal social welfare model as a decision rule. In Kaplow (2012), for example, he considers explicitly how we might incorporate considerations of social welfare into burden of proof rules, including a discussion of how we might reformulate jury instructions to make ex ante considerations more salient. Unlike Kaplow I do not propose the adaptive model as a decision rule. That is, I do not argue that we should instruct judges and juries on how to use the adaptive model in order to improve legal decision making. My approach is descriptive and the model I propose is attitude eliciting.

I propose the model as a way of learning something about what decision makers value when they decide the way they do. Judges and juries will probably not apply Bayes' Theorem explicitly, and they will probably not explicitly consider their relative preference for avoiding Type I and Type II error rates. But the decision they ultimately make tells us something important about their attitudes to risk of error – that is, it enables us to learn something about their relative assessment of the relevant epistemic costs. In this sense, the model elicits or reflects the decision maker's underlying values.

In statistical decision theory, we are often interested in estimating a decision maker's subjective probability. Following De Finetti (1937) and Savage (1971), it is common to assume that subjective probabilities are marginal rates of substitution between contingent claims. To operationalize this idea, scoring rules are used to convert an agent's forecast into a lottery. For example, under the common quadratic score, a report of p would lead to a payoff that is some monotonic function f of the quadratic distance of p from the true outcome, which is $(1-p)^2$ if the outcome occurs and p^2 if it does not.

[57] See e.g., Tribe (1971), Wasserman (1991) and more generally Colyvan et al. (2001).

[58] See e.g., Ferzan (2004) ("[Kaplow and Shavell] define fairness as principles that do not advance welfare. They then walk the reader through hypotheticals to demonstrate that fairness, so defined, does not advance welfare. But what does this ... unabashed tautology ... prove? My dog will always be better than your cat, if the test is whose pet can bark.").

By evaluating pairs of lotteries that an agent is indifferent between, we can infer what her subjective probabilities should be.[59]

There are two ways to interpret the elicitation exercise. On the more extreme interpretation, subjective probabilities are nothing more than the observable behavior they are correlated with. To have a belief of .5 in a coin's bias toward Heads, on this interpretation, just is to be indifferent between receiving $1 for sure and taking a bet that pays $0 on Heads and $2 on Tails on a single toss of the coin. Ramsey (1926) comes close to this extreme. On a less behaviorist interpretation, observable behavior provides us with an imperfect clue about the true underlying doxastic attitude.

In either case, however, the inference we make from observable behavior to the underlying belief will be precise only if we assume the agent is risk neutral. For example, if an agent declines to pay $1 for a bet that pays $0 on Heads and $2 on Tails on a single coin toss, it might be either because she believes that the coin is Heads biased or because her utility function is concave so that the expected utility of the bet is lower than the utility of the sure thing. Risk attitudes interfere with our ability to discern underlying beliefs.

As a result, a common simplifying assumption in the elicitation literature is to assume the agent is risk neutral. By screening off risk, we can draw precise inferences about belief. In reality, the best we can expect is something like an interval based estimate about an agent's beliefs bounded by the information we have about her degree of risk aversion.

My approach in this paper reverses this process. By assuming that agents update their probabilities efficiently by applying Bayes Theorem we are able to learn something about their attitudes to risk in the context of legal decision making. So in Problem 3, for example, when an agent declines to find the defendant liable, where the prior odds are $4:1$, it may be either because her error rates are equal to or greater than $1:4$, or because her subjective posterior odds are less than 4 (i.e., she has failed to some extent to update correctly).

The efficient updating assumption is a simplifying one, and by taking into account the extent of the agent's dynamic incoherence we would get at most an imprecise interval for the values they assign to a and b. For example, it may be that given our best estimate of the divergence of the agent's posterior from the correct Bayesian posterior in Problem 3, $3.5 < a < 4.5$. In subsequent research, it would be interesting to develop a finer grained model that considers decision making under imperfect updating or perhaps even under probabilistic incoherence.

