Foundations of Epistemic Risk

by

Boris Babic

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Philosophy)
in the University of Michigan
2017

Doctoral Committee:

Professor James M. Joyce, Chair
Professor Richard D. Gonzalez
Professor Peter A. Railton
Professor Brian James Weatherson

Boris Babic
bbabic@umich.edu
ORCID iD: 0000-0003-2800-1307

2017

ACKNOWLEDGMENTS

I would like to thank my dissertation chair, Jim Joyce, whose patience, mentorship and support made this project possible. Jim has been an extraordinary advisor.

I would also like to thank the rest of my dissertation committee – Rich Gonzalez, Peter Railton, and Brian Weatherson – for their continued guidance.

Thanks to the graduate students at Michigan for being incredible colleagues.

Thanks to Nevena, to my brother Milan, and to my parents Danka and Pajo for their constant support.

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

My goal in this dissertation is to start a conversation about the role of risk in the decision-theoretic assessment of partial beliefs or credences in formal epistemology. I propose a general theory of epistemic risk in terms of relative sensitivity to different types of graded error. The approach I develop is broadly inspired by the pragmatism of the American philosopher Charles Sanders Peirce and his notion of the "economy of research." I express this framework in information-theoretic terms and show that epistemic risk, so understood, is dual to information entropy. As a result, every unit increase in risk comes with a corresponding unit decrease in information entropy and epistemic risk may be expressed in terms of entropic change. I explain the significance of this for the selection of priors and the Laplacian principle of indifference. I also extend this notion of epistemic risk to the assessment of updating rules, where a similar duality between risk and information holds. In the dynamic context, epistemic risk is given by *cross*-entropic change. Here I explore the relationship between risk, the Value of Knowledge Theorem, dynamic coherence, and the role of expected accuracy in the selection of update rules. Finally, I apply these considerations to a social institution where attitudes to error are especially salient – namely, legal decision-making – and argue that considerations regarding the relative severity of different types of error are central to understanding evidentiary burdens of proof and the probative value of statistical evidence.

CHAPTER 1

Introduction

Riskiness and attitudes to risk play a substantial role in ordinary rational choice. For instance, the notion of risk aversion, captured in the standard framework in terms of diminishing marginal utility, is central to evaluating investments, understanding insurance, and more generally shaping social policy. Recently, philosophers have applied rational choice approaches to epistemology, using a decision-theoretic framework to understand and evaluate how we form our beliefs and revise them in light of new information. This theoretical framework is known as *epistemic utility theory*. My goal in the substantive chapters of this dissertation is to initiate a conversation about the role of risk in this framework. I will motivate several important questions about epistemic risk, develop a general framework for understanding and measuring epistemic risk, and apply some of these notions to a social institution where attitudes to error are especially salient – namely, legal decision-making. The approach I develop is broadly inspired by the pragmatism of the American philosopher Charles Sanders Peirce and his notion of the "economy of research." The more immediate influences on this work are L.J Savage and E.T. Jaynes – the substantial influence of Savage's work on the elicitation of subjective probabilities and Jaynes's information-theoretic understanding of statistical inference will be obvious. What follows is a brief overview.

Chapter 1 is broadly introductory. I motivate the general notion of epistemic risk by drawing on C.S. Peirce's economic approach to scientific inquiry and I situate the project within the contemporary epistemic utility framework. I explain what I mean by epistemic risk – i.e., what exactly is at risk when an agent adopts one set of beliefs instead of another – and I contrast my approach to existing literature on this topic (in particular, Duncan Pritchard's modal analysis of epistemic risk).

Chapter 2 will answer the following questions: What does it mean for one probability distribution to be riskier than another? In particular: What makes it riskier? How do we measure such risk? And how does risk relate to other properties of a probability distribution? The considerations in this chapter are all static. There is no temporal dimension yet,

and we set aside learning and evidence-gathering for the moment. From a Bayesian perspective, this chapter is about the riskiness of an agent's prior beliefs. Meanwhile, Chapter 3 is dynamic. Here I seek to answer the following questions: After an agent receives some information, what does it mean for one update rule, or one posterior probability, to be riskier than another? How do we measure such dynamic epistemic risk? And how does it relate to our static measure of epistemic risk, as developed in Chapter 2?

Finally, Chapter 4 applies considerations of epistemic risk to legal decision-making. In a nutshell, I argue that considerations regarding the relative severity of different types of error are central to understanding legal burdens of proof.

CHAPTER 2

The Economy of Inquiry

2.1 Introduction

I begin by drawing on the work of Charles Sanders Peirce and his notion of the "economy of research" to motivate the basic idea that attitudes to risk of error play an important role in scientific inquiry. Attitudes to risk of error are to be distinguished from Frequentist error probabilities. It is best to think about the notion of epistemic risk I will develop here by analogy to ordinary risk in economic theory: just as ordinary riskiness has something to do with the range of monetary outcomes in a gamble, so epistemic riskiness will have much to do with the range of accuracy outcomes in deciding what to believe. This is why I develop the framework by drawing on the Peircian notion of the economy of research and its central role in inference, rather than focusing on the ordinary error probabilities in the [Neyman and Pearson(1933)] paradigm. Next, I give an overview of epistemic utility theory, which is an application of ordinary rational choice theory to epistemology, and situate my project within it. I also give a brief overview of the role of risk in ordinary rational choice. Finally, I compare my approach to the existing literature on epistemic risk that has been developed by, for example, [Pritchard(2017)].

2.2 C.S. Peirce and the truth-seeking economist

As [Rescher(1976)] puts it, Peirce "gave the place of pride to a theory – indeed a discipline – of his own devising, namely to what he called the *economy of research*." Indeed, he adds, "to this idea of the economy of research . . . Peirce gave as central a place in his methodology of science as words can manage to assign." Unfortunately, the notion has been completely neglected by subsequent developments in the philosophy of science.[1]

[1] "no other part of this great man's philosophizing has fallen on stonier ground" [Rescher(1976)]. This remains true today. Even the Stanford Encyclopedia of Philosophy entry on Peirce, while paying lip service

Peirce concisely describes the basic idea as follows,

> The doctrine of economy, in general, treats of the relations between utility and cost. That branch of it which relates to research considers the relations between the utility and the cost of diminishing the probable error of our knowledge." [Peirce(1879), 643]

The 'probable error of our knowledge' is a treacherous expression. In my framework it will be understood in a very specific way and one that is different from how Peirce would have understood it at the turn of the twentieth century. For Peirce, the probability of error is to be understood roughly the way we would use a confidence interval today, and he goes on to highlight the increasing sampling cost of marginal improvements in its precision.

The idea of characterizing scientific inquiry in terms of error costs was of course carefully developed by [Neyman and Pearson(1933)], and in the philosophy literature the cost of false acceptance/rejection was explored by Kyburg, Levi and others in the context of binary theories of acceptance and belief [Levi(1974), Kyburg(1974)]. This is to be expected, as there is a close relationship between a hypothesis test and a confidence interval. The modern Frequentist confidence interval consists of all possible values of the unknown parameter under which the null would not be rejected. Rather than identifying all such values, it is generally possible to invert the equation for the test statistic instead to find the boundary points of the confidence interval. Since the significance threshold is typically set, for better or for worse, by considering what we would consider to be tolerable observed false positive and false negative error rates, Peirce's use of probable error is quite close to the error probabilities that emerged at the forefront of null hypothesis significance tests.

I will not use probable error quite this way, however. I build my approach on the central insight that gradational attitudes to the possibility of being mistaken, and the severity of the mistake should one be mistaken, affect the course of scientific inquiry. These are the central phenomena that constitute what I will identify as 'epistemic risk'. I will explain these notions carefully, below. The central difference is that the kind of error I am interested

to the importance Peirce placed on economic considerations to scientific inquiry, fails to explain *how* those considerations were supposed to affect the course of inquiry – their specific role in abduction, deduction, and induction, their influence on Peirce's thinking on probability, and his skepticism of so-called inverse inference (what we now call Bayesian inference). The entry states, for example, that with respect to null hypothesis significance testing, Peirce had "worked out the whole matter" before Neyman and Pearson. While the extent to which Peirce anticipated null hypothesis significance testing is extremely impressive, he obviously had not worked out the whole matter. One cannot find the Neyman-Pearson Lemma in Peirce, for example, the relationship between power, significance, sample size, and error rates, the invertability of test statistics, or their asymptotic behavior. Peirce did, however, explicitly setup the issue of what research to pursue as a linear optimization problem, and derived its first order conditions well before contemporary econometric techniques were ordinarily used. See, [Wible(1994), Wible(2008)].

in is the probability of a parameter estimate being inaccurate, from the decision maker's perspective, rather than the probability of falsely rejecting a null hypothesis that is true or erroneously failing to reject a false hypothesis. The latter are evaluated by computing the probability of the data under the assumption that it is true/false. But on my approach, error probabilities are evaluated in terms of self-expectation.

On the Peircian approach, and in its early/mid-twentieth century refinements, our attitudes are not themselves probabilities. Our attitudes are: reject or do not reject the null hypothesis. However, many philosophers have recently moved away from the binary attitudes of acceptance/rejection and belief/disbelief and toward modeling doxastic/belief-like states using [Ramsey(1926)], [De Finetti(1937)] and [Savage(1954)]'s theory of subjective probability. This is the approach I follow. It would be anachronistic to locate *this* in Peirce. Peirce was one of the earliest philosophers to understand sampling theory and by some accounts the progenitor of key results in the Neyman-Pearson paradigm. His conception of probability was Frequentist, he often took a narrow view of empirical learning as given by sampling, and he was a harsh critic of then prevailing Bayesian methods. He was, however, one of the first scholars to work on elicitation of subjective probabilities in experimental psychology [Stigler(1978)] and in that light one might suspect he may not have been such a harsh critic of Bayesian methods as they have been developed in the second half of the twentieth century.

In any case, the insight that we can understand inquiry in broadly economic terms is a substantial one. As [Rescher(1976)] emphasizes, inference is, in Peirce's view, crucially dependent on economic considerations and reasonable assessment of the risk of different types of error as well as the value of correct verdicts. This central insight, as I interpret it, is my jumping off point – namely, *that inquiry depends in part on considering the costs and benefits of small changes in the probability of being mistaken*. Indeed, there are two central pillars to this Peircian insight as I understand it. The first is the notion of abduction as model selection. While the related notion of inference to the best explanation plays a significant role in realist approaches to philosophy of science, this way of thinking about abduction will be novel. The second is the notion of induction as risk management. The first of these insights corresponds to the ideas that will be developed in Chapter 2. The second corresponds to the ideas that will be developed in Chapter 3.

2.2.1 Abduction as model selection

At a very general level, every approach to inference in the subjective or Bayesian tradition must answer two very basic questions: (1) how shall I identify a prior distribu-

tion of beliefs? (2) how shall I update that distribution after receiving new information? [Gelman et. al.(2013)], for example, describe a trichotomy that requires identifying a joint probability distribution (model), drawing inferences from the model, and model assessment. While the third step is important, we will set it aside for now. Similarly, philosophical Bayesians focus on justifying particular ways of selecting priors and identify/defend update strategies for processing new information.

In chapter 2, I will develop a theory of epistemic risk for the selection/assessment of a prior distribution. For this, I draw on Peirce's notion of abduction. In Chapter 3, I will develop a dynamic characterization of epistemic risk for the assessment of an agent's update rule. This stage will draw on Peirce's notion of drawing inferences by analogy to insurance risk management.

Peirce coined the term 'abduction'. In contemporary philosophy of science, this term is often used synonymously with inference to the best explanation (IBE). [Harman(1965)], for example, explicitly identifies IBE with abduction. Indeed, many commentaries on Peirce refer to 'abduction' as a type of inference – suggesting that it is an alternative to induction. Put this way, it appears that one has the option, in making an inference, of using IBE, or abduction. But for Peirce this was precisely what abduction was not. Abduction and induction are two independent stages of inference. IBE for Peirce would have belonged to the inductive stage. Abduction comes earlier.

Peirce drew a distinction between abduction and induction in order to distinguish abductive processes from inferential ones in the context of a unified method of scientific inquiry. For him, abduction is the process of forming an explanatory hypothesis. It is the only logical operation which introduces any new idea (5.171-2) and it includes "all the operations by which theories and conceptions are engendered " (5.590).[2] Induction, on the other hand, is the subsequent process that is used to assess the hypotheses generated by the abductive process.

So what is this process by which theories are engendered before they are assessed – and in what way is it scientific? Indeed, we hear Peirce say that "abduction is an appeal to instinct" (1.630) which critics take to question whether abduction could play a meaningful part in a rational theory of inference. Commentators, such as [Frankfurt(1958)], for example, focus on the fact that Peirce highlighted that abduction may be schematized: it has a logical structure, therefore, it is scientific. For example: we observe some data $\mathbf{X}$. If theory H were true, $\mathbf{X}$ would be unsurprising. Therefore, we have some reason to suspect H might be true. The schematic reads very much like IBE, which has led to abduction's frequent confusion with IBE. After attributing this characterization of abduction to Peirce,

[2] All parenthetical references are to [Peirce(1931-1958)].

Frankfurt goes on to criticize it as insufficiently rational – i.e., there are too many degrees of freedom in which ideas one introduces by this process and no limit on how many ideas might be introduced. But this is not the way Peirce would have defended the rationality of abduction. The emphasis for Peirce is not on the likelihood of H given $\mathbf{X}$ (which is what subsequent commentators highlight), rather it is on the process of generating H to begin with. If we had two competing theories, but one made the data more plausible, then IBE may indeed tell us to choose the first theory. But abduction is not about adjudicating between competing theories in light of data. It is the selection method by which all candidate theories are identified for consideration to begin with. By definition, we have to consider the theory before learning that its likelihood is high under the observed data. But how is this to be done?

For Peirce, the process of abduction was closely related to his pragmatism. He says, for example,

> If you carefully consider the question of pragmatism you will see that it is nothing else than the question of the logic of abduction. That is, pragmatism proposes a certain maxim which, if sound, must render needless any further rule *as to the admissibility of hypotheses to rank as hypotheses*, that is to say, *as explanations of phenomena held as hopeful suggestions*; and furthermore this is all that the maxim of pragmatism pretends to do (5.196).

This is a bit of an overstatement in terms of the identification of pragmatism and abduction, but nonetheless, it shows just how closely related these two ideas were for Peirce. But how shall we put this into practice – what is the pragmatist process by which we establish "the admissibility of hypotheses to rank as hypotheses"?

Peirce's crucial insight, and one that is still overlooked even by commentators who recognize the distinction between abduction and IBE, is that this process is to be governed by economic considerations:"in all cases the leading consideration in Abduction ... is the question of Economy – Economy of money, time, thought, and energy" (5.600). For example, he points out that we may want to include in our problem implausible hypotheses because their truth-value may be easily settled. Critics point out that as a selection method this is over-inclusive and inefficient. After all, we would not want to waste our time disproving implausible hypotheses. But this (including easily disproved hypotheses) is not a selection method. It is an application of a more general economic process of abduction. The time and energy cost of including too many trivial hypotheses is precisely why Peirce's method would not recommend doing so. It is clear in Peirce that admissibility is governed by considerations of expected utility. If the hypothesis is so implausible as to be not worth

investigating, the low-cost of determining its truth-value will be insufficient to admit it into our model.

Therefore, which hypotheses we consider to begin with is not an unadulterated epistemic question – it is a question for expected utility analysis. For example, if I want to learn about the bias of an ordinary looking coin I could start from a stance of perfect neutrality, adopt a uniform distribution with respect to its limiting mean or objective bias, and waste my time tossing the coin many times. If nothing serious turns on this, it is not a very productive use of time. It is more efficient to start by assuming the coin is fair, and toss it a few times to confirm what I strongly suspect. Meanwhile, if I had to bet my life on the coin, or a lot of money, or a research grant, I might want to be more careful and start with a very cautious prior. As a result, the Peircian considerations of time, money, thought, and energy affect my assessment of the seriousness of the mistake that might be made. They reflect my assessment of *epistemic risk*. And in turn, my assessment of epistemic risk affects the prior I identify as appropriate. To paraphrase Rescher again, the process of induction is, for Peirce, crucially dependent on prior intelligent deployment of economic considerations at the abductive stage. Such normative considerations are therefore inseparable from scientific inference. Attitudes to risk of error – epistemic risk – therefore affect an agent's selection of a suitable prior. But that is not their only role.

Before I move on, it is important to highlight here that Peirce was a critic of what was then called inverse inference. Inverse inference roughly corresponds to Bayesian inference today, but it is a particular type of Bayesian inference – it is Laplacian "objective Bayesianism" as it is known today. Indeed, what Peirce took special issue with was the so-called principle of indifference, because it substitutes absence of information for information of equipossibility. Subjectivists relied on the principle since the central issue was how to assign priors in the absence of information. Peirce found the principle to be arbitrary. But had he applied the economy of research to the selection of priors he may have arrived at a different conclusion. Indeed, this is what I hope to do. So while it would be anachronistic to suggest that Peirce had a proto-theory for assigning prior probabilities I think his central insight of characterizing abduction in terms of the economy of research can help us think about the role of epistemic risk in identifying appropriate priors.

2.2.2 Induction as insurance brokerage

Induction for Peirce is the process of evaluating, on the basis of observed samples, hypotheses identified by the abductive process, including the economic considerations described above. Indeed, he equates inductive reasoning with sampling. The crucial aspect of

induction for Peirce is long-run error control. Rather than assigning probabilities to competing hypotheses, we are interested in using a decision procedure which minimizes the probability of mistaken verdicts in the long run. This is of course very familiar now from [Neyman and Pearson(1933)]'s null hypothesis significance testing framework, but the extent to which Peirce anticipated many of the ideas in NHTS is remarkable. As mentioned earlier, probability of error is what Peirce has in mind when he refers to reducing the "probable error of our knowledge" – it is the probability of being mistaken on the assumption that a hypothesis is true (a Frequentist notion) rather than the probability that the hypothesis is false (a Bayesian notion). Indeed, Peirce uses likelihood to describe how we are comparing hypotheses here. What I am interested in highlighting, in particular, is Peirce's focus on long-run error control. For example, Peirce says of induction,

> we cannot say that the generality of inductions are true, but only that in the long run they approximate to the truth ... In fact, insurance companies proceed upon induction – they do not know what will happen to this or that policy-holder; they only know that they are secure in the long run.

The insurance analogy for induction is what I want to highlight. On the Bayesian framework I will develop, there are two central steps that will be addressed in this dissertation: identifying a prior and drawing inferences from the prior. I explained in the previous section how considerations of risk of error are relevant to the identification of a prior. This will be the subject of Chapter 2. What I want to highlight here is that considerations of error are just as relevant in drawing inferences after one has identified a prior. This will be the subject of Chapter 3. Peirce takes the insurance analogy seriously. He says, for example,

> by faithfully adhering to ... [induction], we shall, on the whole, approximate to the truth. Each of us is an insurance company, in short. But, now suppose that an insurance company, among its risks, should take one exceeding in amount the sum of all the others. Plainly, it would have no security whatever.

There are many gaps to be filled here: are we all like insurance companies in the sense that we diversify our portfolio of beliefs by making many conjectures in the long run? This temporal dimension does not seem necessary. We can also diversify our portfolio by adopting many beliefs at any given time. Peirce does not say. Just as importantly, what does it mean to take a smaller or larger risk in the context of induction? How do we measure the riskiness of inductive inference? What exactly makes one inductive process riskier than another? Peirce does not say either. And finally, what would be the equivalent to leverage in the context of inference? Once we pile up a stock of truths, is it permissible to be more

reckless in the inferences we make? Or is it just as bad, perhaps even worse, to make a mistake after one has established a strong reputation as a careful scientist? Peirce assumes the former but I am doubtful this is the case.

These are all questions that will be addressed in Chapter 3, where the subject of epistemic risk will be the agent's update strategy. I will explain what makes one update rule riskier than another and provide a measure of epistemic risk for competing rules for updating (or, equivalently, for posterior probabilities). I will also evaluate the asymptotic behavior of inductive strategies with different degrees of riskiness. The idea, again, is not so much to suggest that Peirce had worked out a contemporary problem of statistical inference but rather that he identified a central insight – namely, to think about the process of evaluating hypotheses in light of new evidence in the way that an insurance company thinks about its policy-holders. The economic approach is just as valuable here as it is in the abductive context and, again, it underscores the inevitability of the normativity of statistical inference.

One concern here might be that my approach to inquiry seems obviously dependent on many practical considerations. The agent's choice of what to believe depends on both practical and epistemic consequences: her resources, time, money, energy, and so forth. My response is: it absolutely does. Indeed, I will further highlight the role of practical considerations at almost every step of scientific inquiry. As we will see below, in order to measure accuracy we need to identify a suitable loss function. I am skeptical one can do this on purely epistemic grounds. [Leitgeb and Pettigrew(2010a)], for example, defend the Brier score because of its symmetry. That is certainly an epistemic reason in favor of the Brier score, though I am not sure how persuasive it is, but in using it we are forced to give up the assumption that information loss is additive. But Shannon found additivity to be essentially an a priori property of information, and it figures as an axiom in his characterization of entropy. Ultimately, I will argue, to identify an appropriate scoring rule we need to consider the relevant risks of error. In a classification task where false positives and false negatives are equal (distinguishing cats from dogs in image search, for example), a symmetric score function seems appropriate (as opposed to, say, a biomedical context where the costs of different types of errors can be much different). Likewise, determining how much epistemic risk is appropriate will depend on the circumstances of the case and the agent's degree of risk aversion. One can certainly construct an objective rule of inference by adopting the maxim that epistemic risk ought to be minimized, and stipulating a particular loss function to be used for evaluating accuracy and measuring divergence between distributions, but this is not rationally required.

2.2.3 Micro or macro economy of research?

There are two ways to develop the Peircian project. One would be to start with a scientist's ordinary utility function and consider how the pursuit of truth interacts with other institutional factors in the modern scientific community to affect her decision of which line of research to pursue including, importantly, the values and behavior of other scientists. This is the project that [Kitcher(1990)] and [Zollman(2017)], for example, develop. The focus of this project is the institution of science within which individual agents interact and the incentive structure that it generates. The focus is on strategic group interaction. We may perhaps helpfully call this the macro-economy of research.

Another would be to take the agent and her pursuit of truth (the way she values truth epistemically) as a starting point, and explore how other considerations affect it, what normative attitudes guide her inquiry, and so forth. The key to this project it that we start from the individual scientist seeking truth or accuracy, and build up a more robust picture by evaluating how other considerations affect *her* pursuits. The emphasis is on the rationality of the individual from a decision-theoretic perspective. This is an individualistic, rational reconstruction of what is going on in science whose focus is on eliciting the attitudes that govern inquiry and providing a rationale for any particular individual. This is the micro-economy of research.

The two approaches are related of course – there is no bright line distinction. The general difference is in our aims and scope. The macro-economic philosopher of science seeks to explain how a rational scientist will behave in her community by modeling the whole group in terms of strategic interaction, highlighting game-theoretic equilibria, or using agent-based networks, for example. The micro-economic philosopher of science seeks to explain whether, for example, it is rational for an individual to hold a particular set of priors, how much accuracy she might sacrifice for additional information, how much risk is rationally permissible, and so forth. Her tools are similar to those of the ordinary microeconomic theorist, focusing on convex optimization in a decision-theoretic context. This dissertation may be thought of as a project in the micro-economy of research.

2.3 Epistemic utility theory

In this section I briefly describe epistemic utility theory, its relationship to ordinary rational choice, and how it compares to traditional analyses of knowledge in epistemology.

The subject of ordinary epistemology is the binary doxastic mental state *belief*. With respect to a proposition, an agent believes it, disbelieves it or, perhaps, is agnostic about

it. In this context, the epistemic value at stake is ordinarily taken to be *knowledge* and, by extension, truth and justification are central to the analysis of belief. Believing a false proposition or holding an unjustified belief is epistemically bad because the agent fails to know that proposition. It is natural to suppose, however, that our epistemic attitudes are often not binary. For example: I am relatively confident that a six-sided die will land on an even or prime number; I suspect it will not rain tomorrow; I am quite confident I will not win the raffle.

To accommodate such finer-grained attitudes, many philosophers take a Bayesian decision-theoretic approach inspired by [Ramsey(1926)], [De Finetti(1937)] and [Savage(1954)] and replace the binary framework with a probabilistic one where credences or subjective probabilities take the role of beliefs as the relevant doxastic attitude under analysis. A person's subjective probability in a proposition corresponds to her subjective degree of confidence. To paraphrase [Joyce(2009)], credences are *inherently gradational*: the range of their strength is continuous between complete certainty of the truth of a proposition to complete certainty of its falsehood – depending on the evidence available to the agent.