The adaptive model also makes several empirically verifiable predictions. If I am correct we should expect a strong correlation between people's sensitivity to risk of error and their responses to hypothetical cases involving statistical evidence. Further, because I suggest that the risk parameters will be context sensitive, we should expect variation in responses to statistical evidence as we change the underlying factual circumstances (from, say, mass exposure class actions to slip and fall cases).[60] It would be worth directly testing these predictions in subsequent empirical work as a way of learning how attitudes to risk of error affect legal decision making. If the adaptive model is correct, it would help us understand why there is so much variation among judges and juries in understanding burdens of proof, as reported in McCauliff (1982), for example. Since decision makers vary widely in their attitudes to risk, if the adaptive model is correct it should not be a surprise that their interpretations of evidentiary standards are correspondingly variable. There is some promising preliminary experimental evidence on the effect of loss aversion, a relative to risk aversion, to legal decision making that strongly supports the adaptive model (Ritov & Zamir, 2012). In subsequent research, it would be worthwhile to put the model to a more direct empirical test.

[59] See generally Winkler & Murphy (1968), Winkler (1969), Winkler (1996), Gneiting & Raftery (2007), Kadane & Winkler (1987).

[60] Current empirical evidence indirectly supports this conjecture. In addition to McCauliff (1982), discussed in §5.2, *supra*, Solan (1999) describes a wide range of probabilities that juries associate with different forms of the "beyond a reasonable doubt" jury instruction, as it varies from context to context.

6.3 The principal-agent choice environment

One might worry that Royall's trichotomy, and in turn my approach here, is fundamentally anti-Bayesian. From Savage (1954)'s perspective, the answers to Royall's three questions are not separable in the way I have separated them here. For example, what you should do depends on what you believe and what you believe depends in part on what we assume you value which means that what the evidence says depends in part on both our assumptions about your beliefs and how you value outcomes. And I certainly do not want to stake out a position here that is incompatible with Savage's approach.

However, the legal context is not an ordinary decision making context and I think it is especially appropriate for drawing Royall's distinction in a way that is not incompatible with the general Bayesian decision making framework. The adaptive model exists in what we may perhaps helpfully call a principal-agent (P-A) environment of choice. The basic idea is that we are often in a position of having to choose, as principals, on behalf of someone else, the agent. These contexts vary in the scope of the principal's authority. On one extreme, we have cases where the agent delegates so much of the decision process that the principal effectively uses her own preferences in place of the agent's. So, for example, a wealthy art patron with little understanding of aesthetic value may hire a curator and tell her "find me something good." In this case, the curator uses her own preferences about what makes good art. On the other extreme, the principal is forced to substitute her preference for someone else's. Suppose we are meeting for dinner and I am running late. I may say, "please order me a nice seafood meal." You might hate seafood, but you still have to try and place yourself in the shoes of someone that likes seafood and identify a preference ranking over the available meals from their perspective.

The nice thing for us about the P-A environment is that it makes room for a variety of attitudes to risk in the context of Bayesian expected utility optimization. For example, as a hedge fund manager, a client may tell you: "I only care about my exposure to loss and I request that you rank investment decisions on that basis alone." Your own decision making is still governed by maximizing expected utility, but when it comes to decisions on behalf of this particular client, the way to maximize expected utility is to rank options exclusively on the basis of her exposure to loss.

In the context of legal decision making, the judge or jury is the principal and the agents are, collectively, the group of people bound by the institution. The cost parameters, then, should be evaluated by reference to whether those bound by the institution (the agents) would find them appropriate. So, again, each individual decision maker is going to choose however they choose. They probably will not apply Bayes' Theorem, and they probably will not explicitly attempt to maximize expected utility. But we can represent them as if they were doing so. This is what the expected utility theorem enables us to do. And we can evaluate their individual or collective decisions by considering the values they reflect. This is what the adaptive model enables us to do. As a result, thinking about the adaptive model as embedded in a P-A choice environment brings together each of its features: (**i**) it is Bayesian; (**ii**) it is compatible with expected utility theory; (**iii**) it is flexible; (**iv**) it is preference eliciting; and (**v**) it does not presuppose that legal decision making takes place in epistemic heaven.