On this approach, *knowledge* is typically not the guiding virtue. My belief that 'the die will land on an even number' is either true or false. But my .5 credence that the die will land on an even number is neither true nor false. It can, however, be more or less accurate. Believing a true proposition to degree .9 or predicting an event which occurs with .9 confidence, is better than doing either with, say, .3 confidence. On this approach, being more accurate is better. So as a first pass we can say that in the context of fine-grained or partial belief epistemology, it seems reasonable to hold accuracy as the guiding normative virtue.[3]

However, we want to assess how well an agent is doing without knowing which propositions are true or false in advance. We want to say, for example, that from the agent's current evidential state, it is rational to hold the credences she holds. Similarly, we want to say that she made a good or reasonable prediction given the evidence she had at the time. The normative role that accuracy plays, therefore, is similar to the normative role of ordinary utility in rational choice. We do not fault an agent for failing to maximize utility – for example, for failing to make a reckless bet that against all odds would have paid handsomely. We do, however, fault the agent for failing to maximize *expected* utility – for failing to take the action that in expectation would have been best for her.

[3]I prefer to think of this as a modeling assumption rather than a fundamental truth of epistemology. In economic theory, it is customary to assume that one's utility is a function of money, but there is no reason why tastes, values, altruistic desires, and so forth cannot affect an agent's utility. The same goes for the epistemic case. We will assume for now that epistemic utility is given in terms of accuracy, but it would be interesting to consider how parsimony, explanatory power, verisimilitude, and other epistemic virtues may affect it.

As a result, maximizing *expected* accuracy is a more accurate description of the governing normative virtue for the epistemology of partial beliefs. The approach is of course most useful for assessing an agent's doxastic state from the internal perspective – i.e., given her evidential state – or from evaluating the quality of a prediction before we learn the actual outcome. Indeed, it is a generalization of ordinary decision theory with accuracy assuming the role of ordinary utility. We will adopt the useful fiction that an agent can choose her credences and in order to evaluate an agent's credences we will consider whether they maximize expected accuracy (or minimize expected inaccuracy).

For this reason, I follow the literature and refer to this approach to epistemology – i.e., the graded approach originating in [Ramsey(1926)], [De Finetti(1937)], and [Savage(1954)] – as epistemic utility theory.[4] It is a decision theory for evaluating an agent's choice of what beliefs to hold, rather than which actions to take.

2.4 Risk and rational choice

From its earliest development, expected utility theory has been bound up with risk. In 1738, Daniel Bernoulli presented what has come to be known as the St. Petersburg Paradox [Bernoulli(1954/1738)]. Bernoulli identifies a gamble with infinite expected value which would be judged by nearly everyone as not reasonably worth more than a modest sum. The puzzle identifies a mismatch between the expected value of a gamble and the amount that a rational person would pay for it.

Bernoulli diagnoses the problem by pointing out that the marginal utility of money diminishes as wealth increases and that our intuitions about the irrationality of paying excessively for the St. Petersburg gamble reflect an aversion to risk. This is the origin for understanding attitudes to risk in terms of the curvature of a person's utility function. A risk averse person would not pay 1 dollar for a gamble that pays 0 or 2 dollars with equal probability (a fair gamble) because from her perspective the utility of the expected value is greater than the expected utility. The diminishment in wealth, if things go poorly, is weighted more than the improvement, if things go well. This is of course an application of the defining property of concave functions. If X is a random variable and f is concave over its support then $f(\mathrm{E}[X]) \geq \mathrm{E}[f(X)]$. In the terminology of modern economic theory, an individual with a concave utility function is risk-averse; an individual with a convex utility function is risk-seeking; and an individual with a linear utility function is risk-neutral.

It is natural to suppose, then, that the degree of concavity of an agent's utility function reflects the extent to which she is risk-averse. This is precisely the insight that [Arrow(1965)]

[4] [Joyce(1998), Joyce(2009), Greaves and Wallace(2006), Pettigrew(2016a)].

and [Pratt(1964)] exploit. We might start by taking the second derivative of the utility function as a measure of an agent's aversion to risk. However, since von-Neumann Morgenstern utility functions are unique only up to affine transformations[5] we should divide the resulting quantity by the first derivative so as to get a measure that is invariant to arbitrary changes in slope and location. To get a number that increases in magnitude as concavity increases we add a negative sign in front of the ratio of derivatives. This is the Arrow/Pratt coefficient of risk aversion.

It became clear relatively early that this notion of risk aversion was extremely important to economic analysis. Understanding attitudes to risk has important implications for understanding the rationality of individual decision-making, market behavior, insurance, and evaluating social policy. But it was never clear in all this what risk itself was. In other words, Arrow and Pratt provide the starting point for an analysis of a decision maker's *attitudes* to risk. The ratio of derivatives model provides a framework for explaining when one person is more risk averse than another or how aversion to risk changes with wealth, for example. But they leave open a more fundamental question: *what is risk*? When we say that an agent is risk-averse what about the gamble is it she she is averse to? Can we line up a set of gambles on a shelf, so to speak, and rank them according to their risk? This is the question that [Rothschild and Stiglitz(1970)] take up.

Their answer, which improves upon the earlier mean-variance approach of [Markovitz(1952)] and [Markovitz(1959)], is that one gamble is riskier than another if it is a mean preserving spread of it. What risk averters are worried about, on their approach, is the increase in the uncertainty of monetary outcomes. A gamble which pays 1 or 2 dollars with equal probability is less risky than a gamble which pays 0, 1, 2, or 3 dollars with equal probability. While they have the same expected value, the second is created from the first by taking the probability mass from each outcome and spreading it over better and worse outcomes in a way that keeps the mean fixed. Risk on this approach provides a partial stochastic ordering of gambles. Importantly, [Rothschild and Stiglitz(1970)] show that if one gamble is a mean preserving spread of another then a risk averse decision maker would prefer the first to the second. This connects up the Arrow/Pratt coefficient of a decision maker's degree of risk aversion with the risk ordering of lotteries themselves.

These notions, both of risk and of attitudes to risk, have been central to the development of the standard model of rational choice. However, considerations of risk are almost nonexistent in the context of epistemic utility theory.[6] My goal in the chapters to follow

[5] [von Neumann and Morgenstern(1944)].

[6] [Fallis(2007)] is really the only paper that explicitly addresses epistemic risk within the epistemic utility paradigm, though the notion comes up in epistemology discussions more generally, e.g. [Maher(1993)] and [Levi(1962), Levi(1974), Levi(1977)] and it also comes up implicitly in some of the literature in epistemic

is to begin to develop a line of research regarding epistemic risk by, hopefully, framing a number of fruitful questions, and developing an approach to epistemic risk that begins to answer at least some of them.

2.5 Value at risk

In financial analyses the expression 'value at risk' denotes the quantity (in monetary terms) that a firm or financial portfolio, say, stands to lose. I use the phrase more literally here to ask what is the normative *value* that may be under risk in epistemic contexts? When we say an agent is epistemically conservative, we presuppose she is being extra careful to avoid some sort of loss. What that loss is will determine our analysis of epistemic risk. There are two competing approaches to the epistemic value at risk: the alethic approach, which I have hinted at above, which follows [Joyce(1998)], [Harman(1986)] and [Goldman(2002)], among others, and the competing modal approach, developed by [Pritchard(2017)].

2.5.1 The alethic approach

I assume, following many others, that an epistemic agent should aim to believe truths and avoid believing falsehoods. Truth seeking is the overarching goal in epistemology. As a result, when we formulate a belief, we do so for the sake of an epistemic benefit – namely, to hold a belief that is true – and despite a potential epistemic cost – namely, to believe something that is false. For example, [Harman(1986)] defends epistemic conservatism – the notion, roughly, that a reason for holding on to a set of beliefs is the fact that we currently hold them. The value at risk here is truth – by changing our beliefs we risk believing an additional falsehood or disbelieving an additional truth. The benefit we seek when we form beliefs, therefore, is to believe the truth.

Meanwhile, in epistemic utility theory, when an agent identifies her credences, something is likewise at stake. Whatever that is – the relevant normative consideration when we form beliefs or credences – will determine how we should proceed. Since the relevant attitude on our approach will be a credence or subjective probability the relevant value is graded accuracy.

Graded accuracy, as we will see, may be evaluated in terms of common families of statistical loss functions. A common approach, for example, is the Brier score, which measures the squared distance between the true value and the assigned probability. An agent's

utility theory, e.g. [Pettigrew(2016b)]. However, neither Pettigrew nor Fallis provide a measure of risk or a stochastic ordering of credence functions of the sort that we find in economic theory.

credence that a die will land on an even number is neither true nor false, but it can be more or less accurate. Therefore, the probabilistic generalization of the veritistic perspective is [Joyce(1998)]'s norm of gradational accuracy – an epistemic agent should aim to assign high probabilities to truths and low probabilities to falsehoods. On the probabilistic approach, the risk we take when we adopt one credence function instead of another is the risk of moving further from the truth. The value at risk, therefore, is truth in the categorical case and graded accuracy in the probabilistic case. Both approaches are alethic.

2.5.2 The modal approach

Alternatively, we might prefer to think that what is at risk is not the risk of error – i.e., the potential of holding a false belief or inaccurate credence – but rather the risk of failing to have knowledge – i.e., the risk of holding a belief that, while true, fails to constitute knowledge. This approach emerges out of anti-luck approaches to epistemology, where safety is central to justification. 'Safety' is a technical notion in modal approaches to epistemology. For [Pritchard(2007)], for example, an agent's belief is safe if it remains true in most nearby possible worlds in which the agent holds the belief in the same way as in the actual world.

In Hohfeldian terms, we might say, risk is a natural correlative to safety. Unsafe beliefs are risky and conservative beliefs are safe. However, because overwhelmingly probable beliefs can be unsafe, the value at risk cannot be the alethic value described above. Some beliefs on modal accounts may well be overwhelmingly probable without being safe, as we will see below. For someone who takes the value at risk to be truth or accuracy, such beliefs cannot possibly be risky. So we must look elsewhere for the value at risk. This approach to risk goes with the position that belief aims, not so much at truth, but at knowledge in particular. The value at risk is not truth or gradational accuracy. It is knowledge. To say that a belief is risky, therefore, is not to say that one risks believing something false, or perhaps something inaccurate (in the case of graded beliefs). Rather, it means that one risks having a true belief that fails to constitute knowledge. On knowledge first approaches to epistemology, this is the main source of risk that believers face [Williamson(2000)].

The main problem with Pritchard's approach from my perspective is that it fails to guide epistemic behavior. It is, at best, a diagnosis of common failures of rationality. As a diagnosis, however, it is not clear that it adds anything to our understanding of cognitive biases beyond what the experimental literature in psychology already provides.

2.5.3 Risk and normativity

[Pritchard(2017)] articulates a modal theory of epistemic risk on which attitudes to risk come apart from probabilistic judgments and can be pretty clearly irrational. As he puts it: "According to the modal account of risk, the level of risk involved is determined by how modally close the target risk event is. In particular, where the risk event is modally close, then it is high-risk (even if it is a probabilistically unlikely event), but where it is not modally close, then it is low-risk" (pg. 13).

For example, people often judge airplane travel to be riskier than driving a car even though statistically one is more likely to be hurt in a car than on an airplane. This attitude may be explained on Pritchard's modal account as follows: A car driver may be subject to cognitive biases that make her conception of the actual world such that the possible world in which she incurs serious injury while driving is not especially close, and hence not a serious risk. For example, Pritchard says, she may be under the illusion of control leading her to judge herself as much more competent than the average driver. Meanwhile, the passenger on the airplane may be subject to a different set of cognitive biases that make her conception of the actual world such that the possible world in which the plane crashes is quite close. For example, the accessibility bias, coupled with the prevalence of reports regarding airplane accidents.

Pritchard, therefore, assumes the task of vindicating such common judgments about risk, as documented in the social psychology literature. I do not. He says, for instance: "subjects might grant that the probabilistic likelihood of two events is broadly the same, and yet nonetheless characterize one of them as being riskier than the other because they regard this event as modally closer." (pg. 11) Therefore, "While subjects will grant that the probability of sustaining serious injury when, say, driving a car is much, much higher than alternative forms of transport, such as taking the train, they nonetheless tend to judge that car driving is not an especially risky activity." (pg. 11)

I think of my project on epistemic risk as having a normative dimension. That is not to say it is normative in the sense that I defend a unique attitude or set of attitudes to risk (e.g., you should minimize it) but in the sense that if you think taking a plane is riskier than driving a car, despite statistical evidence to the contrary, then you are mistaken about the risks involved. Indeed, I would not want a theory on which it comes out to be true that taking a plane is riskier than driving a car. If your insurer would not accept it as true, neither should you, even if a plane crash is more salient to you so that you judge it to be modally closer.

This does not mean that there is no degree of subjectivity on my account. As we will see below, which set of beliefs minimizes epistemic risk will depend on how an agent

measures inaccuracy. So it is possible on my account for two agents in the same evidential circumstances to judge the same credence function differently in terms of risk. However, they will at least be able to come to an agreement that the source of their disagreement is normative – it is a disagreement about value. In particular, it is a disagreement about the relative cost of moving in the direction of different types of error. On Pritchard's account, however, the disagreement does not seem to be a disagreement about value. We disagree because we are irrational in different ways. I believe driving a car is safer than it really is because of the illusion of control and you believe flying in an airplane is more dangerous than it really is because of the accessibility bias. In other words, the source of disagreement on the alethic approach is a difference in value (just how bad is it to make a false positive/ false negative error?), whereas the source of disagreement on the modal approach is a difference in irrationality (which set of cognitive biases is one's judgment predominantly skewed by?).

CHAPTER 3

Generalized Entropy and Epistemic Risk

3.1 Introduction

My goal in this chapter is to provide an account of *epistemic* risk that is analogous in important respects to contemporary approaches to risk in expected utility theory. I develop this approach within the framework of what has recently come to be called epistemic utility theory, following [Joyce(1998)], [De Finetti(1974)] and, ultimately, [Ramsey(1926)]. In particular, I assume that an agent's selection of a subjective probability distribution (or credence function) may be treated as an epistemic act and that the rationality of that act may be evaluated using the tools of ordinary decision theory. To measure epistemic utility I use a familiar class of statistical loss functions known as scoring rules.[1] Whereas an ordinary decision maker seeks to maximize expected utility, the epistemic agent seeks to minimize expected inaccuracy. Accuracy, then, is our primary commodity, and epistemic norms – such as probabilistic coherence, updating by Bayesian conditioning, and chance calibration principles, for example – may be defended on the ground that they promote accuracy.[2] This basic idea originates with [Ramsey(1926)], [Savage(1971)], and [De Finetti(1974)], and it has been developed in important respects by [Lindley(1982)] and [Schervish(1989)].

The risk measure I propose is inspired by [Rothschild and Stiglitz(1970)]'s approach to economic risk. I motivate the idea that one subjective probability distribution is riskier than another if it is a mean preserving spread of it and that the least risky probability assignment is the one that guarantees a particular inaccuracy score regardless of the outcome. We will see that mean preserving spreads may be measured in terms of changes in expectation, and that a plausible measure of risk, therefore, is the difference in expectation from the risk-free probability. Following [Grunwald(2000), Grunwald and Dawid(2004)], I use the term 'general entropy' to refer to the expected inaccuracy of a probability distribution evaluated

[1] See [Gneiting and Raftery(2007)] for an overview.

[2] See [Joyce(1998)], [Greaves and Wallace(2006)], and [Pettigrew(2012)], for an accuracy-based defense of each, respectively.

with respect to itself (we will see why, below) and as a result epistemic risk turns out to be a measure of entropic change.

While the risk function is similar in spirit to the economic notion of risk, it has a uniquely epistemic interpretation, which has its roots in [Peirce(1879)]'s "economy of research". In particular, the shape of the agent's risk function reflects her attitude toward the relative cost of increasing inaccuracy in the direction of false positive (Type I) mistakes against the cost of increasing inaccuracy in the direction of false negative (Type II) mistakes. On larger sample spaces, the agent's risk function reflects her attitude to increasing inaccuracy in the direction of every possible outcome. Meanwhile, the curvature of the risk function encodes attitudes toward marginal changes in inaccuracy and local sensitivity to error.

From every risk function we may derive a unique scoring rule, and the agent's attitude to different types of error will determine the shape of her score. For example, if she considers the different error costs to be equal, her score will evaluate equally changes in inaccuracy in the direction of each outcome. If such an agent seeks to minimize epistemic risk, she will identify a uniform prior by applying the principle of indifference. Plausibly, then, one application of this approach is the selection of priors for Bayesian statistical inference. However, the uniform prior minimizes epistemic risk *only if* the different types of error are treated equally. More generally, the relationship between risk, error costs, and general entropy suggests that there exists a *family* of indifference principles each reflecting a different way of evaluating the error costs of a prospective probability distribution. This highlights the normative commitments that come with endorsing an uninformative or flat prior. These consequences follow from a central duality between epistemic risk and general entropy – namely, the sum of risk and entropy is constant provided the associated scoring rule is proper:

$$Risk + Entropy = k$$

This implies that risk is a scaled reflection of entropy. The agent's risk profile, therefore, is in an important sense epistemically central. Once we know what it is, we can determine the appropriate measure of risk, the associated entropy, the scoring rule, and the measure of divergence to be used for updating. For example, a logarithmic scoring rule will also imply the Kullback-Leibler measure of divergence for updating, and from this perspective Bayesian conditioning is optimal.

The chapter proceeds as follows. In Section Two, I describe the relevant formal concepts. In Section Three, I develop the theory of epistemic risk for the simple case where the agent is interested in a single proposition. In Section Four, I articulate the normative attitudes to the cost of error implied by the location, shape, and curvature of an agent's

epistemic risk function. In Section Five, I develop the duality between risk and entropy, and explain the conceptual difference between minimizing epistemic risk and maximizing information entropy. While the duality between these two concepts shows that they are often co-extensive, they are independently motivated. In Section Six, I extend the measure of epistemic risk proposed to general sample spaces, both continuous and discrete with any number of outcomes. In Section Seven, I explore in more detail the relationship between epistemic risk, the selection of priors, and the principle of indifference, especially as defended by [Jaynes(1957a), Jaynes(1957b), Jaynes(2003)].

3.2 Background

In financial analyses, the expression 'value at risk' denotes the quantity in monetary terms that a firm or, say, investment portfolio stands to lose. As I seek to develop a theory of epistemic risk, the value at risk should be epistemic. While other considerations are often at stake in inquiry – securing research funding, obtaining grants, achieving tenure, and so forth – the epistemic cost is independent of these other commodities. As a result, I develop an approach to epistemic risk within the framework of epistemic utility theory, where accuracy is the primary source of value.

Following the literature, I adopt the useful fiction that an agent is able to choose between competing credence functions. As a result, the credence function is the object of risk – it is credence functions that can be more or less risky. By analogy to economic approaches to risk, what makes one credence function riskier than another is that the agent stands to lose more in terms inaccuracy or that variability in accuracy outcomes is greater. Unlike belief or acceptance, accuracy is a graded notion. Therefore, the theory of epistemic risk I propose is a theory of the alethic sensitivity to (big or small) changes in inaccuracy. This resembles in some respects [Peirce(1879)]'s notion of the "economy of research". While it would be anachronistic to locate the germ of this framework in Peirce, he comes extraordinarily close to developing the idea by applying notions of ordinary utility and the cost of error to the epistemic context.[3] I develop the notion of epistemic risk carefully below. The remainder of this section provides a directed introduction to the relevant epistemic utility background.

[3]Peirce says, for instance: "The doctrine of economy, in general, treats of the relations between utility and cost. That branch of it which relates to research considers the relations between the utility and the cost of diminishing the probable error of our knowledge" (643). As [Rescher(1976)] emphasizes, inductive logic is, in Peirce's view, crucially dependent on economic considerations and reasonable assessment of the risk of different types of error as well as the value of correct verdicts. The difference between Peirce's approach and the approach to be developed is that Peirce did not have at his disposal the notion of *epistemic* utility in terms of graded accuracy. Subsequently, [Levi(1974)], [Maher(1990), Maher(1993)], and [Fallis(2007)] have suggested similar approaches to epistemic risk.

I assume that an epistemically rational agent should adopt as her credence function a probability distribution whose expected inaccuracy is at least as low as any alternative distribution she might adopt. This is the norm of **gradational accuracy**.[4] Therefore, minimizing expected inaccuracy plays a similar role in epistemic utility theory that maximizing expected utility plays in ordinary decision theory.[5] To measure inaccuracy, we use a **scoring rule**. This is a two-place function $s : \{0, 1\} \times [0, 1] \to \mathbb{R}$, denoted by $s_v(p(h))$, that measures the inaccuracy of the probability assigned to h when the true outcome is v, where $v = 1$ if h is true and 0 otherwise. I restrict my attention to coherent agents for whom the credence function is a probability (the account can be modified to apply to non-ideal cases as well).

Five properties of scoring rules will be relevant to my argument: additivity, truth directedness, continuity, propriety, and $0/1$ symmetry. For **additive** scores, the overall inaccuracy of a discrete credence function $\langle p(h_1), ..., p(h_n)\rangle$ is $\sum_{i=1}^{n} s_v(p(h_i))$ (for continuous credence functions we integrate $s_v(p)$ over the sample space where p is a density). Notice that it is possible to have a situation where the rule we use to evaluate the inaccuracy of one proposition may be different from the rule we use to evaluate the inaccuracy of another one. It may be that for every element in the partition a different score is applied. The overall inaccuracy is still given by the sum of individual inaccuracies.

A minimal constraint on the functional form of scoring rules is that they be truth-directed. **Truth-directedness** implies that $s_1(p)$ is a decreasing function of p and $s_0(p)$ is an increasing function of p. Thus, moving closer toward the actual truth-value cannot make an agent worse off. It is also typically assumed that s_1 and s_0 are **continuous** functions of p, so as to avoid arbitrarily small changes in credence leading to big changes in inaccuracy. Truth-directedness and continuity are generally accepted properties of an appropriate measure of inaccuracy.

The expected inaccuracy of a probability distribution is the expectation of $s_v(p)$ evaluated with respect to the agent's beliefs, $b = b(h)$. In the binary case this is,

$$\mathrm{E}_b[s_v(p)] = bs_1(p) + (1 - b)s_0(1 - p) \tag{3.1}$$

If this equation is (uniquely) minimized at $b = p$ the score is **(strictly) proper**. This means that a coherent agent can do no better in expectation, from the perspective of minimizing inaccuracy, than to adopt as her credence function the probability distribution that

[4] [Joyce(1998), Joyce(2009), Pettigrew(2012)].

[5] A thorough (and opinionated) development of epistemic utility theory may be found in [Pettigrew(2016a)]. [Greaves(2013)] presents important objections to modeling epistemic rationality in decision theoretic terms.

corresponds to her sincere degrees of belief. Finally, s_v is **0/1 symmetric** if, given two probabilities for h, $p(h)$ and $q(h)$, that are identical except that $p(h) = 1 - q(h)$, then $s_1(p(h)) = s_0(q(h))$.

I assume that an agent's normative attitudes to risk, if they are to be found anywhere, must be reflected in the prior the agent deems appropriate. As a result, in developing a measure of epistemic risk we set aside for now considerations of updating and ask: regardless of one's evidence about a proposition, what structural features make one credence riskier than another? Of course, it is also important to consider what makes one update riskier than another. Equivalently, how much epistemic risk might be justified by the agent's evidence? These are questions about *dynamic* epistemic risk and I pursue them in a subsequent project.[6]

3.3 Epistemic risk: the simple case

Consider an agent formulating a credence $p(h)$ about a single proposition h. Regardless of h's content or epistemic import, we know that her inaccuracy decreases as her credences get closer to the truth and that it increases as they get further away from it. Since s_1 is continuous and decreasing on $[0, 1]$ with $s_1(1) = 0$, and s_0 is continuous and increasing on $[0, 1]$ with $s_0(0) = 0$, the intermediate value theorem guarantees that there exists a p^* for which $s_1(p^*) = s_0(p^*)$. For $0/1$ symmetric scores, this is $.5$. For asymmetric scores it may be something else. The following figure illustrates this situation. Figure (3.1) depicts a symmetric score whereas Figure (3.2) depicts an asymmetric one.

[6]Relatedly, [Buchak(2010)] uses ordinary attitudes to risk in epistemic contexts to argue against [Good(1967)]'s principle that one should not turn-down cost free evidence. Similarly, [Bradley and Steele(2016)] consider certain rare cases where an agent with imprecise credences seems to be committed to the rationality of paying to avoid cost-free evidence.

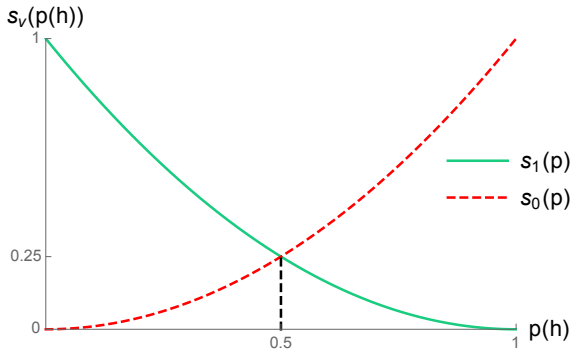

Figure 3.1: Risk-free probability (symmetric score)

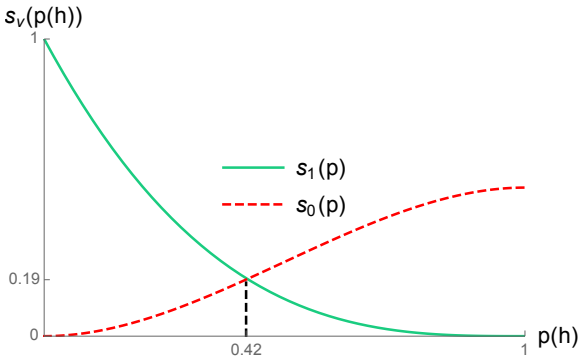

Figure 3.2: Risk-free probability (asymmetric score)

The point p^* may be thought of as the least risky point in the following sense: if the agent's credence for h is given by p^* her inaccuracy will be the same regardless of the actual truth-value for h. As a result, she knows with certainty how inaccurate she will be even before she learns whether h is true or false.

It is natural to think of a guarantee in one's outcome as implying an absence of risk. Indeed, this is the purpose of ordinary insurance: to charge a premium for guaranteeing a particular outcome (and, in turn, removing risk) – hence the term 'risk premium'. The outcome in insurance contexts is given in monetary terms (e.g., one does not have to pay out of pocket costs for a home repair). Here the same idea still applies, but our epistemic commodity is accuracy and therefore the outcome is given in inaccuracy as measured by a scoring rule. Informally, therefore, we might identify p^* as the least *risky* probability in the sense that it *guarantees* a certain inaccuracy score, regardless of the outcome. Since the choice of scale in constructing a risk measure is arbitrary, we may call p^* the risk-free credence, and define it more formally as follows.

> **Risk-free credence.** Given a single proposition h the risk-free credence $p(h) = p^*$ satisfies the equation $s_1(p^*) = s_0(p^*)$.

Now suppose that the agent has a credence for h that is more extreme than the risk-free one, say $p(h) = .8$. Given this credence if h is true her inaccuracy will be very low, but if h is false her inaccuracy will be quite high. Since $p(h) = .8$ creates an opportunity for the agent – the probability of doing better – together with a corresponding potential cost – the probability of doing worse – it is in this sense a riskier credence relative to $p^*(h)$. A natural measure for this increase in risk is the spread between s_1 and s_0, as depicted by the shaded areas in the following figure, because this quantity increases monotonically with shifts of probability to the tails of the risk-free distribution. Figure (3.3) depicts the increase in risk from a $0/1$ symmetric score's risk-free credence whereas Figure (3.4) depicts the increase in risk from an asymmetric score's risk-free credence. Notice, however, that it is

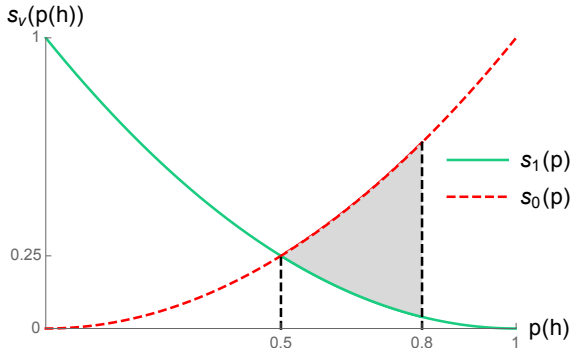

Figure 3.3: Symmetric measure of epistemic risk

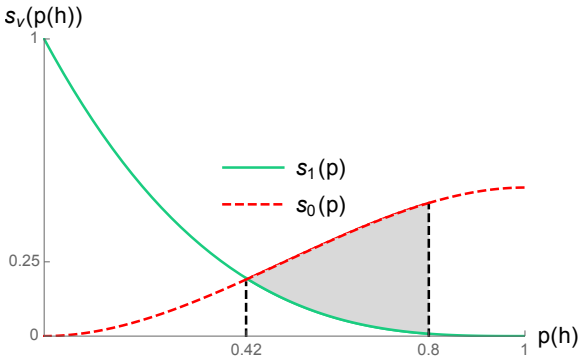

Figure 3.4: Asymmetric measure of epistemic risk

less sensible to speak about one credence function being riskier than another if we vary the number of possible outcomes in the sample space. For example, with three outcomes instead of two, the risk-free probability would occur where $s_1(p) = s_2(q) = s_3(1-p-q)$. Assuming a $0/1$ symmetric score this would be the uniform distribution $p = q = 1/3$. So to compare the increase in risk of another distribution over three outcomes we should measure the "spread" from the risk-free distribution for this larger space (we will see how to do this

later). In light of these remarks, we can define a risk measure for the single proposition case as follows.

> **Epistemic risk.** Given a single proposition h and a risk-free credence p^* the risk associated with investing credence $p < p^*$ in h is,
>
> $$R(p) = \int_p^{p^*} |s_1(t) - s_0(t)|dt$$
>
> For $p > p^*$ the bounds of integration are reversed. For $p = p^*$, $R(p) = 0$.

Provided the scoring rule is continuous, the risk function will be likewise continuous. Its local maxima $\max_p R(p) = s_v(p^*)$ will occur at $p(h) = 0$ and $p(h) = 1$. Since the scoring rule must be monotonically decreasing as the credence approaches the true value, it is easy to tell that risk monotonically increases away from the risk-free credence.

3.4 Risk and normativity

Any move away from the risk-free credence risks increasing inaccuracy by either increasing confidence in h when it is false, or decreasing confidence in h when it is true. Whether or not one deems the direction important reflects a substantial normative attitude toward the cost of approaching different types of error. As $p(h)$ goes up, one risks increasing inaccuracy in the direction of a false positive (Type I) error. Meanwhile, as $p(h)$ goes down, one risks increasing inaccuracy in the direction of a false negative (Type II) error. It is doubtful that the only correct attitude to these types of error is indifference. Being solely concerned with the truth, as [Gibbard(2008)] points out, does not commit one to a particular way of valuing accuracy. As a result, we want our measure of risk (and associated scoring rule) to reflect different trade-offs that agents might make between moving toward either type of error. This will enable us to evaluate the rationality of different attitudes to epistemic risk.

For example, h could be the outcome of a coin toss, where unit increases in inaccuracy in the direction of falsely predicting h (heads) are about as bad as unit increases in inaccuracy in the direction of falsely predicting its negation (tails). This set of attitudes to error is adequately captured by a $0/1$ symmetric score, such as the Brier score where $s_v(p) = (v - p)^2$, because an $\epsilon > 0$ increase in inaccuracy in the direction of either s_1 or s_0 from any point $p(h) = k \in [0, 1]$ leads to a decrease in epistemic utility of $(k - \epsilon)^2$. Figure (3.5) depicts this situation.

Since the score is symmetric a unit move in either direction away from the risk-free credence increases risk by the same amount. As a result, the risk of $p(h) = .8$ (the shaded area to the right of the risk-free point) is equal to the risk of $p(h) = .2$ (the shaded area to its left). Indeed, they are reflections of each other around the risk-free point. As a result, an agent with this risk function is equally sensitive to unit increases in inaccuracy in the direction of either false positive or false negative errors.

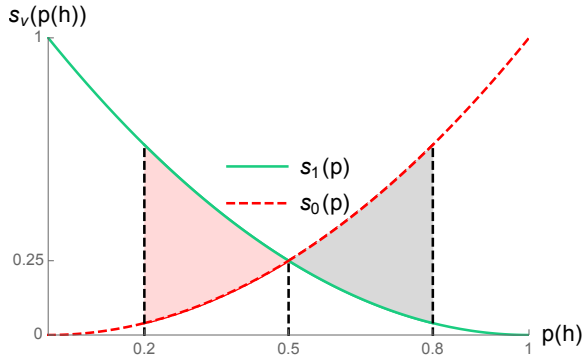

Figure 3.5: Epistemic risk and graded error (symmetric)

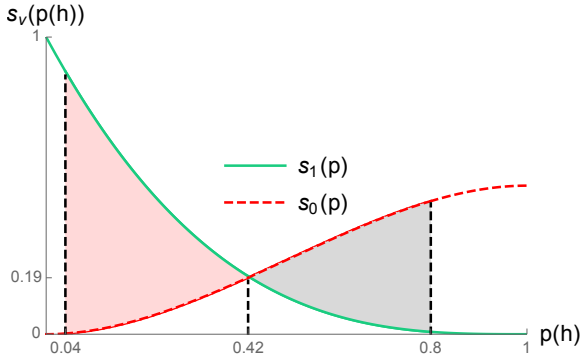

Figure 3.6: Epistemic risk and graded error (asymmetric)

Alternatively, h could be a very informative proposition that the agent is singularly pursuing so that the relevant partition is simply h and its negation. In this case, falsely believing h may be much better than falsely believing $\overline{h}$. The latter may produce an enormous epistemic opportunity cost that delays or more permanently inhibits her search for the truth, whereas the former may take the agent on a misleading line of inquiry that can be corrected through subsequent experimentation. In this example, unit increases in inaccuracy in the false negative error direction are worse than unit increases in inaccuracy in the false positive error direction. In this and similar contexts, a $0/1$ score is inappropriate.

Such attitudes to error are better captured by an asymmetric score whose risk function puts more weight on false negative increases in inaccuracy. An example of this is the score

considered in [Joyce(2009)], where $s_1(p) = (1-p)^3$ and $s_0(p) = (p^2/2)(3-2p)$. Like the Brier score, this score is proper, continuous, and monotonic. But unlike the Brier score an ϵ increase in inaccuracy in the direction of s_1 from $p(h) = k$ leads to a decrease in epistemic utility of $(k-\epsilon)^3$ whereas an increase in inaccuracy in the direction of s_0 leads to a decrease in epistemic utility of $\epsilon^2(3-2\epsilon)$. This situation is depicted in Figure (3.6). For this score, a unit move away from the risk-free credence in the direction of a false positive error leads to a smaller increase in risk (the shaded area to the right) than a correspondingly large move away from the risk-free credence in the direction of a false negative error (the shaded area to the left). As a result, the risk of $p(h) = .8$ is not equal to the risk of $p(h) = .04$ (nor for that matter is it equal to $p(h) = .2$).[7]

The symmetry of the embedded scoring rule is encoded in the risk function itself – since we capture risk by integrating the score's absolute difference. In particular, it is reflected by the risk function's location. As figures (3.7) and (3.8) show, a risk function associated with a $0/1$ symmetric score will reach its minimum at $p(h) = .5$ (left panel) whereas if the risk reaches its minimum elsewhere on the unit interval the embedded score must be asymmetric (right panel). In the example depicted, the risk function in the right panel is slightly shifted to the left.

[7]Note that there will be an equally risky point in the direction of a false negative mistake as $p(h) = .8$ – namely, the point $p(h) = v$ where $\int_v^{.42}(s_1 - s_0)dt = \int_{.42}^{.8}(s_0 - s_1)dt$. But since this particular score is relatively more sensitive to moving in the direction of a false negative error, v will be closer in probability to the risk-free credence than .8 is to the risk-free credence. Therefore, permuting probabilities for symmetric scores does not affect their risk. For asymmetric scores permuting probabilities does not preserve risk, though there exist "risk preserving permutations" which reflect the agent's relative sensitivity to different types of error.

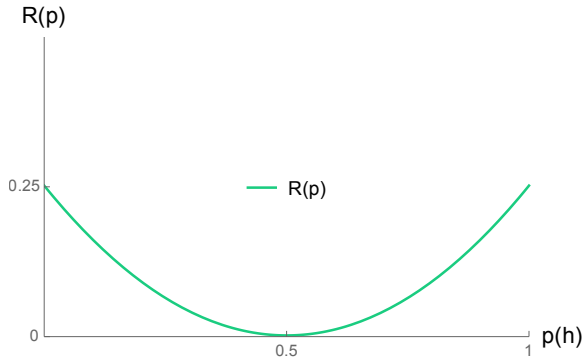

Figure 3.7: Symmetric epistemic risk function

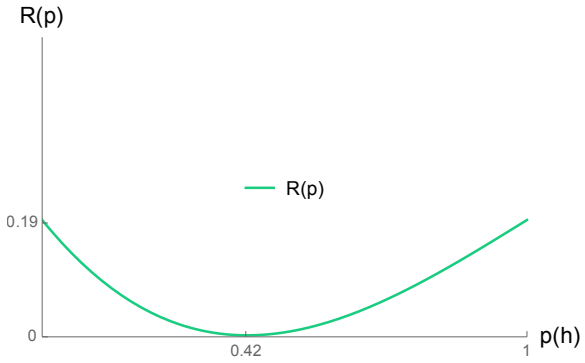

Figure 3.8: Asymmetric epistemic risk function

I refer to risk functions such as the one in Figure (3.7) as **symmetric**: it reaches its minimum at $p(h) = .5$ and its shape on $[0, .5)$ is a reflection of its shape on $(.5, 1]$. Symmetry in the risk function is related to $0/1$ symmetry of the scoring rule: A scoring rule is $0/1$ symmetric *only if* its associated risk function is symmetric.

Therefore, we should distinguish at least two different ways of valuing accuracy: a symmetric risk function corresponds to a way of valuing accuracy in which moving away from the truth in either direction is equally bad whereas an asymmetric risk function implies a way of valuing accuracy where unit changes in the direction of false positives/negatives get weighted differently at different credal values. Indeed they may not be weighted equally at any place. It is not enough, therefore, to claim as [James(1896)] does that we should seek truth and avoid error. Such an epistemic norm is underspecified. We need to decide further how to trade-off the potential costs of different types of mistakes. The epistemic risk function is flexible enough to encode different ways of balancing the competing costs.

So far we have exploited only the location of the risk function. But the risk function in Figure (3.8) is not just shifted to the left. Speaking picturesquely, it is also pressed against the y-axis. As a result, there is both a within and between difference in its *concavity*: it is (a)

steeper to the left of its risk-free point than it is to its right, and (b) it is not equally concave as compared to the risk function in Figure (3.7), whose embedded score is symmetric. These properties add further texture to the proposed measure of risk, reinforcing the idea that risk is a measure of alethic sensitivity to error. To exploit the concavity of the risk function, we need to introduce another quantity.

Let $h(p) = s_1(p) - s_0(p)$. For example, when $p = .8$, $h(p)$ is a measure of the length of the dashed vertical line segment connecting s_1 and s_0 at .8. $R(p)$ is the antiderivative of $h(p)$. As a result, our definition of epistemic risk implies that $R'(p)$ is equal in absolute value to $h(p)$. This means that the rate at which risk increases as we move away from the risk-free point reflects the increase, in absolute value, between the agent's best and worst outcomes. As a result, while the risk function itself reflects the agent's relative sensitivity to unit increases in inaccuracy in the direction of different types of error, its first derivative reflects, instead, the agent's local sensitivity to risk as a function of her current credence. It is a measure of marginal increases/decreases in risk. For example, the derivative of the risk associated with the Brier score is $2p - 1$. As a result, marginal changes in credence away from the risk-free point lead to a constant increase in risk, as Figure (3.9) shows.

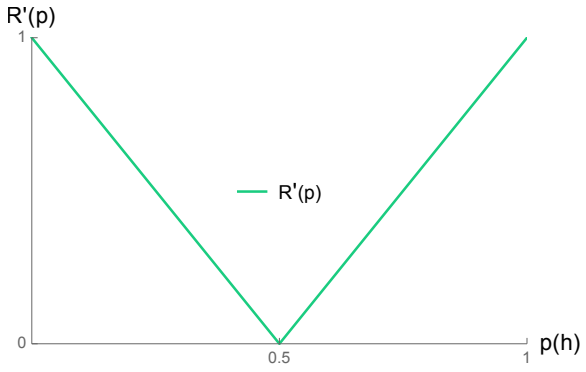

Figure 3.9: Constantly increasing epistemic risk aversion

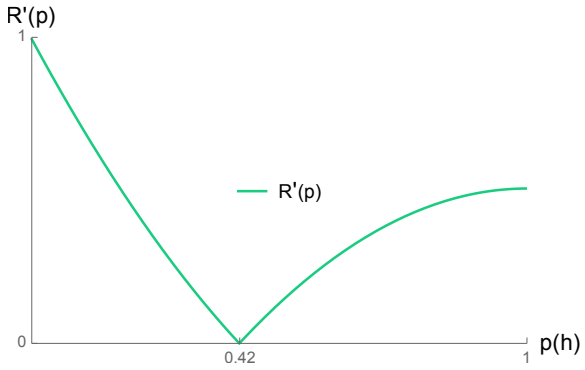

Figure 3.10: Unequally increasing epistemic risk aversion

If we let ΔFP stand for marginal increases in false positive inaccuracy and ΔFN stand for marginal increases in false negative inaccuracy then a symmetric risk function (such as the Brier score's) implies that $\Delta FP = \Delta FN$.

By comparison, the derivative of the risk associated with the asymmetric score we have been considering is $-(3/2)p^2 + 3p - 1$ (Figure 3.10). For this score, marginal changes in credence away from the risk-free point in the direction of a false negative error lead to bigger changes in risk relative to marginal changes in credence away from the risk-free point in the direction of a false positive error. The agent applying this particular asymmetric score is more worried about marginal increases in false negative inaccuracy than she is about marginal increases in false positive inaccuracy. For this particular asymmetric risk function, $\Delta FN > \Delta FP$. This corresponds to the example described above – where h is so important that rejecting it leads to substantial epistemic opportunity cost.

Moreover, marginal increases in risk taper-off as the agent approaches categorical false positive error. This makes sense from a Bayesian perspective of scientific inquiry, since having credence $.05$ in a true and important proposition is not that different from having credence $.01$ in the same proposition. In both cases, the agent will likely not pursue the idea further. There is no hard "cut-off" point of the sort significance levels play in Frequentist inference. Meanwhile, given her concern about false negative error, her anxiety in that direction persists, leading to near constant marginal changes in risk across the whole $[0, .42)$ sub-interval.

We can see this dimension of the agent's attitude to risk in the second derivative of the risk function. $R''(p)$ is equal in absolute value to $h'(p)$. This function, $h'(p)$, is what [Gibbard(2008)] identifies as an indicator of the **urgency** the believer ascribes to getting credences right, by her lights, in the vicinity of p (9). For the Brier score $R''(p) = 2$. No matter where the agent's credence is on the unit interval, her local sensitivity to being mistaken remains the same. For our asymmetric score, $R''(p) = 3 - 3p$. This is exactly what we described in the previous paragraph. This is a constantly decreasing function from 0 to 1. The agent's peak local sensitivity to error occurs at categorical false negative error and slowly tapers off as she approaches false positive error. Given the sensitivity of this particular score to false negatives that is to be expected because $p(h) = 1$ is where false negatives are eliminated altogether.

One might wonder whether this is a reasonable attitude to false positive error. But this example should not be taken as an endorsement of this particular risk function. Rather, I use it to illustrate the flexibility of the proposed approach to capturing a wide range of attitudes to epistemic risk. By comparison, consider a logarithmic risk function, whose second derivative is $1/[p(1-p)]$. In this case, we have increasing marginal risk aversion

as we move away from the risk-free point in either direction. The concavity of the risk function resembles in some respects the Arrow/Pratt measure of risk aversion for ordinary economic prospects, where the normalized second derivative reflects an agent's relative sensitivity to ordinary risk of monetary loss [Pratt(1964), Arrow(1971)].[8]

3.5 Risk and generalized entropy

When Equation (3.1) is minimized at $b = p$ (i.e., the scoring rule is proper) it may be re-written as follows,

$$\mathrm{E}_p[s_v(p)] = ps_1(p) + (1-p)s_0(1-p) \tag{3.2}$$

Following [Grunwald and Dawid(2004)], I refer to this function, $\mathrm{E}_b[s_v(p)]$ in which $b = p$, as $H(p)$, the **general entropy**. Let me explain why, as this will be relevant to extending our measure of risk to larger sample spaces.

Suppose $w(p)$ is a measure of information conveyed by learning that the event h occurs with probability p. What conditions should w satisfy? This is the question [Shannon(1948)] seeks to answer. His famous result is a representation theorem showing that the logarithmic construction $w(p) = k\log(p)$ uniquely satisfies several intuitively plausible constraints on a measure of information – namely, that w should be a decreasing, continuous, and additive function of p. By the same token $-w(p)$ measures a lack of information and Shannon entropy, H, is the expectation of $w(p)$ with $k = -1$.

In the binary case, Shannon entropy becomes $-[p\log(p) + (1-p)\log(1-p)]$. This is equivalent to the expected inaccuracy of the additive log score, defined as $\sum_S \log(v - p)$, which is proper. So Shannon entropy is the entropy associated with the log score in particular. But we can think more generally about the entropy function H associated with other proper scoring rules – the weighted average of a different function of the probability. The notion of entropy is an important building block in epistemic utility theory because [Savage(1971)] gives us a recipe for deriving proper scores from entropy by showing that every twice differentiable concave entropy function corresponds to a proper scoring rule, as follows,

$$s_v(p) = H(p) + (v-p)H'(p) \tag{3.3}$$

where v is the $0/1$ truth-value for the event in question. This relationship is extremely

[8] It has been noted in the literature that the convexity of a scoring rule implies aversion to epistemic risk in the following sense: suppose an agent is offered a pill that would, with equal probability, raise or lower her credence in h by $k \in [0,1]$. If the scoring rule is convex, such a pill would look unattractive in expectation because losses are weighted more heavily than gains [Joyce(2009)].

useful. As long as we start from a twice differentiable $H(p)$ concave on $[0, 1]$ we can derive a continuous, truth-directed, strictly proper score.

The entropy function H is closely related to our measure of epistemic risk, R. For example, for the Brier score, risk is equal to $p^* - p(1-p)$ whereas entropy is $p(1-p)$. This relationship is depicted in Figure (3.11). Meanwhile, for our asymmetric score risk is $p^* - p(p-1)(p-2)$ whereas entropy is $p(p-1)(p-2)$. We can see this in Figure (3.12).

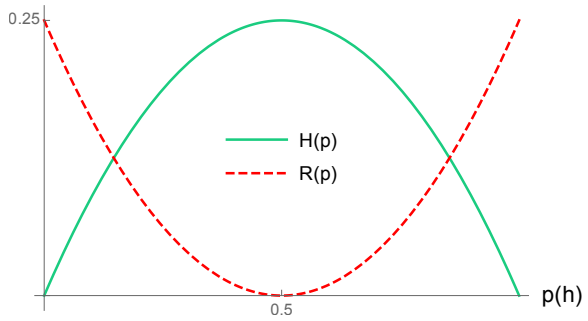

Figure 3.11: Risk/entropy duality (symmetric)

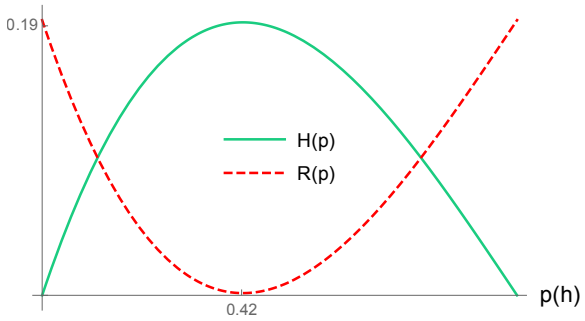

Figure 3.12: Risk/entropy duality (asymmetric)

The following theorem establishes that this duality between entropy and epistemic risk holds for all strictly proper scoring rules.

> **Theorem 1.** For strictly concave and twice differentiable entropy function H and risk function R defined on $[0, 1]$,
>
> $$R(p) + H(p) = k \text{ where } k = \min_p R(p) = \max_p H(p) \tag{3.4}$$
>
> Proof in Appendix.