7 Concluding Remarks

In this article I have developed an adaptive model of the burden of proof as a true Bayesian hypothesis test under which every decision is governed by a comparison of posterior odds to a rejection threshold determined by the ratio of error costs. As I said, this does not mean that that is how legal decision makers actually approach a choice problem. Rather, this gives us a helpful way of framing the legal decision making process. We can better understand legal fact finding by modeling our decision makers as if they were applying the adaptive model. When they appear to ignore strong statistical evidence, for example, the conclusion we draw is not that they are epistemically irrational but rather that they are highly risk sensitive. As a result, we may also apply the adaptive model for the normative

assessment of legal decisions, by attending to the particular values to risk they elicit and considering their reasonableness. Finally, the model may be used as a tool for predicting the resolution of future disputes. To make a prediction we use our best judgment to estimate from the circumstances what the relative cost parameters might be. This is not as difficult as may first appear. As we saw above, the plausible conjecture that $a = 4b$ explains much of the relevant case law. Moreover, our estimate does not need to be precise. It is usually enough to guess an inequality.

Finally, my approach is compatible with proposals like Rosenberg (1984)'s for extending the proportionality approach to civil liability from the very specific DES context for which market share liability was initially fashioned to mass exposure cases more generally, including harmful chemicals like Asbestos, Agent Orange, Tobacco, PCB, PBB, BPA, etc., and pharmaceuticals and medical devices like DES, Vioxx, silicone breast implants and many others. The only addition we would need to make to the adaptive model is to make the proportion of recovery a positive linear function of the posterior odds. If we want true proportionality (rather than discrete cut offs) that function should be continuous.

References

Babic, Boris. 2017. "Generalized Entropy and Epistemic Risk." *Draft.*

Bar-Hillel, Maya. 1980. "The base-rate fallacy in probability judgments." *Acta Psychologica*, vol. 44 (3): 211–233.

Brier, Glenn W. 1950. "Verification of Forecasts Expressed in Terms of Probability." *Monthly Weather Review*, vol. 78 (1): 1–3.

Brook, James. 1985. "The Use of Statistical Evidence of Identification in Civil Litigation: Well Worn Hypotheticals, Real Cases, and Controversy." *St. Louis University Law Journal*, vol. 29 (2): 293–352.

Buchak, Lara. 2014. "Belief, Credence, and Norms." *Philosophical Studies*, vol. 169 (2): 285–311.

Cheng, Edward K. 2013. "Reconceptualizing the Burden of Proof." *Yale Law Journal*, vol. 122 (5): 1254–1279.

Cheng, Edward K. & Michael S. Pardo. 2015. "Accuracy, Optimality, and the Preponderance Standard." *Law, Probability & Risk*, vol. 14 (3): 193–212.

Cohen, Jonathan. 1981. "Subjective Probability and the Paradox of the Gatecrasher." *Arizona State Law Journal*, vol. 1981 (2): 627–634.

Colyvan, Mark, Helen M. Regan & Scott Ferson. 2001. "Is It a Crime to Belong to a Reference Class?" *Journal of Political Philosophy*, vol. 9 (2): 168–181.

De Finetti, Bruno. 1937. "La prévision: ses lois logiques, ses sources subjectives." *Annales de l'institut Henri Poincaré*, vol. 7 (1): 1–68.

DeGroot, Morris H. & Mark J. Schervish. 2012. *Probability and Statistics.* New York: Wiley, fourth edn.

Demougin, Dominique & Claude Fluet. 2008. "Rules of Proof, Courts, and Incentives." *The RAND Journal of Economics*, vol. 39 (1): 20–40.

Diamond, Shari Seidman & Neil Vidmar. 2001. "Jury Room Ruminations on Forbidden Topics." *Virginia Law Review*, vol. 87: 1857–1915.

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

Ferzan, Kimberly Kessler. 2004. "Some Sound and Fury from Kaplow and Shavell." *Law & Philosophy*, vol. 23 (1): 73–102.

Gneiting, Tilmann & Adrian E. Raftery. 2007. "Strictly Proper Scoring Rules, Prediction, and Estimation." *Journal of the American Statistical Association*, vol. 102 (477): 359–378.