In other words, the sum of risk and entropy is constant. Or more informally,

$$Risk + Entropy = k$$

In general, therefore, entropy is a scaled reflection of epistemic risk around the risk-free point $p^*(h) = k$, as figures (3.11) and (3.12) suggest. But the risk-free credence is also the maximum entropy credence. That is, $\min_p R(p) = \max_p H(p)$. Therefore, rearranging the duality equation suggests that epistemic risk may be expressed as a measure of entropic change from the maximum entropy credence to the target credence: $R(p) = H(p^*) - H(p)$. We will use this definition below to develop a general measure of epistemic risk.

Since epistemic risk is dual to entropy one might question whether we need to introduce a notion of risk, given the already large literature on entropic inference.[9] Rather than speaking in terms of increases in epistemic risk, we could instead describe the same changes in terms of decreases in entropy. While this is true for proper scoring rules, with the effect that risk and entropy are co-extensive, they are independently motivated. We saw this while developing the notion of epistemic risk in terms of sensitivity to different types of graded error. That is, I am not arguing that the risk-free distribution is risk-free *because* it maximizes entropy. Rather, it is risk-free, as we saw, because it eliminates variability in terms of epistemic outcome. Strictly proper scoring rules have the feature that these two properties do not come apart. For many other scoring rules, we could eliminate variability without maximizing entropy. In such cases, the duality would not apply and we could not measure epistemic risk in terms of entropic change.

Therefore, even though risk and entropy are extensionally equivalent for proper scoring rules, thinking in terms of risk minimization is conceptually very different from thinking in terms of entropy maximization. An agent might prefer risk-free credences not because they do not go beyond the evidence, even though that might be true, but because from her perspective they give her the best balance of graded error costs, a uniquely epistemic concern. There is a conceptual difference between thinking in terms of minimizing the amount of information an agents brings into the inference problem (the entropic interpretation) and identifying an appropriate trade-off between different types of potential mistakes (the risk interpretation). As a result, we should not think of one concept being reducible to the other. The duality theorem shows that entropy and risk are two different ways of conceptualizing

[9]For example, [Jaynes(1957a), Jaynes(1957b), Jaynes(2003)] defends maximum entropy methods for identifying priors, whereas [Williamson(2010)] goes further and defends updating by maximizing entropy as well. [Gaifman and Vasudevan(2012)] object to using entropic approaches to understand (dynamic) epistemic risk, using [Van Fraassen(1981)]'s Judy Benjamin Problem to argue that the maximum entropy posterior is not least risky. [Seidenfeld(1986)] contains a thorough discussion of the relationship between Bayesian epistemology and entropic methods.

the same underlying epistemic facts.

Indeed, insofar as proponents of entropic methods reference risk, it is assumed that a credence function is risk averse *because* it maximizes *Shannon* entropy. Jaynes is the most vivid proponent of this position. For Jaynes, the maximum entropy distribution is the most conservative distribution in the sense that it does not permit us to draw any evidentially unwarranted conclusions because it is "as smooth and spread out as possible" subject to the data [Jaynes(1963)]. But consider an entropy function that reaches its maximum at $p(h) = .9$. An entropy maximizing agent with this function would not be conservative at all in Jaynes's sense. In the absence of *any* data, she would predict h's occurrence with high confidence. Therefore, for asymmetric risk functions the least risky distribution will not be maximally uniform.

3.6 Epistemic risk: the general case

So far we have considered credence functions for a single proposition h. Now, let the sequence $\{h\}_{i=1}^{n}$ form a partition on sample space S. The objects of epistemic risk are probability distributions on S. The risk-free distribution becomes the distribution which solves the equation $s_v(p_i) = s_w(p_j)$ for all i, j and indicators of truth-value v, w. Since this is difficult to visualize (and calculate) beyond three dimensions we will use Theorem (**??**) to identify this point as the point of maximum general entropy. Since entropy is the expected inaccuracy of a strictly proper scoring rule, expressing risk in terms of entropic change enables us to harness helpful properties of expectation.[10]

To make use of these properties, we will need to introduce the notion of a random variable and its cumulative distribution function (cdf). A cdf is just a different way of expressing a probability distribution. Let $X : S \to \mathbb{R}$ be a random variable that maps outcomes in the sample space to the real numbers, and whose mass/density is given by $f(X = x)$. For example, if the random quantity X represents the numerical outcome of a single toss of a die, then the realized outcome x may take on integer values from 1 to 6. If the die is fair, then $f(X = x) = 1/6$ for every value of x. Meanwhile, for each value of x the cdf, defined as $F(X \leq x)$, tells us the probability that X is less than or equal to that value. That is, the cdf $F(X \leq x) = \sum_{x_i \leq x} f(x)$ (for discrete X) and $\int_{-\infty}^{x} f(t)dt$ (for continuous X), where $f(x)$ is the mass/density.

Notice that for our purposes every outcome may be described in terms of the agent's

[10] As propriety is generally accepted in the literature, I proceed by limiting my attention to strictly proper scoring rules. But we should keep in mind that the general approach to epistemic risk I have developed here does not necessarily depend on propriety. Rather, these scoring rules have some nice simplifying properties that enable us to express epistemic risk in terms of entropic change.

inaccuracy if that outcome occurs, where inaccuracy is measured by a scoring rule. Therefore, we can define outcomes in terms of random variables as follows: let X be a random variable that maps outcomes from the sample space to the real numbers, where the real numbers represent inaccuracy given by s_v. For every valid probability distribution on S, call it $p(h)$, there exists an induced probability distribution on X, which we will call $f(x)$, that is likewise valid. $f(x)$ is the ordinary mass/density function for random variable X. The possible values of the random variable now represent inaccuracy scores. Many scoring rules will take values on a small sub-interval of $\mathbb{R}$. For example, under the Brier score all outcomes are mapped to $[0, 1]$.

Changing the underlying scoring rule will rescale the random variable. Therefore, when evaluating distributions in terms of their epistemic risk, we need to identify a random variable X which describes outcomes in terms of some particular measure of inaccuracy s_v. We can now define the risk-free cdf.

> **Risk-free probability distribution.** Let $W \subseteq \mathbb{R}$ be the set of values that a specific scoring rule s_v can take. Given a random variable mapping outcomes from the sample space S to inaccuracy given by s_v, $X : S \to W$, the risk free cdf P^* satisfies $\arg\max_F H_F(X)$.

To simplify expression, I will denote the entropy of a distribution P as $H(P)$, keeping in mind that this is the entropy of X whose distribution is given by P. As I emphasized above, P^* is not risk-free because it maximizes entropy. Rather, this is the probability assignment that eliminates variability in terms of epistemic outcome, which is how we defined the risk-free credence in the simple case. We can now extend our definition of epistemic risk as follows.

> **General epistemic risk.** Given a random variable $X : S \to W$, where W is defined as above, let $P^* = \arg\max_P H(P)$. Then Theorem (1) implies that the epistemic risk of another cdf P is given by $R(P) = H(P^*) - H(P)$.

Recall that in the simple case, this definition was motivated as a measure of the "spread" between the agent's inaccuracy if the proposition is true, and her inaccuracy if the proposition is false. It remains to be shown that the general definition given here is motivated by the same underlying conceptual framework.

To see that this is indeed the case, I draw on [Rothschild and Stiglitz(1970)]'s notion of a mean preserving spread. Informally, one probability distribution is a mean preserving spread of another if the second is a transformation of the first obtained by pushing probability mass/density to the tails of the distribution without affecting its expected value. In

the case of ordinary economic lotteries, distributions are given in terms of wealth. For example, a lottery that pays $\$0$ or $\$10$ with equal probability is a mean preserving spread of one that guarantees a payment of $\$5$.

In the epistemic context, however, the outcomes of a "lottery" can not be specified exogenously. Rather, the scale (i.e., scoring rule) is exogenous, but the outcome, given in terms of that scoring rule's inaccuracy, depends on the probability assignment itself. For example, assuming the Brier score, a credence $p(h) = .8$ in a single proposition h is effectively an epistemic lottery that pays $(1 - .8)^2 = .04$ if h is true and $(0 - .8)^2 = .64$ if h is false (this is a lottery where less is more). Now consider a riskier credence like $p(h) = .9$. The latter is a probabilistic spread of the former because it is a transformation accomplished by taking the probability assigned to h and making it even more extreme while at the same time taking the probability assigned to $\overline{h}$ and making it correspondingly more extreme in the other direction. Assuming the agent is coherent, as we have been doing, there is a quantity that is preserved every time we make a credence riskier as we just did – namely, the simple mean given by $1/|S|$, where $|S|$ is the length of the partition. As long as we keep this quantity fixed, every probabilistic spread as just described guarantees an increase in risk. In this sense, a mean preserving spread of a credence function implies an increase in that credence function's epistemic risk. By expressing a credence function in terms of its cdf, we can give a general definition of mean preserving spreads and prove this relationship.

For example, suppose $\{h_1, h_2, h_3\}$ is a partition on S and we want to measure the epistemic risk of credence function p given by $\langle 1/5, 3/5, 1/5\rangle$ under the Brier score. Since the Brier score is $0/1$ symmetric we know that its risk-free credence function $p*$ is the uniform $\langle 1/3, 1/3, 1/3\rangle$. To evaluate the spread of our target credence function from the risk-free credences, we write the cdfs of both credence functions, P and P^*, as follows.

$$P^* = \begin{cases} 0 & \text{for } x < (1/3)^2 \\ 2/3 & \text{for } (1/3)^2 \leq x < (2/3)^2 \\ 1 & \text{for } x \geq (2/3)^2 \end{cases}$$

$$P = \begin{cases} 0 & \text{for } x < (1/5)^2 \\ 4/5 & \text{for } (1/5)^2 \leq x < (4/5)^2 \\ 1 & \text{for } x \geq (4/5)^2 \end{cases}$$

Figure (3.13) below depicts the plot of each cdf. The arrows indicate the spread in probability generated by moving from P^* to P. This is harder to visualize for discrete cdfs. To make the idea more intuitive, Figure (3.14) depicts two arbitrary cdfs of a continuous

random variable X where one is a mean preserving spread of the other.

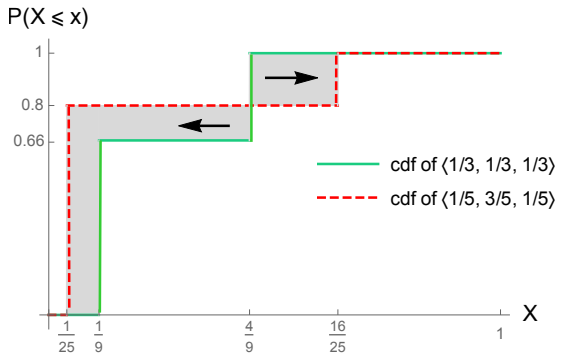

Figure 3.13: Mean preserving epistemic spread (discrete)

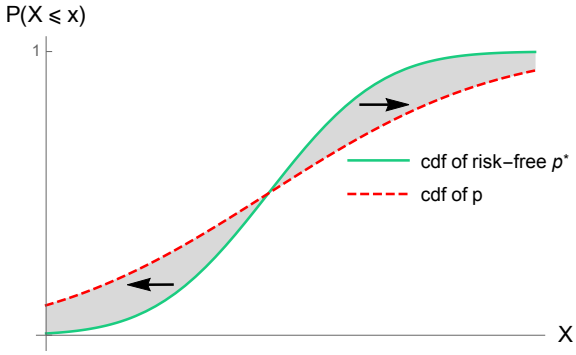

Figure 3.14: Mean preserving epistemic spread (continuous)

Notice that for any value of X, representing inaccuracy, the area underneath the dashed (risky) curve is greater than or equal to the area underneath the solid (safe) curve. Following [Rothschild and Stiglitz(1970)], we can use this quantity to define mean preserving epistemic spreads.

> **Mean preserving epistemic spread.** Given a random variable $X : S \to W$, where W is defined as above, let P and Q be two cdfs. Then Q is a mean preserving epistemic spread of P if, for all x, $\sum_{i=0}^{x} P(t_i) \leq \sum_{i=0}^{x} Q(t_i)$ (if X is discrete) and $\int_0^x P(t)dt \leq \int_0^x Q(t)dt$ (if X is continuous).

In the single proposition case this implies that one probability $q(h)$ is a mean preserving epistemic spread of another probability $p(h)$ if $|s_1(p) - s_0(p)| < |s_1(q) - s_0(q)|$. This is consistent with our definition of epistemic risk in the simple case as the integral of the absolute difference between s_1 and s_0. Therefore, by using mean preserving epistemic spreads to measure risk, we measure the difference in area underneath the risk-free cdf and

the target cdf. In Figure (3.15), below, this is the difference of the two rectangles labeled A and B.

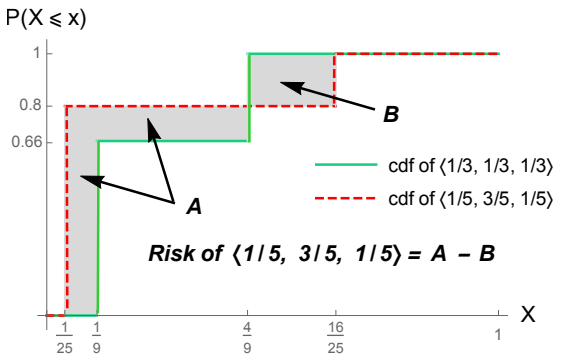

Figure 3.15: Epistemic risk as entropic change

This measure of epistemic risk, in terms of the change in area underneath the cdf, developed by analogy to [Rothschild and Stiglitz(1970)]'s approach to ordinary risk, preserves the motivation given for measuring epistemic risk in the simple case as sensitivity to approaching different types of error. In the general case, however, epistemic risk reflects an agent's sensitivity to graded inaccuracy with respect to any given outcome in the sample space. As a result, we no longer have Type I and Type II errors only. Instead, we have n error types for $|S| = n$ possible outcomes.

We are now in a position to show that our definition of epistemic risk in terms of entropic change corresponds to the general interpretation of epistemic risk given in terms of mean preserving epistemic spreads. For any given cdf P, as the area underneath it, given by $\sum_{i=1}^{n} P(x_i)$ (for discrete X) or $\int_X P(x)dx$ (for continuous X), decreases, the quantity $1 - \sum_{i=1}^{n} P(x_i)$ (for discrete X) or $1 - \int_X P(x)dx$ (for continuous X), increases. In Figure (3.15), for example, for each cdf, this is the area to its left and bounded above by the line $P(X \leq x) = 1$. This quantity, which is essentially the anti-cumulative given by $P(X > x)$, is equal to the expectation of X, ordinarily defined as $\sum_{i=0}^{n} x_i p(x_i)$ (for discrete X) or $\int_X xp(x)dx$ (for continuous X) where p is the ordinary mass/density. This relationship is a consequence of Fubini's Theorem. Importantly for us, since X maps outcomes to inaccuracy scores, the expectation of a random variable X with cdf P is precisely the entropy of P, $H(P)$, provided the underlying inaccuracy scale given by s_v is proper. Furthermore, on any given sample space S, the risk-free cdf will be the cdf that has the smallest area underneath it. Equivalently, it will be the cdf that has the *largest* area to its left. We can see this in Figure (3.15). In other words, the risk-free credence is the maximum entropy credence. Again, however, it is risk-free not because it maximizes entropy, but rather because this is the point where the agent's sensitivity to graded error in the di-

rection of every possible outcome in the sample space is equal. And again it turns out, as in the simple case, that for strictly proper scores this point is also the point that maximizes entropy. Therefore, as measured in terms of mean preserving epistemic spreads, risk may be given as the difference between the entropy of the risk-free cdf and the target cdf. This is precisely the quantity $A - B$ in Figure (3.15) and it corresponds exactly to how we have defined epistemic risk, as $H(P^*) - H(P)$. This leads to the following theorem, which is now unsurprising.

> **Theorem 2.** Given a random variable $X : S \to W$, where the underlying scoring rule s_v is proper, and two cdfs P and Q, if P is a mean preserving epistemic spread of Q then $R(P) > R(Q)$.
>
> Proof in Appendix.

As a result, every mean preserving epistemic spread increases variability in the underlying outcomes, increases risk, and (if s_v is proper) decreases entropy.

For example, take the credences we have been considering on a sample space with three outcomes, P^* and P, whose cdfs are given above. To measure the risk of P we first determine the entropy of the risk-free P^*. The area to the left of its cdf is a sum of two rectangles: one of length $1/3$ and width $2(1/3)$ and another of length $2(1/3)$ and width $1/3$. This is $4/9$. Next, we determine the entropy of P. Tracing the same approach, we get $8/25$. Since risk is given in terms of entropic change the risk of P is $4/9 - 8/25 = .12$. We would get the same result by integrating the density between the scores.

Since the approach we have developed requires identifying an inaccuracy scale before evaluating the risk of a credence function, one might reasonably wonder how general the risk ordering of credence functions will be. For example, suppose we have the same two credence functions as in the previous paragraph, $P^* = \langle 1/3, 1/3, 1/3\rangle$ and $P = \langle 1/5, 3/5, 1/5\rangle$, but we define epistemic outcomes logarithmically instead of quadratically. That is, the x-axis now measures inaccuracy in terms of the log score. The y-axis remains the same. Would it still be the case that $R(P) > R(P^*)$? If so, would the risk order be preserved for any arbitrarily chosen set of cdfs?

For most families of scoring rules considered in the literature, including some improper scores, the risk ranking of credence functions will be consistent. This includes the Brier, log, spherical, and absolute value scores. But it does not include the asymmetric score we have been considering throughout. This is because the asymmetric score has a different risk-free point and risk is measured in terms of deviation from the risk-free point. Of course, if we take two asymmetric scores with the *same* risk-free point, wherever it happens to be, then it is very likely the risk ordering will be consistent between them. More

generally, for any two scoring rules, if they reach their local minimum on the unit interval at the same point (i.e., they have the same risk-free point), *and* their risk function is convex, then the risk-order of credence functions between them will be consistent.

> **Theorem 3.** Given a random variable $X : S \to W$, where $W \subseteq \mathbb{R}$ contains inaccuracy scores measured by a scoring rule s_v, let $V = \{P_1, ..., P_n\}$ be a set of cdfs for X. Given a random variable $Y : S \to W^*$, where $W^* \subseteq \mathbb{R}$ contains inaccuracy scores measured by a different scoring rule s_v^*, let $U = \{Q_1, ..., Q_n\}$ be a set of corresponding cdfs for Y. This means that for each outcome $h \in S$, the probability assigned to h by P_i is equal to the probability assigned to h by Q_i, but whereas in the first case the outcome h is described by s_v in the second case it is described by s_v^*. Suppose (1) s_v and s_v^* are truth directed scoring rules, whose risk functions R and R^* are such that (2) $R'' > 0$, $R^{*''} > 0$, and (3) $\arg\min R = \arg\min R^*$ on the unit interval. Then $R(P_i) > R(P_j)$ if and only if $R(Q_i) > R(Q_j)$.
>
> Proof in Appendix.

This result expands the reach of the approach to epistemic risk we have developed to the vast majority of commonly considered families of scoring rules.

That is not to say, however, that all information encoded in the risk function will be preserved across different scoring rule transformations of it. Take the Brier and log risk functions, for example. While they are both convex and share the same risk-free point, their derivatives look very different. As a result, while a Brier-to-log transformation preserves an agent's risk ordering it does not preserve their attitudes to unit changes in inaccuracy nor does it preserve their local sensitivity to marginal changes in risk. We could have two agents who rank two prospective credence functions equally in terms of risk, yet while one agent finds that degree of risk tolerable, the other considers it to be inappropriate, because of differences in the way in which they evaluate the potential cost of increasing graded inaccuracy in the direction of any given outcome. This is to be expected, however. We would not want a risk function that erases well-known differences between these scores. As [Selten(1998)] emphasizes, the log score is hypersensitive in the sense that one's inaccuracy goes to infinity as the probability assignment goes to 0 or 1. This hypersensitivity is reflected in the curvature of its associated risk function.

3.7 Risk, priors, and the principle of indifference

By developing a theory of accuracy dominance [Joyce(1998), Joyce(2009)] gives us a powerful tool for evaluating the quality of an agent's beliefs. The theory of epistemic risk we have developed here enables us to go further in terms of our understanding of the normative dimensions of an agent's doxastic state. One might ask how these attitudes to risk will manifest themselves, however. After all, nearly everyone in the literature agrees that an agent should choose the credence function that, in light of her evidence, minimizes her expected inaccuracy. As a result, attitudes to risk are not going to play a direct role in one's choice (fictional or otherwise) of what to believe. But this is not the role of risk in ordinary utility theory, either. We do not consult our sensitivity to risk in order to make a choice. Instead, our choice reflects our attitudes to risk. Roughly the same is true in the epistemic case. However, epistemic risk attitudes can play a more direct role at the beginning of one's epistemic journey: in particular, they affect the agent's selection of an appropriate prior. But since an agent's prior influences their subsequent beliefs, it is in this sense that the agent's attitudes to epistemic risk *will* ultimately affect the beliefs they hold at any given time – though evidence will gradually dilute their effect.

Identifying an appropriate prior can be difficult, especially in the absence of any information. The so-called Laplacean principle of indifference (or, principle of insufficient reason) is often given as a crude guide for selecting priors under conditions of ignorance: if you do not have information to privilege one outcome over others given a partition of the sample space, you should assign equal probability to each outcome. It is assumed, therefore, that given an appropriate partition of the sample space the POI recommends the uniform distribution. The most well-known problems with this principle stem from its association with uniformity. In particular, the guideline is not partition invariant. The uniform distribution over one partition may be logically inconsistent with the uniform distribution over a simple transformation of that partition.[11] Epistemic risk provides a new perspective on the principle of indifference. In particular, whether or not the distribution recommended by the POI is uniform depends, as I emphasize below, on the agent's underlying risk function. By recasting the POI as a risk minimization principle we can identify the normative commitments presupposed in its endorsement.

[Pettigrew(2016b)] argues for the POI from considerations of accuracy and a minimax decision rule. In particular, he shows that if an agent seeks to minimize her worst case inaccuracy under the Brier score, then in the absence of information she should apply the

[11] For example, as John Venn first observed a uniform distribution over X is not a uniform distribution over X^2. [Van Fraassen(1989)] makes this point vividly using the example of a box whose dimensions are unknown and may be measured in terms of side length or volume.

POI and select a uniform prior. On the approach we have developed, a more general result follows. In particular: if an agent seeks to minimize a symmetric and convex epistemic risk function then in the absence of information she too should apply the POI and adopt a uniform prior. Our result holds not only for the Brier score but also for the log, spherical, absolute value, and many other families of scoring rules. The proof of this is trivial. Risk-free credences are minimax optimal, i.e., they minimize worst case inaccuracy, because they *guarantee* a certain inaccuracy outcome. And a convex symmetric risk function implies a scoring rule that minimizes risk at $p(h) = 1/n$ for all h where $n = |S|$. Therefore, for all such scores, the risk-free credence function will assign equal probability to every outcome in the partition. But this is not an argument for the uniform prior in the absence of information. Rather, it tells us that every epistemic risk function has its own scoring rule and its own principle of indifference. Sometimes the POI endorsed prior is uniform, other times it is not. Therefore, whether or not an agent identifies the uniform prior as optimal – where optimal can mean minimax optimal, risk-free, or more generally rationally permissible – depends on the shape of her epistemic risk function.

Suppose we start with the asymmetric risk function we have been considering throughout, given by $p^* - p(p-1)(p-2)$, and seek to identify an appropriate prior. Given this risk function, the risk-free prior in a single proposition case, as we saw, is $p(h) = .42$. *This* is the maximally non-committal prior for such an agent, because it guarantees a particular outcome in terms of inaccuracy. If an epistemically risk averse agent has a risk function that takes this form, she will not adopt a uniform prior. From the epistemic risk perspective, therefore, having uniform credences is not the same as having maximally non-committal credences. As a result, the agent's attitudes toward epistemic risk will determine the prior she deems appropriate and her interpretation of what the POI recommends. Uniformity is a special case of agents with symmetric risk functions who do not discriminate among mistakes. Such indifference between different ways of being mistaken is sometimes appropriate but more often it is not.

The notion, due especially to [Jaynes(1957a), Jaynes(1957b)], that the right prior is to be found by identifying the maximum entropy distribution is a combination of two separate normative principles: (a) that one ought to minimize epistemic risk, and (b) that one ought to evaluate epistemic risk using a convex symmetric risk function. What I have done here is to distinguish the two principles: *even if* we agree that minimizing epistemic risk is desirable, the appropriate prior may not be uniform. Therefore, the Jaynesian commitment to maximum entropy priors is a commitment to a particular attitude to how much risk is rationally permissible (as little as possible) and how different types of errors are to be evaluated (equally). These are very strong normative assumptions which, despite the size

of the literature on the problem of the priors and the principle of maximum entropy, have not been adequately addressed.

CHAPTER 4

Dynamic Epistemic Risk

4.1 Introduction

In the preceding chapter, I argued that entropic change is a plausible measure of risk associated with a probability distribution. Let P be a probability distribution on a discrete finite sample space S and p its density where it exists. Then P's riskiness is given by $H(P^*) - H(P)$ where H is a generalized measure of entropy and P^* is the corresponding maximum entropy probability distribution.

In this chapter, I address a dynamic question: given a prior distribution P, what is the least risky posterior probability Q, after undergoing some learning experience? More generally, how can we measure the epistemic riskiness of an update rule or, equivalently, the epistemic riskiness of a posterior probability relative to some fixed prior?

4.2 The formal framework

I investigate learning in the context of acquiring information about a simple binary process. To fix ideas, suppose three agents, A, B, and C, are interested in formulating a credence about the bias of a coin. Before seeing any evidence, they estimate the coin's bias as $1/2$ (i.e., they start with a weak assumption of fairness). However, while their valence about the coin's bias is the same, the resilience of their credence, as we will see below, may be different. Moreover, while A updates by Bayesian conditioning, B and C apply a different rule for updating beliefs.

Together, they will witness a sequence of coin tosses with a coin whose objective chance of landing on heads or tails is unknown. We want to explore how prior information and updating rules affect the riskiness of their learning procedures and their estimates of the coin's bias. We have not said what it means for learning or, more specifically, updating, to be more or less risky yet. This remains to be seen.

Suppose more generally that $\{X_n\}$ is an independent and identically distributed (iid) sequence of Bernoulli random variables that can take the value 0 (e.g., for heads) or 1 (e.g., for tails). Then $\Omega = \{0,1\}_{n\geq 1}$ is the set of all possible sequences that might be observed.

One might object to the iid assumption. In general, this assumption will not be necessary for the account of dynamic risk I intend to develop. Rather, I use it as a heuristic for parametric modeling of stylized learning behavior. The model is not a realistic description of learning, but it is useful for helping us to consider our judgment about different ways of processing information. The iid assumption does, however, affect the asymptotic behavior of different updating rules, as we will see below. We can, however, recover most of the results with exchangeability alone.

In any case, let $Y = \sum_i X_i$ be a random variable that represents the sum of successes (in this example, the frequency of tails). Then Y follows a Binomial(n, p) distribution. Let A, B, and C be Bayesian agents endowed with coherent prior probability distributions about the bias of the coin which they will revise in response to new evidence. The coin's bias corresponds to the limiting mean of the sampling distribution: $EX \to p$ in probability as $n \to \infty$. It is a consequence of DeFinetti's representation theorem that this quantity exists with probability 1 [Zabell(2005b)]. Therefore, the unknown parameter of interest θ is equal to p in the binomial example we are considering. From here on, we will be talking about estimating θ, which should be understood as the limiting mean of the sequence. Or the true mean. Or the population mean. How we conceptualize the quantity to be estimated depends on one's metaphysical understanding of probability, but I would like to set that aside for now. In any case, θ is the unknown quantity we wish to estimate.

There is a more important metaphysical issue, however. Extreme subjectivists often avoid talk of estimating 'true but unknown' quantities altogether, and replace them with inter-subjective agreement in degree of belief. I will not do this, but I do not think anything in my argument turns on it. Another way of putting the objection might be to say that a point estimate is conceptually not Bayesian. But we can imagine situations where a Bayesian may need to make such an estimate: for example, an expert has tossed the coin many thousands of times, and is offering A, B, and C a prize for making the most accurate guess about the expert's observed large sample mean. This way of framing the problem treats objective chance as an expert, which is compatible with Bayesian characterizations of the Principal Principle [Lewis(1980)]. In any case, our Bayesian estimator, $\hat{\theta}$, will be a function of the posterior distribution. Our agents will each have an estimator, given by $\hat{\theta}_i$ for agent i that gives their best estimate of θ. This gives us a very simple example to work with.

The conditional distribution of X given $Y = y$ is equal to $1/\binom{n}{y}$. This can be interpreted

as the probability of tails on the next toss, given the proportion of heads that has occurred up to now. This quantity does not depend on $\theta = p$. Applying the Factorization Theorem, Y is therefore a sufficient statistic for estimating θ [Halmos and Savage(1949)]. In general, the Bayesian estimate given a sample of the data is equivalent to the Bayesian estimate given a sufficient statistic of the data. Therefore, in thinking about how our agents will revise their beliefs about θ, we can confine our attention to updating on Y. Indeed, the claim that our estimate of θ can be made on the basis of Y is really a special case of W.E. Johnson's sufficientness postulate [Johnson(1924), Zabell(1982)].

This revision of beliefs will be the subject of our investigation. A will revise her beliefs by Bayesian conditioning whereas B and C will apply a different update rule. Let $\pi_i(\theta)$ represent agent i's prior distribution for θ (subscripts are omitted where unnecessary). Further, let $f(\mathbf{x}|\theta)$ represent the empirical distribution. Then,

$$\pi_i(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta)\pi_i(\theta) \tag{4.1}$$

is agent i's posterior distribution after seeing the data. A Bayesian agent's estimate of the coin's bias after seeing the data is typically taken to be the posterior mean. This is the ordinary Bayes estimator, given by,

$$\hat{\theta} = \mathrm{E}[\theta|y] = \int \theta\pi(\theta|y)d\theta \tag{4.2}$$

It is worth investigating whether we should indeed apply the ordinary Bayes estimator, as the usual justification for it is that it minimizes a quadratic loss function. Indeed, it is well-known that if we do not square the loss function then the Bayes estimator would be the median of the posterior distribution – this is to be expected. Since quadratic loss is more sensitive to outlying/extremal observations those observations will drag the optimal estimate up or down. But what happens if the loss function takes a different functional form – such as the logarithmic distance? Perhaps unsurprisingly, we would still get the mean of the posterior. Indeed, for all strictly proper scoring rules in a binary process the Bayes estimator is the mean of the posterior. From an accuracy-theoretic perspective this makes sense. The Bayes estimator is essentially the estimator that maximizes posterior expected accuracy.

In any case, as we said earlier, the estimate produced by $\pi_i(\theta|\mathbf{x})$ is equal to the estimate produced by $\pi_i(\theta|y)$. Since we are interested in a binary random process, it is reasonable

to suppose that $\pi(\theta) \sim Beta(\alpha, \beta)$, given by,

$$\pi(\theta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \tag{4.3}$$

where $B(\alpha, \beta)$ is the complete Beta function. If $\alpha, \beta \in \mathbf{N}^+$ then $B(\alpha, \beta) = (\alpha-1)!(\beta-1)!/(\alpha+\beta-1)!$. We also know that the sampling distribution is,

$$f(y|\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y} \tag{4.4}$$

The joint distribution is therefore,

$$f(y, \theta) = \left[\binom{n}{y} \theta^y (1-\theta)^{n-y}\right] \left[\frac{1}{B(\alpha, \beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}\right] \tag{4.5}$$

Applying Bayes' rule, this means that the posterior is,

$$\pi(\theta|y) = \frac{1}{B(y+\alpha, n-y+\beta)} \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1} \tag{4.6}$$

In the usual terminology, the beta prior is conjugate for the binomial – i.e., the posterior is isomorphic to the prior and is given by $Beta(y+\alpha, n-y+\beta)$.

Takeaway: with a beta-binomial conjugate distribution, the hyperparameters of the prior, (α, β), are equal to pseudo trials (e.g., imagined coin tosses) which determine the valence and resilience of the agent's credences before seeing any data. "Imagined" is not quite the correct term here, though I hope it helpfully illustrates the point. It may be, instead, that a person's prior about this coin toss is informed by previous coin tosses she has seen. This will depend on the context and on our judgment of whether previous experience is relevant to the experiment about to be performed. In simple coin toss cases, it probably is.

In any case, the posterior distribution is a beta distribution whose parameters are given by the sum of pseudo and real successes $(y+\alpha)$ and pseudo and real failures $(n-y+\beta)$. The posterior credence for θ (i.e., the agent's best guess about the coin's bias after Bayesian conditioning on the evidence) is then given by the posterior mean,

$$\mathrm{E}[\theta|y] = \frac{y+\alpha}{\alpha+\beta+n} \tag{4.7}$$

This gives us a very easy way to think about Bayesian updating in the context of binary sequences with Beta priors. If an agent starts with a uniform Beta$(1, 1)$ prior, and observes

two heads and three tails, then after Bayesian conditioning on the data she will end up with a Beta$(4, 3)$ posterior. Her initial estimate for θ would have been $1/2$ whereas her new estimate is that the coin has a $4/7$ bias in favor of tails.

This also gives us a very transparent way of thinking about the resilience of a credence function. If we have two agents, A and B, one with a Beta$(1, 1)$ and another with a Beta$(3, 3)$ distribution, they will agree in their estimate of the coin's bias $(1/2)$ (valence) but they will not agree on how to revise that estimate after seeing an additional toss or sequence of tosses (resilience). Indeed, A's credence will be more sensitive (less robust) to new evidence. If they see one additional toss, and it lands heads up, A will move from $\hat{\theta}_A = .5$ to $\hat{\theta}_A = .33$ whereas B's estimate will shift from $\hat{\theta}_B = .5$ to $\hat{\theta}_B = .42$. As we will see below, α and β provide a very good measure of the resilience of the estimate.

4.2.1 Beta priors and Carnap's continuum of inductive methods

The machinery we have developed here reflects Carnap's continuum of inductive methods. Note that the prior mean, i.e., the mean of the pseudo trials, is given by $\alpha/(\alpha + \beta)$. Meanwhile, the mean of the empirical distribution is simply y/n. We can write $\hat{\theta} = \mathrm{E}[\theta|y]$ as follows,

$$\hat{\theta} = \left(\frac{\alpha + \beta}{\alpha + \beta + n}\right)\left(\frac{\alpha}{\alpha + \beta}\right) + \left(\frac{n}{\alpha + \beta + n}\right)\left(\frac{y}{n}\right) \tag{4.8}$$

Since the weights on the prior and sample mean sum to 1, the posterior mean is a convex combination of the prior mean and the maximum likelihood estimator (MLE),[1]

$$\hat{\theta} = \lambda\Big(pseudo\ mean\Big) + (1 - \lambda)\Big(MLE\Big) \tag{4.9}$$

where $\lambda \in [0, 1]$. The pseudo counts determine how much weight we put on the prior – i.e., they determine the value of λ. It is easy to see from the expression of the posterior as a convex combination of the prior and the MLE how data washes out the prior. Let $S = \alpha + \beta$. Then $\lambda = S/S + n$. As $n \to \infty$, $\lambda \to 0$.

With good prior information λ should be high. With poor prior information, it should be weak or uninformative. But we have a menu of options corresponding to all values on the unit interval for determining how much weight to give to the prior. For every Beta prior, there exists a value of λ in Carnap's framework. Therefore, the pseudo trials of the

[1]Recall that $f(\mathbf{x}|p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$. The log likelihood is then $l(p|\mathbf{x}) = \log p \sum_{i=1}^{n} x_i + \log(1 - p)(n - \sum_{i=1}^{n} x_i)$, and the first derivative $\frac{d}{dp} l(p|\mathbf{x}) = \sum_{i=1}^{n} x_i/p - (n - \sum_{i=1}^{n} x_i)/(1 - p)$ which equals 0 at $p = 1/n \sum_{i=1}^{n} x_i = \overline{X}$. Checking the second derivative and boundary points will confirm this is indeed a maximum.

Bayesian's prior essentially provide a smoothing function on the MLE. In other words, there exists a function $g : \mathbf{R}^2 \rightarrow [0, 1]$ that assigns to every point in the 2-dimensional $\alpha - \beta$ plane a unique value of λ in the unit interval. The proof of this is easy to see given the expression above – λ is equal to the proportion of pseudo trials to all trials. For example, if $\alpha = \beta = 1$ then $\lambda = 2/(2+n)$. At every step in the sequence this is a determinate value. For example, if $n = 10$, then $\lambda = 1/6$: a weak Beta prior is equivalent to a low λ value. However, notice that g is not bijective. While it is true that for every value of S there is a unique λ, it is still the case that there are many different ways to sum up α and β to any particular value of S. So, for example, at $n = 10$ Beta$(2, 2)$ and Beta$(1, 3)$ assign the same value to λ, namely $2/7$. So the weight being put on the pseudo mean is the same. However, the posterior is different because the pseudo mean itself is different.

The takeaway from this discussion is that if we want to vary the amount of weight we put on the prior we need to vary the absolute value of S. The absolute value of S encodes the resilience of the prior. In Carnap's framework, therefore, λ corresponds to the resilience of the prior. There is still no flexibility, on this approach, to vary the update rule.

4.2.2 A generalized Bayes estimator

But what if we want to vary the update rule? In other words, holding the pseudo mean fixed, what if we want to update more or less quickly on the evidence, relative to Bayes' rule?

This kind of framework would enable us to describe the behavior of imperfect Bayesian agents: agents who shoot past the Bayesian update or fall short of it. Realistically, most decision-makers do something like this, and while it is possible to explain the discrepancy by postulating a difference in their priors, I want to investigate what happens if someone applies a different rule of updating without attributing the difference to the prior. Indeed, I will argue that different update rules can be more or less risky. Dynamically conservative agents will be close to the Bayesian ideal whereas dynamically liberal agents will significantly depart from it. I will introduce a measure for such risk, below.

So, focusing for now on the case of beta-binomial credence functions, how can we tune the update rule? Consider again our characters A, B, and C. Suppose they each start, at t_o, with a uniform Beta$(1, 1)$ prior. So the number of pseudo tosses is fixed at $S = 2$ and $\alpha = \beta = 1$. A updates by Bayes' rule. Now suppose B does the following: for every trial (i.e., every new data point in the sequence), B double counts it. So B is the type of character who is too quickly persuaded by empirical evidence and jumps to hasty conclusions. If they observe a success (tails) at t_1 A moves to a Beta$(2, 1)$ posterior whereas

B moves to a Beta$(3, 1)$ posterior. I have used the term 'pseudo' to refer to prior trials but we may suppose that what B does is for every trial she observes, she adds a matching 'ersatz' trial. Ersatz trials are different from pseudo trials because they are a function of Y. They are super data. Meanwhile, C exhibits the opposite flaw – C is too dogmatic to give the evidence enough weight. As a result, relative to A, C counts each toss for half the weight. In the case above, C would move to a Beta$(1.5, 1)$ posterior. C essentially adds 'antidata' to her sample, to borrow a phrase from [Rodriguez(2006)].

Let η (eta) represent ersatz successes and σ represent ersatz failures. Then the total number of ersatz trials is given by $W = \eta + \sigma$. When $W = 0$, as in the case of A, the agent is an ordinary Bayesian updater. When $W > 0$, as is the case for B, the agent is adding ersatz trials (super data). Accordingly, we will call them a super-updater. When $W < 0$, as is the case for C, the agent is likewise adding ersatz trials, but in the form of antidata. We will call them a sub-updater.

All we are doing here is treating ersatz trials exactly how pseudo trials are treated. We can now give the expression of a new, generalized Bayes estimator, as follows,

$$\hat{\theta} = \left(\frac{S}{S + W + n}\right)\left(\frac{\alpha}{S}\right) + \left(\frac{n + W}{S + W + n}\right)\left(\frac{y + \eta}{n + W}\right) \tag{4.10}$$

This is a convex combination of the prior mean and the new empirical/ersatz hybrid mean. So for B, the distribution of η is given by $2Y$. Reasonably, one might wonder how we are going to treat C? There is a natural extension, again in terms of Y. We will say that for C the distribution of η is given by $-1/2Y$. Similarly, $W = 2n$ for B and for C, $W = -1/2n$. Suppose we see the following sequence of trials,

$$1, 0, 0, 1, 1, 0, 1, 1, 1, 0$$

(6 successes and 4 failures). Assuming each agent started with a uniform prior, A will move to Beta$(7, 5)$, B will move to Beta$(13, 9)$ and C will move to Beta$(4, 3)$.

Takeaway: re-parameterizing the Beta-binomial distribution with ersatz trials allows us in effect to vary the update rule in Bayesian inference. This distribution should be considered a generalization of the Beta distribution because, again, Bayesian updating is a special case of it with $W = 0$. Indeed, the distribution takes the form Beta$(y + \eta + \alpha, n + \eta - y + \beta)$, with the following density,

$$\pi(\theta|y) = \frac{1}{B(y + \eta + \alpha, n + \eta - y + \beta)}\theta^{y+\eta+\alpha-1}(1 - \theta)^{n+\eta-y+\beta-1} \tag{4.11}$$

This expression still integrates to 1 and is a valid pdf. It is important to highlight the following here: in Bayesian inference we are treating θ as random and the evidence Y as fixed. In a sense, then, Y plays the role of a parameter. Since η is a function of Y, η is likewise a parameter in the posterior distribution. It is a parameter that tracks ersatz evidence, whereas the hyperparameters α and β track pseudo evidence, and the "parameter" Y tracks actual evidence. Every part of the distribution is, therefore, accounted for: we have the random variable θ, the hyperparameters of the prior (α, β), the parameter of the sample (Y), and the ersatz parameters (η, σ). This gives us the sought after machinery and provides more flexibility than was available either under Carnap's continuum of inductive methods or with the original beta-binomial framework: we can now tune the resilience of the prior with the hyperparameters *and* we can tune the update rule with the ersatz parameters.

One might wonder whether the approach we have developed is still Bayesian. It is certainly subjective but if we are no longer updating by applying Bayes rule in what sense is this Bayesian? From a slightly different perspective, we can see that it is, by highlighting that η plays a role that is analogous to Carnap's λ. Recall that since $\lambda = S/(S+n)$ when we vary λ we end up with agents who are Bayesian conditioning on different priors. Now consider η. For example, if $\eta = 2Y$, then the set of all possible paths that might be observed is given by $\Omega = \{0,1\}^{\infty}_{2n}$. More generally, $\Omega = \{0,1\}^{\infty}_{n+W}$. This is still a σ-algebra. It is the σ-algebra that corresponds to the generalized beta distribution we have developed. So, when we talk about our Bayesian sub-conditioners and super-conditioners, another way to describe them is to say that they are Bayesian conditioning on a different σ-algebra. So if one wants to describe such agents in terms of Bayesian conditioning then we would say that η is a tuning parameter for the underlying algebra of events. But it remains true that *from A's perspective*, B and C are not conditioning. And that is all we need to have a discussion about the riskiness of different update rules: the fact that A can meaningfully ask herself: should I condition? Or Should I do what B and C are doing?

4.2.3 Asymptotics of the generalized Bayes estimator

It is worth investigating what happens with the generalized Bayes estimator asymptotically. We can write the ordinary Bayes estimator as follows,

$$\hat{\theta} = \frac{y + \alpha}{S + n} \tag{4.12}$$

In other words, the posterior mean is the sum of pseudo and actual successes, over the total number of pseudo and actual trials. As n goes to infinity y and n swamp α and S and by the weak law of large numbers y/n converges to θ in probability.

Meanwhile, the generalized Bayes estimator we have developed here will be the sum of pseudo, actual, and ersatz successes over, again, the sum of all three types of trials, given as,

$$\hat{\theta} = \frac{y + \alpha + \eta}{S + W + n} \tag{4.13}$$

Asymptotically, this will approach $(y+\eta)/(W+n)$. A necessary condition for convergence is that $W \to 0$ as $n \to \infty$. But W is not any scalar. It is a function of S. So letting $\eta = ky$ we can more helpfully write this expression, as follows,

$$\hat{\theta} = \frac{ky + \alpha}{S + kn} \tag{4.14}$$

This will converge to ky/kn – i.e., to y/n. It must, because both estimators satisfy the Martingale condition. For example, suppose we have an ordinary and super updater. Both start from a Beta$(1, 1)$ prior and their prior estimate is $\hat{\theta} = .5$. The ordinary updater may end up with a Beta$(1, 2)$ or a Beta$(2, 1)$ prior. The super updater may end up with a Beta$(1, 3)$ or a Beta$(3, 1)$ prior. From their prior perspective, the two possibilities are equiprobable. Therefore, their expected estimate after seeing the next toss – $2/4$ and $3/6$, respectively – is equal in valence to their current estimate. Clearly, then, the generalized Bayes estimator is consistent. It converges in probability to the true mean under the same conditions as the ordinary Bayes estimator. (both estimators, by the way, are biased, and the ersatz parameters of the generalized Bayes estimator introduce additional bias.)

However, for any finite n notice that the rate of convergence depends on the update rule and the kind of evidence the agents will encounter. For example, suppose we have a fair coin. And the evidence is minimally misleading in that the outcomes alternate between heads and tails without exception. Then the agent for whom $\eta > Y$ will accurately estimate the population mean more confidently. The opposite occurs when $\eta < Y$. However, if the agents encounter misleading evidence (for example, a sequence of all heads when a coin is fair), then the agent for whom $\eta > Y$ will also be misled more quickly. Again, the opposite occurs when $\eta < Y$. In the limit, of course, the detours will wash out.

We can be more specific about the rate of convergence of the generalized Bayes estimator. Applying Chebyshev's inequality, we find that,

$$P(|kY_n/nk - \theta| \geq \epsilon) \leq \frac{Var(kY_n/nk)}{\epsilon^2} = \frac{\sigma^2}{kn\epsilon^2} \tag{4.15}$$

Which means that the estimator converges linearly in n. So the rate of convergence of the super/sub updater is a linear function of the rate of convergence of the ordinary Bayes estimator. From a slightly different perspective, we know from the Central Limit Theorem that

$\sqrt{n}(Y_n/n - \theta) \xrightarrow{d} N(0, \sigma^2)$ so $Y_n/n \xrightarrow{d} N(\theta, \sigma^2/n)$. This means that for the generalized Bayes estimator, $kY_n/n \xrightarrow{d} N(\theta, \sigma^2/kn)$. In other words, it is likewise approximately normal with smaller (larger) variance. This is to be expected because the sample size has been artificially inflated (deflated) with super data (antidata) to account for the rate of the update rule. So depending on the size of k the generalized Bayesian is more/less sure of herself in her estimate of the true mean at every finite point in the sequence.

Importantly, the following is true of super/sub updaters:

1. **Potential gain**: She stands to gain something: namely, faster convergence under sufficiently non-misleading evidence or slower convergence under sufficiently misleading evidence;
2. **Potential loss**: She stands to lose something: namely, slower convergence under sufficiently misleading evidence or faster convergence under sufficiently non-misleading evidence.

We have said before that this is the hallmark of increasing risk: namely, the prospect of a gain coupled with a chance of loss. Indeed, the point has often been made that the most conservative posterior probability is the one that moves as little as possible from the prior subject to the constraints imposed by the observed evidence. As we will see, there is good reason to take this to be the Bayesian posterior. As a result, sub- and super- updaters are risk increasing transformations of the Bayesian updater.

Before we get there, though, does the asymptotic result imply that A, B, and C cannot expect asymptotic inter-subjective agreement? Well, that depends. If agreement means consensus then they cannot expect it. But for large enough n the valence of the credence they assign to any value of θ is going to be very, very close. However, since B is essentially doubling the sample size, the resilience of her credences will be stronger – she will have a more sharply peaked distribution around the posterior mean. Meanwhile, C will exhibit the opposite behavior – since she is diluting the evidence with antidata, her distribution will be stronger tailed. Notice, however, that these are very small differences, appreciable only at a very high resolution. They are, however, small differences that persist after a very large body of evidence. So perhaps it is misleading to call this agreement at all. Despite so much evidence to update on, they will still fail to see eye to eye on the coin's bias. Of course, this is just as true for any finite n in the case of the ordinary Bayes estimator. But for the generalized Bayes estimator, as compared to the ordinary, the effect is even more pronounced.

There is something we can say, however: for any group of agents, *if* they apply the same update rule, then their estimators will converge in probability to the population mean.

What we mean by 'population' will depend on the update rule the group is using. For super-updaters, their population will include super data. For sub-updaters, their population will include antidata. For ordinary Bayesian updaters, their population will include only pseudo and actual data.

This completes our development of the basic framework. We can now turn more directly to considerations of epistemic riskiness in updating.