Greaves, Hilary & David Wallace. 2006. "Justifying Conditionalization: Conditionalization Maximizes Expected Epistemic Utility." *Mind*, vol. 115 (459): 607–632.

Hacking, Ian. 1965. *Logic of Statistical Inference*. Cambridge: Cambridge University Press.

Hay, Bruce & Kathryn E. Spier. 1997. "Burdens of Proof in Civil Litigation: An Economic Perspective." *The Journal of Legal Studies*, vol. 26 (2): 413–431.

Kadane, Joseph B. & Robert L. Winkler. 1987. "de Finetti's Method of Elicitation." In *Probability and Bayesian Statistics*, R. Viertl, editor. New York: Plenum.

Kahneman, Daniel & Amos Tversky. 1972. "On Prediction and Judgment." Tech. Rep. 12(4), Oregon Research Institute Bulletin.

—. 1982. *Judgment under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.

Kaplan, John. 1968. "Decision Theory and the Factfinding Process." *Stanford Law Review*, vol. 20 (6): 1065–1092.

Kaplow, Louis. 1994. "The Value of Accuracy in Adjudication: An Economic Analysis." *The Journal of Legal Studies*, vol. 23 (1): 307–401.

—. 2011. "On the Optimal Burden of Proof." *Journal of Political Economy*, vol. 119 (6): 1104–1140.

—. 2012. "Burden of Proof." *Yale Law Journal*, vol. 121 (4): 738–859.

—. 2014. "Likelihood Ratio Tests and Legal Decision Rules." *American Law and Economics Review*, vol. 16 (1): 1–39.

Kaplow, Louis & Steven Shavell. 2001. "Any Non-Welfarist Method of Policy Assessment Violates the Pareto Principle." *Journal of Political Economy*, vol. 109 (2): 281–286.

—. 2006. *Fairness versus Welfare*. Cambridge: Harvard University Press.

Koehler, Jonathan J. 2002. "When Do Courts Think Base Rate Statistics are Relevant?" *Jurimetrics*, vol. 42 (4): 373–402.

Laudan, Larry & Ronald J. Allen. 2008. "Deadly Dilemmas." *Texas Tech Law Review*, vol. 41 (1): 65–93.

Lehmann, Erich L. & Joseph P. Romano. 2005. *Testing Statistical Hypotheses*. New York: Springer.

Lindley, Dennis V. 1982. "Scoring Rules and the Inevitability of Probability." *International Statistical Review/Revue Internationale de Statistique*, vol. 50 (1): 1–11.

Lipsey, R.G. & Kelvin Lancaster. 1956. "The General Theory of Second Best." *The Review of Economic Studies*, vol. 24 (1): 11–32.

McCauliff, C.M.A. 1982. "Burdens of Proof: Degrees of Belief, Quanta of Evidence, or Constitutional Guarantees." *Vanderbilt Law Review*, vol. 35 (6): 1293–1336.

Merkle, E. C., M. Steyvers, B. Mellers & P. E. Tetlock. 2016. "Item Response Models of Probability Judgments: Application to a Geopolitical Forecasting Tournament." *Decision*, vol. 31 (1): 1–19.

Miceli, Thomas J. 1990. "Optimal Procecution of Defendants Whose Guilt is Uncertain." *Journal of Law, Economics, & Organization*, vol. 6 (1): 189–201.

Nesson, Charles. 1985. "The Evidence or the Event? On Judicial Proof and the Acceptability of Verdicts." *Harvard Law Review*, vol. 98 (7): 1357–1392.

—. 1986. "Agent Orange Meets the Blue Bus: Factfinding at the Frontier of Knowledge." *Boston University Law Review*, vol. 66 (4): 521–539.

von Neumann, John & Oskar Morgenstern. 1944. *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.

Neyman, J. & E.S. Pearson. 1933. "On the Problem of the Most Efficient Tests of Statistical Hypotheses." *Philosophical Transactions of the Royal Society of London. Series A*, vol. 231: 289–337.

Posner, Richard. 1999. "An Economic Approach to the Law of Evidence." *Stanford Law Review*, vol. 51 (6): 1477–1546.