4.3 Dynamic epistemic risk and cross-entropy

The takeaway from GER was the following: epistemic risk is given by entropic change. Our goal is, ultimately, to develop a general theory of epistemic risk. As a result, we will now generalize this idea to a dynamic context, evaluating the riskiness of a posterior probability (or, equivalently, of an update rule). It turns out, fortunately, that the concept I seek to capture is again measurable in information-theoretic terms. The takeaway will be: *dynamic* epistemic risk is given by *cross*-entropic change. As before, however, the notion of risk will be motivated independently of its information-theoretic expression.

4.3.1 Measuring dynamic epistemic risk

Recall from earlier that where s is a score function for a probability distribution P the general entropy is the negative expectation of $s(P)$ given by $-\mathrm{E}_P[s(P)]$. Where the underlying score is logarithmic and additive, this becomes the well-known Shannon measure of information entropy.

Now suppose we want to compare the riskiness of an update rule – for example, we want to compare the learning behavior of our ordinary, super and sub-conditioners, A, B, and C. In other words, when it comes to processing information, who among them is conservative? Who is reckless? Can we rank-order their update strategies in terms of risk?

A natural way to proceed would be as follows: (1) identify some benchmark "safe" posterior distribution, relativized to a prior, and (2) measure the risk of the target posterior in terms of some notion of comparative "distance" that encodes the severity of potential gains/losses in accuracy, as described above. That is, if we want to know how risky A's update rule is, then we compare an appropriate divergence between A's prior and her posterior against the divergence between the safest, benchmark posterior from its prior.[2]

[2] As we will see below, the most common approaches to measuring the divergence between two probability distributions are not proper distances.

As we are now in a dynamic context, we first need an appropriate notion of divergence between two probability distributions, P (for prior) and Q (for posterior). Following [Savage(1971)], we will generate such a notion from the scoring rule itself. Notice that the scoring rule is really a measure of divergence of a probability distribution from the true distribution which assigns probability 1 to the true event.[3] We have so far denoted it as $s_v(p(x))$ which is the accuracy score of $p(x)$ when the true value is given by the indicator v. The measure can be quadratic, as in the Brier score, logarithmic, as in the log score, or it can take many other forms. As long as the underlying risk is convex the score will be strictly proper.

Therefore, we can think at a very general level about the score $s_{q(x)}(p(x))$, which is the score of $p(x)$ evaluated from the perspective of $q(x)$ rather than the true value v. For example, the Brier score is given by $(v-p)^2$, the Squared Euclidean distance from the true value. But we can also think about the Brier score between two distributions – i.e., their squared Euclidean distance – given by $D(P||Q) = \sum_{i=1}^{n}(p_i - q_i)^2$ where P and Q are probability vectors or $\int_\chi f(x) - g(x)dx$ where f and g are densities.

If we use the Euclidean norm to express such a "distance" we get the following expression,

$$\begin{aligned} D_{Brier}(P||Q) &= ||P-Q||^2 \\ &= ||P||^2 - ||Q||^2 - \langle 2P, P-Q\rangle \end{aligned} \tag{4.16}$$

where $\langle P, Q\rangle$ is the inner product between P and Q. Since the derivative of $||p||^2$ is $2p$ the term $||q||^2 - \langle 2p, p-q\rangle$ is the value of the tangent line to $||p||^2$ evaluated at q. Squared Euclidean distance is therefore the vertical distance at p between the graph of $F(p) = ||p||^2$ and the tangent to the graph of F in q. If F is convex, this gives us a general definition of a Bregman divergence.

Let $F : W \to \mathbb{R}$ be a differentiable real-valued function defined on a convex set W. Then the Bregman divergence is given by,

$$D_F(P, Q) = F(p) - F(q) - \Delta F(p)^T(q-p) \tag{4.17}$$

It is the difference between the value of F at p and the value of the first-order Taylor expansion of F around q evaluated at p, as shown in Figure (4.1), below.

[3] [Buja et al.(2005)Buja, Stuetzle, and Shen].

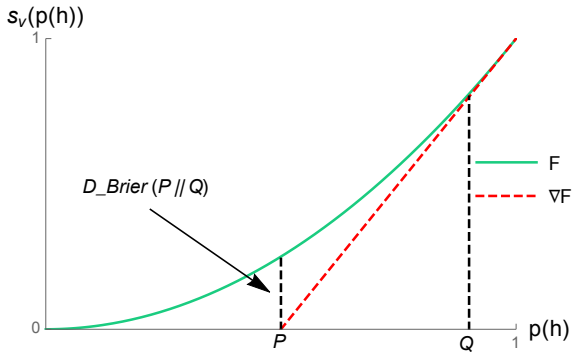

Figure 4.1: Geometric expression of divergence

The solid green line is $||P||^2$. The dashed red line is the tangent to $||P||^2$ evaluated at Q. The black vertical line segment is the Brier divergence from Q to P.

When $F(P) = ||P||^2$ we get squared Euclidean distance between two distributions (above). Now, if we let $F(P) = -H(P)$ (Shannon entropy – i..e., re-scaled logarithmic risk, as developed in GER) then the associated divergence is the Kullback-Leibler divergence, given by,

$$D_{KL}(P, Q) = \mathrm{E}[\log P - \log Q|P] \tag{4.18}$$

In other words, KL is the expected difference in inaccuracy between P and Q with the expectation taken under P. Of course we can think about divergences more generally in this way, as expected differences in inaccuracy for every strictly proper score S, as follows,

$$D_S(P, Q) = \mathrm{E}[S(P) - S(Q)|P] \tag{4.19}$$

Since expectations of strictly proper scores are concave their risk, as developed previously, is convex and it is the risk that is equal up to an additive constant to the F function in the definition of a Bregman divergence. Therefore, if the score is strictly proper the risk is convex and the associated divergence is a Bregman divergence. As a result, it is in keeping with the risk approach previously developed to use Bregman divergences as a measure of how "far apart" in terms of riskiness a posterior is from its prior, which is essentially expected difference in inaccuracy using a strictly proper score. The greater the expected difference in inaccuracy, the more salient the opportunity/chance of loss become with the posterior distribution, which is how we have been motivating increases in epistemic risk.

Indeed, most statistical measures of divergence are Bregman divergences including, for example, the commonly used Mahalanobis distance. The Bregman divergence is not generally a metric. While it is true for all divergences that $D_F \geq 0$, many are not symmetric (as in KL distance), and while some are symmetric, like squared Euclidean distance, they

do not satisfy the triangle inequality. In any case, we can now express our measure of dynamic risk.

$$DR_S(Q) = D_S(P||Q) - D_S(P||R) \tag{4.20}$$

where R is the safest or risk-free posterior and S is the underlying scoring rule. In other words, it is the 'extra' divergence that is required to go from Q to P instead of going the minimum 'distance' from R to P. The following property of Bregman divergences will be important. For all strictly proper scoring rules,

$$\begin{aligned} D_S(P||Q) &= \int p(x)(s(p(x)) - s(q(x)))dx \\ &= \left[\int p(x)s(p(x))dx\right] - \left[\int p(x)s(q(x))dx\right] \\ &= \left[-\int p(x)s(q(x))dx\right] - \left[-\int p(x)s(p(x))dx\right] \\ &= H(P,Q) - H(P) \end{aligned} \tag{4.21}$$

where $H(P,Q)$ is the cross-entropy between P and Q – i.e., the entropy of Q evaluated from $P's$ perspective. Therefore, D_S is a measure of the difference between prior entropy and the entropy of the posterior from the perspective of the prior. Our measure can then be simplified, as follows,

$$\begin{aligned} DR_S(Q) &= D(P||Q) - D(P||R) \\ &= \big[H(P,Q) - H(P)\big] - \big[H(P,R) - H(P)\big] \\ &= H(P,Q) - H(P,R) \end{aligned} \tag{4.22}$$

For all strictly proper scoring rules, $D_S \geq 0$. Dynamic epistemic risk as we saw is motivated by measuring the extra divergence one needs to go, as measured from the perspective of the prior, as against the divergence required to get to the risk-free posterior. As it turns out, this is equivalent to cross-entropic change: the cross-entropy between the prior and the target posterior against the cross-entropy between the prior and the risk-free posterior.

4.3.2 The risk-free posterior

I have so far avoided what is essentially the most important question: What is the risk-free posterior? Measuring dynamic risk depends on a risk-free benchmark. I have suggested this to be the Bayesian posterior without arguing for this claim. I will now give four arguments for the normative claim that the Bayesian posterior ought to be set as least risky.

4.3.3 An argument from accuracy

Suppose $\pi(\theta)$ is our prior. And suppose $X_1, ..., X_n$ is the data to be observed which follows $f(\mathbf{x}|\theta)$. Let $\pi^*(\theta)$ be our posterior. Let $x_1, ...x_n$ be some observed evidence. In this case, the evidence amounts to learning with probability 1 for each random variable which member of a partition is true – i.e., what is the value that it in fact has taken on. For example, if the evidence is three trials with the sequence $0 - 0 - 1$ then the constraint allows all and only posteriors that assign probability 1 to $X_1 = 0$, $X_2 = 0$, and $X_3 = 1$.

There will be a set of distributions compatible with the evidence that has been observed, namely, $W = \{\pi^*(\theta|\mathbf{x}) : m(\mathbf{X} = \mathbf{x}|\theta) = 1\}$ where $m(x) = \int \pi(\theta)f(x)d\theta$ is the marginal distribution of the evidence – i.e., the prior probability of the evidence. Again, in this context the lower-case x are supposed to be values of X that have in fact been observed, rather than hypothetically realized values of a sample.

It seems reasonable for an agent to adopt an updating rule which advises her to choose as her posterior the distribution from W that she currently believes will maximize her expected accuracy after learning that $\mathbf{X} = \mathbf{x}$. In other words, $\pi^*(\theta) = \arg\max_W \mathrm{E}_\pi[s(\pi^*)]$ where s is the scoring rule. In other words,

$$\begin{aligned} \pi^*(\theta) &= \arg\max_{\pi' \in W} \int_\Omega \pi(\theta)s(\pi'(\theta))d\theta \\ &= \arg\max_{\pi' \in W} -H(\pi, \pi') \end{aligned} \tag{4.23}$$

This quantity is therefore the negative of generalized cross-entropy. So to maximize posterior expected accuracy, we should minimize cross-entropy. That is, from an accuracy-centered point of view, the safest posterior is the posterior R which satisfies $\arg\min_{P \in W} H(P, \cdot)$. Therefore, the risk-free posterior is the minimum cross-entropy posterior. As before, however, it is not risk-free *because* it minimizes cross-entropy. Rather, it is risk-free because the agent stands to gain most by way of accuracy, in expectation, by adopting cross-entropy minimization as her update strategy.

So what is the minimum cross-entropy posterior? It has been shown that under reasonably general conditions – namely, where W is a closed and convex set – then for all strictly proper s, if the evidence will be observed with certainty and we stand to learn which element of a partition is true, then $\pi^*(\theta) = \pi(\theta|\mathbf{x}) \propto \pi(\theta)f(\mathbf{x}|\theta)$.[4] In other words, updating by Bayesian conditioning maximizes one's present expected accuracy of their posterior credences. In the early literature, the accuracy dimension of cross-entropy was not empha-

[4] [Shore and Johnson(1980), Williams(1980)]. For more recent philosophical literature, see [Oddie(1997), Greaves and Wallace(2006), Leitgeb and Pettigrew(2010b)].

sized. [Shore and Johnson(1980)], for example, aim to give an axiomatic development of a general information-theoretic method of inductive inference. This is in part because cross-entropy is uniformly assumed to be cross Shannon entropy. I do not make this assumption. Just as we can speak very generally about entropy we can speak just as generally about cross-entropy – as the expected accuracy of one distribution evaluated under another.

[Oddie(1997)], [Greaves and Wallace(2006)], [Leitgeb and Pettigrew(2010b)], and [Myrvold(2012)] develop the consequences of cross-entropy minimization explicitly in accuracy-theoretic terms. Under reasonably general conditions on the scoring rule and the type of learning experience, Bayesian conditioning indeed minimizes cross-entropy. For this reason, from an accuracy maximizing perspective, the Bayesian posterior is least risky.

4.3.4 An argument from the value of knowledge

[Good(1967)] shows that in an ordinary context, where the agent is deciding whether to act now, or perform a cost-free experiment and act later, it always pays in expectation to perform the experiment, provided she updates by Bayesian conditioning. That is, the expected utility of acting now is less than or equal to our current expectation of the utility of acting after performing the experiment.

It is straightforward to generalize this idea to the context of epistemic utility theory. Since our agent is an expected accuracy maximizer she should identify the distribution $\pi(\theta)$ in a way that maximizes expected accuracy, given by,

$$\max_{\theta \in \Omega} \int_{\Omega} s(\pi(\theta))\pi(\theta)d\theta \tag{4.24}$$

We already know from [Williams(1980)], among others, that Bayesian conditioning maximizes expected epistemic accuracy provided the constraint set is closed and convex and the evidence is learned with probability 1. It should not be surprising, therefore, that it pays in expectation to conduct an experiment and then produce one's forecast, provided the cost of the experiment is small enough. In this case, the cost will be given in terms of accuracy.

Suppose we are in a binary context again and our agent is given the chance to perform an experiment and observe $X = x$, after which she will reevaluate her assignment of $\hat{\theta}$. Since the case is binary, $s(\theta)$ is the agent's score. Is it reasonable to perform the experiment in expectation? To answer this question let $m(x) = \int \pi(\theta)f(x)d\theta = \int f(x,\theta)d\theta$ be the marginal distribution of X. If she identified $\hat{\theta}$ without observing $X = x$ then her present

expected epistemic utility can be re-written as follows,

$$\begin{aligned}
\hat{\theta}_{old} &= \max_{\theta\in\Omega} \mathrm{E}[s(\theta)|\pi(\theta)] \\
&= \max_{\theta\in\Omega} \int_{\Omega} s(\theta)\pi(\theta)d\theta \\
&= \max_{\theta\in\Omega} \int_{\Omega} s(\theta) \int_{\chi} \pi(\theta|x)m(x)dxd\theta \\
&= \max_{\theta\in\Omega} \int_{\Omega}\int_{\chi} s(\theta)\pi(\theta|x)m(x)dxd\theta \\
&= \max_{\theta\in\Omega} \int_{\Omega}\int_{\chi} s(\theta)\frac{\pi(\theta)f(x)}{m(x)}m(x)dxd\theta \\
&= \max_{\theta\in\Omega} \int_{\Omega}\int_{\chi} s(\theta)\pi(\theta)f(x)dxd\theta \\
&= \max_{\theta\in\Omega} \int_{\Omega}\int_{\chi} s(\theta)f(x,\theta)dxd\theta \\
&= \max_{\theta\in\Omega} \int_{\chi}\int_{\Omega} s(\theta)f(x,\theta)d\theta dx
\end{aligned} \tag{4.25}$$

Meanwhile, the post-experiment value of assigning $\hat{\theta}$ after observing $X = x$ is given as follows. Let θ^* be the estimate after the learning experience.

$$\begin{aligned}
\hat{\theta}_{new} &= \max_{\theta\in\Omega} \mathrm{E}[s(\theta^*)|\pi(\theta|x)] \\
&= \max_{\theta\in\Omega} \int_{\Omega} s(\theta^*)\pi(\theta|x)d\theta
\end{aligned} \tag{4.26}$$

Therefore, the expected value now of $\hat{\theta}_{new}$, i.e., $\hat{\theta}_{old}$ conditional on learning that $X = x$ becomes,

$$\begin{aligned}
\mathrm{E}[\hat{\theta}_{\mathrm{new}}|\pi(\theta)] &= \max_{\theta\in\Omega} \mathrm{E}[s(\theta^*)|\pi(\theta] \\
&= \max_{\theta\in\Omega} \int_{\Omega} s(\theta^*)\pi(\theta)d\theta \\
&= \int_{\chi} m(x)dx \max_{\theta\in\Omega} \int_{\Omega} s(\theta)\pi(\theta|x)d\theta \\
&= \int_{\chi} \max_{\theta\in\Omega} \int_{\Omega} s(\theta)\pi(\theta|x)m(x)d\theta dx \\
&= \int_{\chi} \max_{\theta\in\Omega} \int_{\Omega} s(\theta)f(x,\theta)d\theta dx
\end{aligned} \tag{4.27}$$

By Jensen's Inequality, $\int \max_t g(t)dt \geq \max_t \int g(t)dt$. Therefore, $\mathrm{E}[\hat{\theta}_{\mathrm{new}}|\pi(\theta)] \geq \mathrm{E}[\hat{\theta}_{\mathrm{old}}|\pi(\theta)]$.

This means that it is always better in expectation to observe that $X = x$ provided that after making this observation one will update by Bayesian conditioning. As is clear above, the proof relies on the fact that $\pi(\theta|x) = [\pi(\theta)f(x)]/m(x)$ (Bayes' Rule).

How much accuracy is it worth sacrificing to perform the experiment? Since the agent is updating by Bayes' Rule, letting $P = \pi(\theta)$ and $Q = \pi(\theta|x)$, the divergence between the prior and the posterior will be given by $D_{KL}(P||Q) = H(P,Q) - H(P)$. Therefore, it would be worth performing the experiment provided the cost does not exceed $\delta = H(P,Q) - H(P)$.

The argument from expected accuracy and the argument from the value of knowledge are related. Making an observation is reasonable from an epistemic perspective if one is going to condition on the evidence observed, since the update rule that maximizes the expected accuracy of one's posterior credences is the update rule given by Bayes' Theorem. Therefore, her present expected accuracy is less than or equal her expectation of her posterior accuracy. Using the language of entropy, the result comes to just this: the entropy of the prior is less than or equal to the cross-entropy between the prior and the posterior. This is true for all strictly proper scoring rules. Since their risk function is convex they all induce a Bregman divergence between a prior and a posterior, which as we saw can be decomposed into a difference of cross-entropy and entropy.

4.3.5 A prudential argument

It is well-known that an expected utility maximizer whose beliefs are probabilistically incoherent is vulnerable to accepting a set of bets which jointly leave her with a sure-loss. This is the classic dutch-book argument, originating in [Ramsey(1926)]:

> Any definite set of degrees of belief which broke [the laws of probability] would be inconsistent in the sense that it violated the laws of preference between options . . . If anyone's mental condition violated these laws . . . He could have a book made against him by a cunning bettor and would stand to lose in any event. (p. 41)

[Teller(1973)] introduces an argument scheme, which he attributes to David Lewis, showing that an agent who fails to update by Bayes' Rule leaves herself vulnerable, at least in principle, to accepting a sequence of bets which lead to a sure-loss. This is the well-known diachronic dutch book argument for Bayesian conditioning. In what follows, I describe the diachronic dutch-book argument through a simple example then quickly sketch a more formal version of this argument, owing to its presentation in [Skyrms(1987), Skyrms(1993)], but using the parametric notation I have been using throughout.

The diachronic dutch-book proceeds in terms of conditional and called off bets. For example, suppose I offer you the following series of bets regarding a baseball game.

> *Bet 1*: If the Tigers game is played tonight and the Tigers win you get 1 dollar. If the game is played and the Tigers lose, you get nothing.

The price I offer you for Bet 1 is equal to your prior probability that the game is played and the Tigers win.

> *Bet 2*: If the game is not played, you receive a dollar amount equal to your prior probability that the tigers win conditional on the game being played. Otherwise you get nothing.

The price I offer you for this bet is equal to your prior probability that the Tigers win, conditional on the game being played, multiplied by your prior probability that the game is not played. Next,

> *Bet 3*: If the game is played, you receive a dollar amount equal to your conditional probability the Tigers win given that the game is played, minus your (non-Bayesian) posterior probability that the Tigers will win after learning that the game will be played. Otherwise you receive nothing.

Now, if the game is not played the value you end up with is proportional to the difference between your conditional probability that the Tigers win given that the game is played and your posterior probability that the Tigers win. If your actual posterior probability is less than the Bayesian posterior probability you end up with a net loss. If, however, the game is played, then I offer you the following final bet:

> *Bet 4*: If the Tigers win you receive 1 dollar. Otherwise you receive 0.

The price of this bet is equal to your posterior probability that the game is played. Bets 1 and 2 jointly constitute a conditional bet on the Tigers winning, which is called off if the game is not played. But even if the game is played you again suffer a net loss proportional to the discrepancy between the Bayesian posterior and the posterior you are using. A similar series of bets can be presented to an agent whose actual posterior probability is greater than the Bayesian posterior probability that the Tigers win. Therefore, if an agent fails to update by Bayesian conditioning, she leaves herself vulnerable to sure-loss.

More formally, the diachronic dutch-book argument proceeds as follows. Let $P = \pi(\theta)$, $Q = \pi(\theta|x)$, and $R = \pi^*(\theta)$ where $Q, R \in W$. Suppose that $\pi(\theta|x) > \pi^*(\theta)$ for $\theta \in V \cap \Omega$ so that $D_{KL}(P||Q) > D_{KL}(P||R)$. We will suppose that $\pi(\theta|x) = \pi^*(\theta)$ for

$\theta \in V^c \cap \Omega$. That is, there is a subset of the parameter space where the alternative update rule and Bayes' Rule do not agree; in the remainder of the space they do agree. This is not supposed to be a controversial assumption – it's certainly possible that $V^c \cap \Omega = \emptyset$ and if $V \cap \Omega = \emptyset$ then $\pi^*(\theta) = \pi(\theta|x)$. Let $\delta = \pi(\theta|x) - \pi^*(\theta)$. Then the bookie can offer our agent the following sequence of bets.

Before observing that $X = x$, she will offer,

1. \$1 if $\theta = k$ and $X = x$, 0 otherwise;
2. $\$\pi(\theta = k|X = x)$ if $X \neq x$, 0 otherwise;
3. $\$\delta$ if $X = x$, 0 otherwise.

After making the observation, she will offer,

4. if $X = x$, [\$1 if $\theta = k$, \$0 otherwise] for its current fair price of $\pi^*(\theta) = \pi(\theta|x) - \delta$.

Under this sequence of bets the agent will lose δ regardless of what happens. A similar argument demonstrates vulnerability to a sure loss if $\pi(\theta|x) < \pi^*(\theta)$. From a pragmatic perspective, therefore, if the agent seeks to avoid vulnerability to sure-loss, the safest way to update is by Bayes' Rule – i.e., by adopting the posterior Q which satisfies $\min_{Q \in W} D_{KL}(P||Q)$.

4.3.6 An information-theoretic argument

The preceding arguments seek to persuade the reader that the Bayesian posterior is the least risky posterior on grounds that are conceptually independent from dynamic risk's information-theoretic expression. Of course, many authors have defended Bayesian updating explicitly on information-theoretic grounds.

Since entropy is a measure of expected uncertainty, evaluated with respect to itself, its negative is often called expected self-information. It seems natural, therefore, to extend this concept and evaluate the expected informativeness of a posterior distribution, from the perspective of the prior distribution. For example, this might be the expected informativeness of a predictive distribution from the perspective of the data generating distribution in a machine learning context. In the updating context, this lends itself to a reasonable rule for processing information: we should move to the posterior that is consistent with the evidence that has been observed, but that introduces as little additional information as possible [Jaynes(1957a), Jaynes(1957b), Jaynes(1963)]. In other words, if $H(P, Q) - H(P) = \delta$ we want an update rule from the constraint set W that minimizes δ. Naturally this amounts to

minimizing the associated Bregman divergence. Minimizing a Bregman divergence from Q to P with respect to Q is equivalent to minimizing the cross-entropy of P and Q, since as we saw above $H(P,Q) = H(P) + D_S(P||Q)$. Provided the constraint set is closed, convex and involves learning which element of a partition is true with probability 1, this is the Bayesian posterior. It is the posterior that adds as little information as possible while still being consistent with what has been learned. It is in this sense that we may think of it as the most conservative update rule.

This mirrors nicely our earlier result, where we found that epistemic risk is given by entropic change. So this is the key point: if we think about dynamic epistemic risk in terms of actual KL divergence minus minimum KL divergence, then we get an account of epistemic risk in terms of cross-entropic change. The least risky distribution under a strictly proper score is the maximum general entropy distribution. Meanwhile, the least risky posterior under a Bregman divergence induced by a strictly proper score is the minimum cross-entropy posterior.

Notice, however, that the divergence is given in terms of the change in divergence 'from Q to P' and 'from R to P'. We are measuring 'distance' from the perspective of the posterior, and dynamic epistemic risk is given by the change from the 'closest' posterior and the posterior whose dynamic risk we are interested in. We could have done it the other way around, measuring divergence from the prior to the least risky posterior and the target posterior. [Caticha and Giffin(2006)] suggest this alternative. If we did it this way, then the dynamic risk of Q would be equal to $H(R) - H(Q) + H(Q,P) - H(R,P)$. In other words it is the sum of the static risk increase between Q and R (entropic change) and their cross-entropic change. Or, we could take the sum of risk in both directions, in which case we would get $[H(P,Q) - H(P,R)] + [H(Q,P) - H(R,P)] + [H(R) - H(Q)]$. That is, it is the sum of changes in cross entropy in both directions and the difference in posterior entropy. I do not presently have a way of adjudicating between these competing measures. It is worth investigating further whether we should privilege one over the other.

4.4 Dynamic risk and the generalized beta distribution

Let us now return to our agents A, B, and C with beta-binomial distributions. We have established that the risk-free posterior is the Bayesian posterior. In order to calculate dynamic risk we also need to specify a scoring rule. I have argued that the score should be strictly proper without privileging any particular form. To illustrate dynamic risk with some examples, however, we will have to pick one from the strictly proper class. So for convenience let us assume that the applicable scoring rule is the additive log score which induces the

KL divergence.

For a random variable X following a Beta distribution f in interval I, Shannon entropy is given by,

$$\begin{aligned} H(X) &= \int_I f(x) \log f(x) dx \\ &= -\int_I \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} \log\left[\frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}\right] dx \\ &= \log(B(\alpha,\beta)) - (\alpha-1)\frac{\Gamma'(\alpha)}{\Gamma(\alpha)} - (\beta-1)\frac{\Gamma'(\beta)}{\Gamma(\beta)} + (\alpha+\beta-2)\frac{\Gamma'(\alpha+\beta)}{\Gamma(\alpha+\beta)} \end{aligned} \tag{4.28}$$

where $\Gamma(n)$ is the Gamma function given by $(n-1)!$ for $n \in \mathbb{N}^+$. Meanwhile, given two random variables $X \sim f(x) = \text{Beta}(\alpha, \beta)$ and $Y \sim g(x) = \text{Beta}(\alpha', \beta')$ the cross-entropy is given by,

$$\begin{aligned} H(X,Y) &= -\int_I f(x) \log g(x) dx \\ &= -\int_I \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} \log\left[\frac{1}{B(\alpha',\beta')} x^{\alpha'-1}(1-x)^{\beta'-1}\right] dx \\ &= \log(B(\alpha',\beta')) - (\alpha'-1)\frac{\Gamma'(\alpha)}{\Gamma(\alpha)} - (\beta'-1)\frac{\Gamma'(\beta)}{\Gamma(\beta)} + (\alpha'+\beta'-2)\frac{\Gamma'(\alpha+\beta)}{\Gamma(\alpha+\beta)} \end{aligned} \tag{4.29}$$

Let $Z \sim r(x) = \text{Beta}(\alpha'', \beta'')$ be the distribution from which KL divergence to $f(x)$ is minimized. Then the risk of $Y \sim g$ is $DR(Y) = H(X,Y) - H(X,Z)$. This is,

$$\begin{aligned} DR(Y) = &\log(B(\alpha',\beta')) - (\alpha'-1)\frac{\Gamma'(\alpha)}{\Gamma(\alpha)} - (\beta'-1)\frac{\Gamma'(\beta)}{\Gamma(\beta)} \\ &+ (\alpha'+\beta'-2)\frac{\Gamma'(\alpha+\beta)}{\Gamma(\alpha+\beta)} \\ &- \log(B(\alpha'',\beta'')) - (\alpha''-1)\frac{\Gamma'(\alpha')}{\Gamma(\alpha')} - (\beta''-1)\frac{\Gamma'(\beta')}{\Gamma(\beta')} \\ &+ (\alpha''+\beta''-2)\frac{\Gamma'(\alpha'+\beta')}{\Gamma(\alpha'+\beta')} \end{aligned} \tag{4.30}$$

For example, suppose each of our agents starts with a Beta$(1,1)$ (uniform) prior. I will refer to this prior as U. They observe six tosses: four tails and two heads. A moves to a Beta$(5,3)$ posterior, B moves to a Beta$(3,2)$ posterior, and C moves to a Beta$(9,5)$ posterior. Their posterior distributions are depicted in Figure (4.2), below.

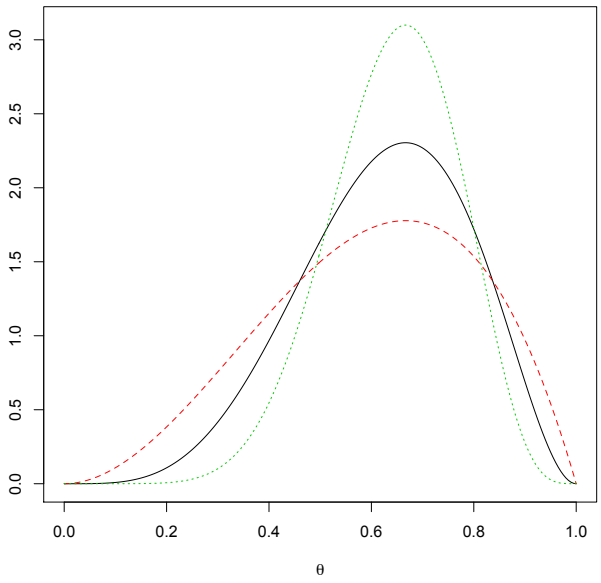

Figure 4.2: Posterior beta distributions

In Figure (4.2), the posterior of A is depicted in black, B is dashed red, and C is dotted green.

As expected, the least resilient distribution belongs to agent B, the agent who is adding antidata to the empirical sample. Next is the Goldilocks Bayesian agent A, who adds neither antidata nor super data to her sample. And finally we have C, the agent who is adding super data. Notice that just from eyeballing the means, it appears that the extent to which C deviates from A is greater than the extent to which B deviates from A. C's updating behavior is riskier than A's. We will explore this insight further, below.

Let $\mathbf{D}$ be the 4×4 D_{KL} divergence matrix. We know that $x_{ii} = 0$ since $D_{KL}(P||P) = 0$ and $x_{ij} > 0$ for $i \neq j$. Numerically, we find that,

$$\mathbf{D} = \begin{bmatrix} & U & A & B & C \\ U & 0.00 & 1.25 & 0.56 & 2.34 \\ A & 0.44 & 0.00 & 0.06 & 0.11 \\ B & 0.24 & 0.08 & 0.00 & 0.45 \\ C & 0.69 & 0.07 & 0.22 & 0.00 \end{bmatrix} \tag{4.31}$$

The way to read this matrix is as follows: $D_{KL}(i||j) = x_{ij}$. So, for example, $D_{KL}(U||A) =$

1.25 and $D_{KL}(A||U) = .44$.[5] Now we can compute their risks as follows,

$$\begin{aligned} DR(A) &= D_{KL}(U||A) - D_{KL}(U||A) = 0 \\ DR(B) &= D_{KL}(U||A) - D_{KL}(U||B) = .69 \\ DR(A) &= D_{KL}(U||C) - D_{KL}(U||A) = 1.09 \end{aligned} \tag{4.32}$$

As suggested by Figure (4.2) the risk of B is lower than the risk of C. This is a product of the asymmetry of the KL divergence.

Arguably, shooting past the evidence is riskier than failing to give the evidence enough weight. The latter seems like skepticism bordering on dogmatism, which is not ideal, but the former amounts to making up evidence, which does seem worse. To put this in our previous terminology: adding antidata to a sample is not as bad as adding the same amount of ersatz data to it.

On the other hand, the antidata/super data distinction does suggest a sort of parallel between the riskiness of the super updater's behavior and the riskiness of the sub updater's behavior. Why should it be a priori obvious that adding n observations is worse than adding $-n$ observations? Is not there a parity between these two ways of increasing risk? Perhaps not. There is an additional substantive reason for thinking that adding super data is worse than adding antidata: the super updater will be slower to respond to additional subsequent evidence because her new prior will be more resilient. This seems like an additional risk of error that the sub updater does not have to take – namely, making oneself less open to changing beliefs at a subsequent time. In other words, the super-updater leaves herself open to future dogmatism in a way that the sub-updater does not, which may increase the probability of error after subsequent learning experiences.

It would be interesting to explore this further – i.e., whether diluting a sample is as bad as overloading it – because the position we take here will reflect our attitude to different types of scoring rules. If we wish to maintain that antidata is indeed as risky as ersatz data then we need a symmetric Bregman divergence like squared Euclidean distance which corresponds to the Brier score. If we reject the notion that there is a parity in riskiness between the sub and super updater, and if we argue that the super updater is taking a bigger risk, as I am inclined to do, then this suggests an asymmetric score like the additive log score and the associated Shannon measure of entropy (and KL divergence).

It is also worth noting that with the KL divergence risk increases at an increasing rate. Suppose we had a risk increasing transformation which adds more ersatz data than C, call it C^*. So whereas C is speedy Bayes, C^* is super speedy Bayes. For example, if C^* is such

[5]The divergences are approximated by generating sequences of Beta distributed values for θ.

that $\eta = 3Y$ and $\sigma = 3(1-Y)$, then C^* ends up with a Beta$(13, 7)$ posterior (Figure (4.3)) , its risk is $DR(C^*) = 3.08$. While it is true that $DR(C^*) > DR(C) > DR(A)$, as we would like, the risk of super-speedy is almost three times as the risk of speedy even though the pace of updating is only increased by 50%.

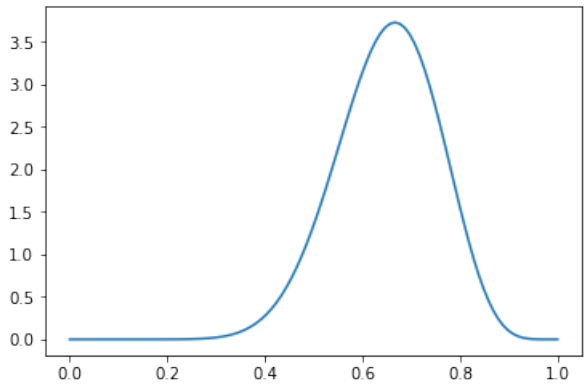

Figure 4.3: Posterior distribution of C^*

4.5 Multinomial dynamic epistemic risk

I developed the generalized Beta distribution as a way to model *both* the resilience *and* the update rule that underwrite a Bayesian agent's posterior credences. The two hyperparameters, α and β, tune the resilience of the prior by tracking pseudo tosses and to this we add η and σ, ersatz parameters for tuning the speed of updating. This was a case where the possible number of outcomes $K = 2$. But suppose $K \geq 3$, as in a die or a classification problem with many possible discrete outcomes. For example, an algorithm must classify an animal with known height and weight in to one of 100 numerically identified bins corresponding to different species. To model our Bayesian agent in the more general case, we can generalize the multinomial analogue to the Beta prior – namely, the Dirichlet prior.

Consider, first, the kernel of the Beta density, given by $\theta^{\alpha-1}(1-\theta)^{\beta-1}$. Clearly we can generalize this notion for $K \geq 3$ as $\prod_{i=1}^{K} \theta_i^{\alpha_i - 1}$ for $\alpha_1, ..., \alpha_K$. Now consider the normalizing constant $1/Beta(\alpha, \beta)$ where $Beta(\alpha, \beta) = [\Gamma(\alpha)\Gamma(\beta)]/\Gamma(\alpha + \beta)$. We can likewise generalize this as the product of Gammas divided by the Gamma of the sum by letting $Beta(\alpha) = \prod_{i=1}^{K} \Gamma(\alpha_i)/\Gamma(\sum_{i=1}^{K} \alpha_i)$. Putting this together we have a valid density for assigning prior probabilities for each θ_i outcome, given by,

$$\pi(\theta|\alpha) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{K} \alpha_i)} \prod_{i=1}^{K} \theta_i^{\alpha_i - 1} \tag{4.33}$$

where θ is a vector of probabilities for each outcome. We say that θ follows a Dirichlet(K, α) distribution where α_i is the hyperparameter for the i-th outcome. For example, for a fair three-sided die, $\theta = 1/3, 1/3, 1/3$. If $\alpha_1 = 2$, $\alpha_2 = 2$, $\alpha_3 = 2$ then our prior is a three-dimensional density concentrating around $1/3$, which may be represented in the ordinary 2-simplex, as follows.

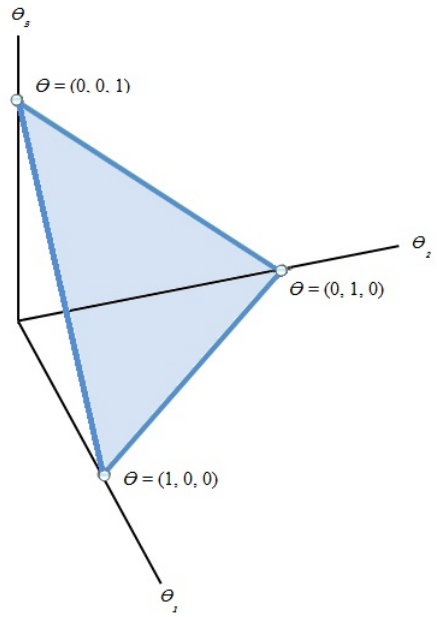

Figure 4.4: Dirichlet prior space for three-sided die

What about the empirical distribution of the outcomes? Recall that in a coin case, this was given by the binomial $f(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$.

Since $x! = \Gamma(x+1)$, the binomial coefficient can be expressed in terms of the Gamma function as $\Gamma(n+1)/[\Gamma(x+1)\Gamma(n-x+1)]$. We can generalize this to $x_1, ...x_k$ outcomes as $\Gamma(\sum_{i=1}^k x_i + 1)/\prod_{i=1}^k \Gamma(x_i+1)$. Even more intuitively, we can generalize the kernel of the distribution as $\prod_{i=1}^K \theta_i^{x_i}$. This gives us the multinomial empirical distribution for $K \geq 2$ outcomes, written as,

$$f(\mathbf{x}|\theta) = \frac{\Gamma(\sum_{i=1}^k x_i + 1)}{\prod_{i=1}^k \Gamma(x_i+1)} \prod_{i=1}^K \theta_i^{x_i} \tag{4.34}$$

Multiplying these together gives us the posterior,

$$\begin{aligned}\pi(\theta|\mathbf{x}) &= \left[\frac{\prod_{i=1}^{K}\Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{K}\alpha_i)}\prod_{i=1}^{K}\theta_i^{\alpha_i-1}\right]\left[\frac{\Gamma(\sum_{i=1}^{k}x_i+1)}{\prod_{i=1}^{k}\Gamma(x_i+1)}\prod_{i=1}^{K}\theta_i^{x_i}\right] \\ &= \frac{\prod_{i=1}^{K}\Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{K}\alpha_i)}\frac{\Gamma(\sum_{i=1}^{k}x_i+1)}{\prod_{i=1}^{k}\Gamma(x_i+1)}\prod_{i=1}^{K}\theta_i^{\alpha_i-1}\prod_{i=1}^{K}\theta_i^{x_i} \\ &= \frac{\prod_{i=1}^{K}\Gamma(\alpha_i+x_i+1)}{\Gamma(\sum_{i=1}^{K}\alpha_i+x_i)}\prod_{i=1}^{K}\theta_i^{\alpha_i+x_i-1}\end{aligned} \tag{4.35}$$

Then the posterior above is Dirichlet$(K, \alpha + \mathbf{x})$. Now we can apply the same approach as before to develop a multinomial ersatz parameter to vary the speed of the update rule. Let $\eta = (\eta_1, ..., \eta_k)$ be a vector of ersatz parameters, corresponding to each outcome. For example, if our super updater observes x_2, she also increases the value of η_2 by one unit. So the posterior distribution will be Dirichlet$(W + K, \alpha + \mathbf{x} + \eta)$. Again where $W = 0$ this reduces to the ordinary Dirichlet posterior. Like the generalized Beta distribution the generalized Dirichlet distribution is a valid pdf and exhibits the same asymptotic behavior.

Consider our agents, A, B, and C again. A is an ordinary updater, B is a super updater that adds one ersatz success η_i for every real success x_i (one unit of super data), and C is a sub updater that subtracts half a success for every real success (one half unit of antidata). Suppose we have a three-sided die whose bias is unknown and our agents start with a uniform prior, given by $Dir(1, 1, 1)$ Figure(4.5). Suppose our agents observe one toss of the die, and it lands on 3.

Our sub updater will move to $Dir(1, 1, 3/2)$ Figure(4.6), the ordinary updater will move to $Dir(1, 1, 2)$ Figure(4.7), and the super updater will move to $Dir(1, 1, 3)$ Figure(4.8).[6]

[6]The Python script for visualizing Dirichlet distributions in the 2-simplex is based on a script by Thomas Boggs, available at http://blog.bogatron.net/blog/2014/02/02/visualizing-dirichlet-distributions/.

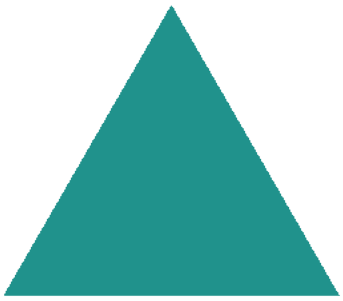

Figure 4.5: Dir$(1, 1, 1)$ posterior

Figure 4.6: Dir$(1, 1, 3/2)$ posterior

Figure 4.7: Dir$(1, 1, 2)$ posterior

Figure 4.8: Dir$(1, 1, 3)$ posterior

The Bayes estimator for the Dirichlet distribution has the same properties as the Bayes estimator for Beta distributions. The bias now is given by the expected value of a toss of the die. For a fair three-sided die, this is $6/3 = 2$. To estimate unknown bias we use the same logic as before. For example, super we start with a $Dir(1, 1, 1)$ prior, observe one 3, and update by ordinary Bayesian conditioning to a $Dir(1, 1, 2)$ posterior. Then our Bayesian estimate of the die's bias is $\hat{\mu} = 9/4$ – the die is estimated to be slightly biased in favor of 3. Notice that since the prior here is so weak even the ordinary updater's estimate is quite responsive to the evidence. For our super updater, her estimate of the bias here would have been $12/5 = 2.4$ and for our sub updater it would have been 2.14. In general,

$$\hat{\mu} = \mathrm{E}[\theta|\mathbf{x}] = \frac{\sum_{i=1}^{K} c_i x_i + \alpha_i + \eta_i}{\sum_{i=1}^{K} \alpha_i + W + n} \tag{4.36}$$

where $(c_1, ..., c_k)$ is the number of successes of $(x_1, ..., x_k)$. This may again be written as a linear combination of the pseudo expectation and MLE/ersatz hybrid expectation.

Now compare for illustration a case where we start with a much stronger prior. For example, suppose we start with a $Dir(10, 10, 10)$ prior. In this case, the valence of the prior Bayes estimate is the same as it would have been with the uniform $Dir(1, 1, 1)$ – $\hat{\mu} = 2$ – but the prior is now much stickier – as depicted by the concentration of density toward the center of the simplex in Figure (4.9). Suppose we then observe ten tosses, as follows:

1 2 1 3 2 3 1 3 2 3

(three 1's, three 2's and four 3's). As we said, the prior estimate of the bias is 2 (the die was assumed to be fair) and the maximum likelihood estimate after observing this sequence is 2.2. We know that our agents' estimates will be between these values – we expect to find that $2 < \hat{\mu}_B < \hat{\mu}_A < \hat{\mu}_C < 2.2$. The ordinary Bayesian updater will move to a $Dir(13, 13, 14)$ posterior (Figure 4.11). Her estimate of the die's bias will be $\hat{\mu} = 2.025$ – i.e. slightly biased toward three. The sub updater will move to $Dir(11.5, 11.5, 12)$ (Figure 4.10) and her estimate of the bias will be $\hat{\mu} = 2.014$. As we expect, she is less responsive to the evidence. She moves away from her prior, but begrudgingly so. Meanwhile, the super updater will move to $Dir(16, 16, 18)$ (Figure 4.12). Her estimate of the bias becomes $\hat{\mu} = 2.04$. Her update rule is most sensitive to evidence and she is quickest in moving toward the MLE.

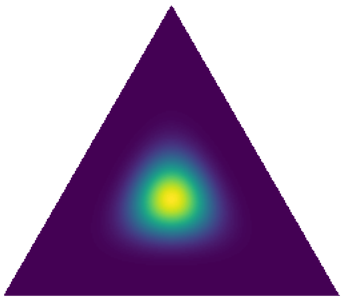

Figure 4.9: Dir$(10, 10, 10)$ posterior

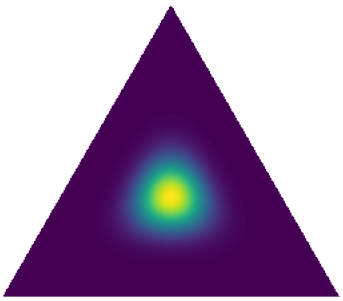

Figure 4.10: Dir$(11.5, 11.5, 12)$ posterior

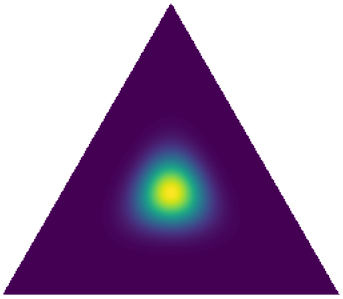

Figure 4.11: Dir$(13, 13, 14)$ posterior

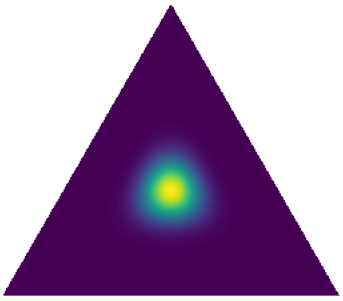

Figure 4.12: Dir$(16, 16, 18)$ posterior

It is obvious from the figure here that the changes from a Dir$(10, 10, 10)$ prior are much more subtle as compared to the updating behavior of our agents when they started with a Dir$(1, 1, 1)$ prior. In each case, the posterior estimate satisfies,

$$\hat{\mu}_{\pi(\theta)} < \hat{\mu}_B < \hat{\mu}_A < \hat{\mu}_C < \hat{\mu}_{MLE} \tag{4.37}$$

This is because of our update tuning parameter η. However, because the hyperparameters $(\alpha_1, \alpha_2, \alpha_3)$ are different in the two cases the distance between each agent's prior and posterior is greater in the Dir$(10, 10, 10)$ case than it is in the Dir$(1, 1, 1)$ case. For example, in the former case, agent A moves from 2 to 2.25 after only observing one toss. In the latter case, she moves from 2 to only 2.025 despite a ten-fold increase in the stock of her evidence. She moves from a centered estimate of the bias to one that just slightly favors 3's. This is not easy to see in the figure because the changes are so subtle, but consider for example a Dir$(10, 10, 50)$ posterior, in Figure (4.13), below.

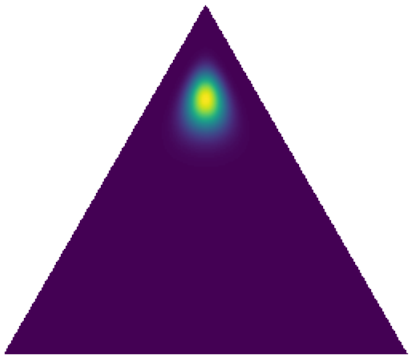

Figure 4.13: Dir$(10, 10, 50)$ density

Dir$(10, 10, 10)$ credences are much more resilient than Dir$(1, 1, 1)$ credences. Therefore, as in the simple beta case, α tunes resilience and η tunes the speed of the update. The only difference is that α and η are now vectors of length K.

The notion of dynamic epistemic risk is still given in terms of cross-entropic change though I will avoid expressing it as it is really no longer feasible to compute by hand. The mathematical machinery remains the same, however, as does the underlying conceptual interpretation.

4.6 Conclusion

This chapter builds on the preceding one by developing a measure of dynamic risk that is analogous to the general theory of epistemic risk motivated in terms of sensitivity to error and expressed in information-theoretic terms. Whereas static epistemic risk is given by entropic change, dynamic epistemic risk is given by cross-entropic change. This gives us a unified theory of epistemic risk for Bayesian inference – the static dimension for the assessment of priors and the dynamic dimension for the assessment of updating procedures.

CHAPTER 5

Adaptive Burdens of Proof

5.1 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)Tetlock, Kristel, Elson, Green, and Lerner] 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

[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)Colyvan, Regan, and Ferson] and [Schauer(2003)].

[3] Competing models are developed in [Posner(1999)], [Kaplow(2014)], [Cheng(2013)], and [Cheng and 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.

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 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 Chapter 2 – 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 chapter 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 and 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

[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 and Zamir(2012)].

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 and 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 and Schervish(2012)] 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 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.[7]

[7] *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.