Ramsey, Frank Plumpton. 1926. *Truth and Probability*. In *The Foundations of Mathematics and other Logical Essays*, ed. R.B. Brathwaite. New York: Harcourt Brace, 1931: pp. 156-198 (1999 Electronic Edition).

Redmayne, Mike. 2008. "Exploring the Proof Paradoxes." *Legal Theory*, vol. 14 (4): 281–309.

Ritov, Ilana & Eyal Zamir. 2012. "Loss Aversion, Omission Bias, and the Burden of Proof in Civil Litigation." *The Journal of Legal Studies*, vol. 41 (1): 165–207.

Rosenberg, David. 1984. "The Causal Connection in Mass Exposure Cases: A 'Public Law' Vision of the Tort System." *Harvard Law Review*, vol. 97 (4): 849–929.

Royall, Richard. 1997. *Statistical Evidence: A Likelihood paradigm*. London: Chapman & Hall.

Rubinfeld, Daniel L. & David E.M. Sappington. 1987. "Efficient Awards and Standards of Proof in Judicial Proceedings." *The RAND Journal of Economics*, vol. 18 (2): 308–315.

Savage, Leonard J. 1954. *The Foundations of Statistics*. New York: Dover.

—. 1971. "Elicitation of Personal Probabilities and Expectations." *Journal of the American Statistical Association*, vol. 66 (336): pp. 783–801.

Schauer, Frederick. 2003. *Profiles, Probabilities, and Stereotypes*. Cambridge: Harvard University Press.

Schervish, Mark J. 1989. "A General Method for Comparing Probability Assessors." *The Annals of Statistics*, vol. 17 (4): 1856–1879.

Sober, Elliott. 2008. *Evidence and Evolution: The Logic Behind the Science*. Cambridge: Cambridge University Press.

Solan, Lawrence M. 1999. "Refocusing the Burden of Proof in Criminal Cases: Some Doubt About Reasonable Doubt." *Texas Law Review*, vol. 78: 105–148.

Spanos, Aris. 2013. "Who Should Be Afraid of the Jeffreys-Lindley Paradox?" *Philosophy of Science*, vol. 80 (1): 73–93.

Tetlock, Philip E., Orie V. Kristel, S.B. Elson, Melanie C. Green & Jennifer S. Lerner. 2000. "The Psychology of the Unthinkable: Taboo Trade-Offs, Forbidden Base Rates, and Heretical Counterfactuals." *Journal of Personality and Social Psychology*, vol. 78 (5): 853–870.

Thomson, Judith Jarvis. 1986. "Liability and Individualized Evidence." *Law & Contemporary Problems*, vol. 49 (3): 199–219.

Tribe, Laurence H. 1971. "Trial by Mathematics: Precision and Ritual in the Legal Process." *Harvard Law Review*, vol. 84 (6): 1329–1393.

Wald, Abraham. 1945. "Statistical decision functions which minimize the maximum risk." *The Annals of Mathematics*, vol. 46 (2): 265–280.

Wasserman, David. 1991. "The Morality of Statistical Proof and the Risk of Mistaken Liability." *Cardozo Law Review*, vol. 13 (2-3): 935–977.

Wells, Gary. 1992. "Naked Statistical Evidence of Liability: Is Subjective Probability Enough?" *Journal of Personality & Social Psychology*, vol. 62 (5): 739–752.

Winkler, Robert L. 1969. "Scoring Rules and the Evaluation of Probability Assessors." *Journal of the American Statistical Association*, vol. 64 (327): 1073–1078.

—. 1996. "Scoring Rules and the Evaluations of Probabilities." *Test*, vol. 5 (1): 1–60.

Winkler, Robert L. & Allan H. Murphy. 1968. "'Good' Probability Assessors." *Journal of Applied Metereology*, vol. 7 (5): 751–758.

Wright, Richard. 1988. "Causation, Responsibility, Risk, Probability, Naked Statistics, and Proof: Pruning the Bramble Bush by Clarifying the Concepts." *Iowa Law Review*, vol. 73 (5): 1001–1077s.

Zabell, Sandy L. 2005. "Fingerprint Evidence." *Journal of Law and Policy*, vol. 13 (1): 143–179.