5.2 Modeling burdens of proof

There is a substantial literature in economic and statistical analyses of evidence law on modeling burdens of proof.[8] Or, to be specific, the burden of persuasion.[9] 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.

5.2.1 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.[10] 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.[11] Other economic approaches are more general and consider costs that, strictly speaking, may not be reducible to their effects on individual utilities.[12] On the accuracy-first approach, correctness of verdicts is the overarching consideration.[13] 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.[14] As a result, as [Kaplow(2012)] puts it, the optimal threshold is de-

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

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

[10] See also [Kaplow(2012)].

[11] See [Kaplow and 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.

[12] See e.g., [Miceli(1990)] (considering the value of retribution and proportionality). Miceli builds proportionality into the utility function. [Kaplow and 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 and Shavell(2006)] is to argue against non-consequentialist approaches to the assessment of legal standards.

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

[14] That is, behavior changes in response to changes in the values of the parameters in the burden of proof

termined 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. 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 and 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 and 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.[15] 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 considera-

model. But see [Rubinfeld and 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.

[15]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 and Pardo(2015)], he criticizes [Kaplow(2012)] for proposing a rule that is difficult to apply in practice.

tions?"[16] 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.[17]

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

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

Table 5.1: 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

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

[17]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 and Lancaster(1956)]'s general theory of second best.

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?[18] 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).

5.2.2 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 com-

[18] 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.

pany 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 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).[19]

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.[20] 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,

$$\frac{L(\mathbf{X}|H)}{L(\mathbf{X}|\overline{H})} > k \tag{5.1}$$

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

[19]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.

[20]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 and 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.

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

$$\frac{P(H)}{P(\overline{H})} \frac{L(\mathbf{X}|H)}{L(\mathbf{X}|\overline{H})} > k \tag{5.2}$$

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,

$$\frac{P(H|\mathbf{X})}{P(\overline{H}|\mathbf{X})} = \frac{P(H)}{P(\overline{H})} \frac{L(\mathbf{X}|H)}{L(\mathbf{X}|\overline{H})} \tag{5.3}$$

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 (5.2) 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 (5.2) 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.

5.3 Naked statistics: the phantom menace[21]

5.3.1 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 (5.3) 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 [Posner(1999)] 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" (1508). Indeed, in *Branion v. Gramly*, 855 F.2d 1256, 1263-64 (7th Cir. 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

[21] The title is inspired by [Hershovitz(2002)].

[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.

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

[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)Colyvan, Regan, and Ferson], [Schauer(2003)], [Redmayne(2008)], [Kaplow(2012)], [Buchak(2014)], [Cheng(2013)] and [Cheng and 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).

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 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.

5.3.2 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]

[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)Enoch, Spectre, and Fisher] (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).

[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.

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 $posterior\ odds = prior\ odds \times likelihood\ ratio$ (5.3), we have,

$$\frac{P(G|g)}{P(B|g)} = \frac{P(G)}{P(B)} \times \frac{L(g|G)}{L(g|B)} = \frac{1}{4} \times \frac{.8}{.2} = 1 \tag{5.4}$$

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 and Tversky(1972)] and [Bar-Hillel(1980)], among others, report finding that the average value for $P(G|g)$ among their subjects is approximately $.8$, 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.

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

[33] For experimental evidence on legal hypotheticals in particular, see [Wells(1992)].

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

[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.

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 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

[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 and Vidmar(2001)] (describing cases where legal decision makers ordinarily incorporate prior information).

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).

5.4 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

[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.

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 and 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

[38]For example, [Thomson(1986)], [Buchak(2014)], and [Enoch et al.(2012)Enoch, Spectre, and Fisher] 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 and Murphy(1968)], [Savage(1971)], [Schervish(1989)], and [Lindley(1982)] for classic discussions and [Gneiting and 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)Merkle, Steyvers, Mellers, and Tetlock] (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.

is greater than the likelihood of the data under the defendant's hypothesis. In other words, find in favor of the plaintiff if,

$$\frac{L(\mathbf{X}|H)}{L(\mathbf{X}|\overline{H})} > 1 \tag{5.5}$$

This is effectively the proposal [Cheng(2013)] puts forth.[41] It is equivalent to (5.1) 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 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 (5.5) 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 and 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.

5.4.1 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 $\theta \neq 0$.

[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.

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 and 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.

5.4.2 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 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,

$$\min_{\delta \in \Delta} a\alpha(\delta) + b\beta(\delta) \tag{5.6}$$

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

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

[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.

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 and 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.

5.4.3 Epistemic risk and the adaptive model

[DeGroot and Schervish(2012)] show that if our goal is to minimize (5.6) 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,

$$b + \sum_{\mathbf{x} \in S_1} [af(\mathbf{x}|H) - bf(\mathbf{x}|\overline{H})] \tag{5.7}$$

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 (5.7) 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,

$$\frac{L(\mathbf{X}|H)}{L(\mathbf{X}|\overline{H})} \geq \frac{b}{a} \tag{5.8}$$

Notice that if we let $k = b/a$ then (5.8) is equal to (5.1). What we get from [DeGroot and 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 (5.1), 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,

$$\frac{P(H)}{P(\overline{H})} \frac{L(\mathbf{X}|H)}{L(\mathbf{X}|\overline{H})} \geq \frac{P(H)}{P(\overline{H})} \frac{b}{a} \tag{5.9}$$

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.

$$\frac{P(H)}{P(\overline{H})} \frac{L(\mathbf{X}|H)}{L(\mathbf{X}|\overline{H})} \geq \frac{b^*}{a^*} \tag{5.10}$$

The left hand side should be familiar now – by (5.3) 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]

[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

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

$$P(H|\mathbf{X})/P(\overline{H}|\mathbf{X}) > b/a \tag{5.11}$$

This is identical to our statement of the adaptive model in (5.2), 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 and Pardo(2015)]'s alternative accuracy-first model.

5.5 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 (5.11).

5.5.1 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,

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.

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

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 and 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 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" [**?**,]1268, n. 26]Cheng2013. 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" [**?**,]1267, n. 24]Cheng2013.

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 and 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

[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.").

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 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 and 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.

5.5.2 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]

5.5.3 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 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

[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.

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(2005a)]. 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).

[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).

5.5.4 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.

[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)Tetlock, Kristel, Elson, Green, and Lerner] 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. $10/11$ths) – 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 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

[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).

[52] NAT'L RESEARCH COUNCIL, COMMITTEE ON FORENSIC DATA TECHNOLOGY: AN UPDATE, THE EVALUATION OF FORENSIC DNA EVIDENCE 185 (1996). See also [Zabell(2005a)] 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.

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

[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.

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 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

[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).

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.

5.6 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 and Morgenstern(1944)].

5.6.1 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 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 and 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

[57] See e.g., [Tribe(1971)] and [Wasserman(1991)].

[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

approach fail to capture salient moral considerations. This disagreement is part of a broader debate about the moral foundations of legal institutions [Kaplow and 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 and 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 and 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.

5.6.2 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.

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.").

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

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 chapter 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).[59] 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 and Zamir(2012)]. In subsequent research, it would be worthwhile to put the model to a more direct empirical test.

5.6.3 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

[59] 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.

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.

5.7 Conclusion

In this chapter I 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.

APPENDIX A

Appendix of Proofs

Theorem 1. For strictly concave and twice differentiable entropy function H and risk function R defined on $[0, 1]$,

$$R(p) + H(p) = k \text{ where } k = \min_p R(p) = \max_p H(p)$$

Proof.

Recall that $h(p) = s_1(p) - s_0(p)$ and $P(p) = \int_p^{p^*} |h(t)|dt$ where $p^* \arg\max_{p\in[0,1]} H(p)$. As a result, $H(p^*) = k$, $H'(p^*) = 0$ and $H''(p^*) < 0$.

Existence of risk free point.
Since $s_1(p)$ is continuous and decreasing on $[0, 1]$ with $s_1(1) = 0$, and $s_0(p)$ is continuous and increasing on $[0, 1]$ with $s_0(0) = 0$, the intermediate value theorem guarantees that a risk free point p^* exists. Alternatively, since $H(p)$ is closed and bounded on $[0, 1]$, the extreme value theorem guarantees that p^* exists.

Duality of risk and entropy.
Recall [Savage(1971)] shows that we can express $s_v(p)$ in terms of $H(p)$ as follows,

$$s_1(p) = H(p) + (1 - p)H'(p) \qquad\qquad s_0(p) = H(p) - pH'(p)$$

As a result, we can expand $h(p)$ in terms of the entropy $H(p)$,

$$\begin{aligned} h(p) &= [H(p) + (1 - p)H'(p)] - [H(p) - pH'(p)] \\ &= (1 - p)H'(p) + pH'(p) \\ &= H'(p) \end{aligned}$$

Therefore,

$$\int_p^{p^*} h(t)dt = \int_p^{p^*} H'(t)dt = H(p^*) - H(p)$$

This implies that,

$$R(p) = \int_p^{p^*} |h(t)|dt$$

Which we can evaluate in parts.

For $s_1(p) > s_0(p)$,

$$\begin{aligned} R(p) &= \int_p^{p^*} h(t)dt \\ &= H(p^*) - H(p) \\ &= k - H(p) \end{aligned}$$

For $s_0(p) > s_1(p)$,

$$\begin{aligned} R(p) &= -\int_{p^*}^{p} h(t)dt \\ &= -[H(p) - H(p^*)] \\ &= k - H(p) \end{aligned}$$

For $s_0(p) = s_1(p)$,

$$\begin{aligned} R(p) &= \int_p^{p^*} h(t)dt \\ &= k - k = 0 \end{aligned}$$

□

Theorem 2. Given a random variable $X : S \to W$, where the underlying scoring rule s_v is proper, and two cdfs P and Q, if P is a mean preserving epistemic spread of Q then $R(P) > R(Q)$.

Proof.
Suppose P is a mean preserving epistemic spread of Q. Then $H(Q) > H(P)$. Let P^* be the risk-free probability so that $H(P^*) = R(P^*) = 0$. By Theorem (1), $H(P^*) - R(Q) > H(P^*) - R(P)$. Therefore, $R(P) > R(Q)$. □

Theorem 3. Given a random variable $X : S \to W$, where $W \subseteq \mathbb{R}$ contains inaccuracy scores measured by a scoring rule s_v, let $V = \{P_1, ..., P_n\}$ be a set of cdfs for X. Given a random variable $Y : S \to W^*$, where $W^* \subseteq \mathbb{R}$ contains inaccuracy scores measured by a different scoring rule s_v^*, let $U = \{Q_1, ..., Q_n\}$ be a set of corresponding cdfs for Y. This means that for each outcome $h \in S$, the probability assigned to h by P_i is equal to the probability assigned to h by Q_i, but whereas in the first case the outcome h is described by s_v in the second case it is described by s_v^*. Suppose (1) s_v and s_v^* are truth directed scoring rules, whose risk functions R and R^* are such that (2) $R'' > 0$, $R^{*''} > 0$, and (3) $\arg\min R = \arg\min R^*$ on the unit interval. Then $R(P_i) > R(P_j)$ if and only if $R(Q_i) > R(Q_j)$.

Proof.
Sufficiency: assume $R(P_i) > R(P_j)$ for arbitrary $i \neq j$. Recall that $R(P) = \mathrm{E}[P^*] - \mathrm{E}[P]$ where $P^* = \max_{P \in V} \mathrm{E}[P]$ is the risk-free cdf. Conditions (2) and (3), together with the extreme value theorem, imply that P^* exists. Condition (3) implies that $P^* = Q^*$. Finally, condition (1) implies that if $\mathrm{E}[P_i] > \mathrm{E}[P_j]$ then $\mathrm{E}[Q_i] > \mathrm{E}[Q_j]$. Therefore,

$$
\begin{aligned}
& R(P_i) > R(P_j) \\
& \leftrightarrow \mathrm{E}[P^*] - \mathrm{E}[P_i] > \mathrm{E}[P^*] - \mathrm{E}[P_j] \\
& \leftrightarrow \mathrm{E}[P^* - P_i] > \mathrm{E}[P^* - P_j] \\
& \leftrightarrow \mathrm{E}[Q^* - P_i] > \mathrm{E}[Q^* - P_j] \\
& \rightarrow \mathrm{E}[Q^* - Q_i] > \mathrm{E}[Q^* - Q_j] \\
& \leftrightarrow R(Q_i) > R(Q_j)
\end{aligned}
$$

Necessity: The procedure above is reversible. Everything we have said remains true if we swap Q's for P's and W for V. □

BIBLIOGRAPHY

[Arrow(1965)] Arrow, Kenneth J. 1965. *Aspectes of the Theory of Risk Bearing*. Helsinki: Yrjõ Jahnsson Sããtiõ.

[Arrow(1971)] —. 1971. *Essays in the Theory of Risk Bearing*. Chicago: Markham.

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

[Bernoulli(1954/1738)] Bernoulli, Daniel. 1954/1738. "Exposition of a New Theory on the Measurement of Risk." *Econometrica* 22:23–36.

[Bradley and Steele(2016)] Bradley, Seamus and Steele, Katie. 2016. "Can Free Evidence Be Bad? Value of Information for the Imprecise Probabilist." *Philosophy of Science* 83:1–28.

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

[Brook(1985)] 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* 29:293–352.

[Buchak(2010)] Buchak, Lara. 2010. "Instrumental Rationality, Epistemic Rationality, and Evidence Gathering." *Philosophical Perspectives* 24:85–120.

[Buchak(2014)] —. 2014. "Belief, Credence, and Norms." *Philosophical Studies* 169:285–311.

[Buja et al.(2005)Buja, Stuetzle, and Shen] Buja, Andreas, Stuetzle, Werner, and Shen, Yi. 2005. "Loss Functions for Binary Class Probability Estimation and Classification: Structure and Applications." Manuscript.

[Caticha and Giffin(2006)] Caticha, Ariel and Giffin, Adom. 2006. "Updating Probabilities." *AIP Conference Proceedings* 872:31–42.

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

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

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

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

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

[De Finetti(1974)] —. 1974. *Theory of Probability*, volume 1. New York: John Wiley and Sons.

[DeGroot and Schervish(2012)] DeGroot, Morris H. and Schervish, Mark J. 2012. *Probability and Statistics*. New York: Wiley, 4th edition.

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

[Diamond and Vidmar(2001)] Diamond, Shari Seidman and Vidmar, Neil. 2001. "Jury Room Ruminations on Forbidden Topics." *Virginia Law Review* 87:1857–1915.

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

[Fallis(2007)] Fallis, Don. 2007. "Attitudes Toward Epistemic Risk and the Value of Experiments." *Studia Logica* 86:215–246.

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

[Frankfurt(1958)] Frankfurt, Harry. 1958. "Peirce's Notion of Abduction." *Journal of Philosophy* 55:593–597.

[Gaifman and Vasudevan(2012)] Gaifman, Haim and Vasudevan, Anubav. 2012. "Deceptive Updating and Minimal Information Methods." *Synthese* 187:147–178.

[Gelman et. al.(2013)] Gelman et. al., Andrew. 2013. *Bayesian Data Analysis*. New York: CRC Press (Taylor & Francis), 3rd edition.

[Gibbard(2008)] Gibbard, Allan. 2008. "Rational Credence and the Value of Truth." In *Oxford Studies in Epistemology*, volume 2. Oxford: Oxford University Press.

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

[Goldman(2002)] Goldman, Alvin I. 2002. *Pathways to Knowledge: Private and Public*. Oxford: Oxford University Press.

[Good(1967)] Good, I.J. 1967. "On the Principle of Total Evidence." *The British Journal for the Philosophy of Science* 17:319–321.

[Greaves(2013)] Greaves, Hilary. 2013. "Epistemic Decision Theory." *Mind* 122:915–952.

[Greaves and Wallace(2006)] Greaves, Hilary and Wallace, David. 2006. "Justifying Conditionalization: Conditionalization Maximizes Expected Epistemic Utility." *Mind* 115:607–632.

[Grunwald(2000)] Grunwald, Peter. 2000. "Maximum entropy and the glasses you are looking through." In *Proceedings of the sixteenth conference on uncertainty in artificial intelligence*, 238–246. Morgan Kaufmann Publishers Inc.

[Grunwald and Dawid(2004)] Grunwald, Peter and Dawid, A.P. 2004. "Game Theory, Maximum Entropy, Minimum Discrepancy and Robust Bayesian Decision Theory." *The Annals of Statistics* 32:1367–1433.

[Hacking(1965)] Hacking, Ian. 1965. *Logic of Statistical Inference*. Cambridge: Cambridge University Press.

[Halmos and Savage(1949)] Halmos, Paul R. and Savage, Leonard J. 1949. "Application of the Radon-Nikodym Theorem to the Theory of Sufficient Statistics." *The Annals of Mathematical Statistics* 20:225–241.

[Harman(1965)] Harman, Gilbert. 1965. "The Inference to the Best Explanation." *Philosophical Review* 74:88–95.

[Harman(1986)] —. 1986. *Change in View*. Cambridge: MIT Press.

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

[Hershovitz(2002)] Hershovitz, Scott. 2002. "Wittgenstein on Rules: The Phantom Menace." *Oxford Journal of Legal Studies* 22:619–640.

[James(1896)] James, William. 1896. "The Will to Believe." *The New World* 5:327–347.

[Jaynes(1957a)] Jaynes, Edwin T. 1957a. "Information Theory and Statistical Mechanics. I." *Physical Review* 106:620–630.

[Jaynes(1957b)] —. 1957b. "Information Theory and Statistical Mechanics. II." *Physical Review* 108:171–190.

[Jaynes(1963)] —. 1963. "Brandeis Summer Institute Lectures in Theoretical Physics." In R. Rosenkrantz (ed.), *Papers on Probability, Statistics and Statistical Physics*. Dordrecht: Reidel (1983).

[Jaynes(2003)] —. 2003. *Probability Theory: The Logic of Science*. Cambridge: Cambridge University Pressniversity press.

[Johnson(1924)] Johnson, W.E. 1924. *Logic, Part III. The Logical Foundation of Science*. Cambridge: Cambridge University Press.

[Joyce(1998)] Joyce, James M. 1998. "A Nonpragmatic Vindication of Probabilism." *Philosophy of Science* 65:575–603.

[Joyce(2009)] —. 2009. "Accuracy and Coherence: Prospects for an Alethic Epistemology of Partial Belief." In Franz Huber and Christoph Shmidt-Petri (eds.), *Degrees of Belief*, 263–300. Springer.

[Kahneman and Tversky(1972)] Kahneman, Daniel and Tversky, Amos. 1972. "On Prediction and Judgment." Technical Report 12(4), Oregon Research Institute Bulletin.

[Kahneman and Tversky(1982)] —. 1982. *Judgment under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.

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

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

[Kaplow(2011)] —. 2011. "On the Optimal Burden of Proof." *Journal of Political Economy* 119:1104–1140.

[Kaplow(2012)] —. 2012. "Burden of Proof." *Yale Law Journal* 121:738–859.

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

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

[Kaplow and Shavell(2006)] —. 2006. *Fairness versus Welfare*. Cambridge: Harvard University Press.

[Kitcher(1990)] Kitcher, Philip. 1990. "The Division of Cognitive Labor." *The Journal of Philosophy* 87:5–22.

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

[Kyburg(1974)] Kyburg, Henry E. 1974. *The Logical Foundations of Statistical Interference*. Dordrecht: Reidel.

[Laudan and Allen(2008)] Laudan, Larry and Allen, Ronald J. 2008. "Deadly Dilemmas." *Texas Tech Law Review* 41:65–93.

[Lehmann and Romano(2005)] Lehmann, Erich L. and Romano, Joseph P. 2005. *Testing Statistical Hypotheses*. New York: Springer.

[Leitgeb and Pettigrew(2010a)] Leitgeb, Hannes and Pettigrew, Richard. 2010a. "An Objective Justification of Bayesianism I: Measuring Inaccuracy." *Philosophy of Science* 77:201–235.

[Leitgeb and Pettigrew(2010b)] —. 2010b. "An Objective Justification of Bayesianism II: The Consequences of Minimizing Inaccuracy." *Philosophy of Science* 77:236–272.

[Levi(1962)] Levi, Isaac. 1962. "On the Seriousness of Mistakes." *Philosophy of Science* 29:47–65.

[Levi(1974)] —. 1974. *Gambling with Truth*. Cambridge: MIT Press.

[Levi(1977)] —. 1977. "Epistemic Utility and the Evaluation of Experiments." *Philosophy of Science* 44:368–386.

[Lewis(1980)] Lewis, David. 1980. "A Subjectivist's Guide to Objective Chance." In Richard C. Jeffrey (ed.), *Studies in Inductive Logic and Probability*, volume 2, 263–293. Berkeley: University of California Press.

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

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

[Maher(1990)] Maher, Patrick. 1990. "Why Scientists Gather Evidence." *British Journal for the Philosophy of Science* 41:103–119.

[Maher(1993)] —. 1993. *Betting on Theories*. Cambridge: Cambridge University Press.

[Markovitz(1952)] Markovitz, H. 1952. "Portfolio Selection." *Journal of Finance* 7:77–91.

[Markovitz(1959)] —. 1959. *Portfolio Selection: Efficient Diversification of Investment*. New Haven: Yale University Press.

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

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

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

[Myrvold(2012)] Myrvold, Wayne C. 2012. "Epistemic Values and the Value of Learning." *Synthese* 187:547–568.

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

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

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

[Oddie(1997)] Oddie, Graham. 1997. "Conditionalization, Cogency, and Cognitive Value." *British Journal for the Philosophy of Science* 48:533–541.

[Peirce(1879)] Peirce, Charles Sanders. 1879. "Note on the Theory of the Economy of Research." Technical report, United States Coast Survey, US Government Publishing Office (Reprinted in *Operations Research 15(4)* (1967): 643-648).

[Peirce(1931-1958)] —. 1931-1958. *Collected Papers of Charles Sanders Peirce*, volume 1-8. Cambridge: Harvard University Press.

[Pettigrew(2012)] Pettigrew, Richard. 2012. "Accuracy, Change, and the Principal Principle." *Philosophical Review* 121:241–275.

[Pettigrew(2016a)] —. 2016a. *Accuracy and the Laws of Credence*. Oxford: Oxford University Press.

[Pettigrew(2016b)] —. 2016b. "Accuracy, Risk, and the Principle of Indifference." *Philosophy and Phenomenological Research* 92:35–59.

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

[Pratt(1964)] Pratt, John W. 1964. "Risk Aversion in the Small and in the Large." *Econometrica* 32:122–136.

[Pritchard(2007)] Pritchard, Duncan. 2007. "Anti-Luck Epistemology." *Synthese* 158:277–297.

[Pritchard(2017)] —. 2017. "Epistemic Risk." *The Journal of Philosophy* Forthcoming.

[Ramsey(1926)] 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(2008)] Redmayne, Mike. 2008. "Exploring the Proof Paradoxes." *Legal Theory* 14:281–309.

[Rescher(1976)] Rescher, Nicholas. 1976. "Peirce and the Economy of Research." *Philosophy of Science* 43:71–98.

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

[Rodriguez(2006)] Rodriguez, Carlos C. 2006. "Antidata." *AIP Conference Proceedings* 872:161–178.

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

[Rothschild and Stiglitz(1970)] Rothschild, Michael and Stiglitz, Joseph E. 1970. "Increasing Risk: I. A Definition." *Journal of Economic Theory* 2:225–243.

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

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

[Savage(1954)] Savage, Leonard J. 1954. *The Foundations of Statistics*. New York: Dover.

[Savage(1971)] —. 1971. "Elicitation of Personal Probabilities and Expectations." *Journal of the American Statistical Association* 66:pp. 783–801. ISSN 01621459.

[Schauer(2003)] Schauer, Frederick. 2003. *Profiles, Probabilities, and Stereotypes*. Cambridge: Harvard University Press.

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

[Seidenfeld(1986)] Seidenfeld, Teddy. 1986. "Entropy and Uncertainty." *Philosophy of Science* 53:467–491.

[Selten(1998)] Selten, Reinhard. 1998. "Axiomatic Characterization of the Quadratic Scoring Rule." *Experimental Economics* 1:43–61.

[Shannon(1948)] Shannon, Claude E. 1948. "A Mathematical Theory of Communication." Technical report, Bell Systems Technical Journal.

[Shore and Johnson(1980)] Shore, J.E. and Johnson, R.W. 1980. "Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy." *IEEE Transactions on Information Theory* 26:26–37.

[Skyrms(1987)] Skyrms, Brian. 1987. "Dynamic Coherence and Probability Kinematics." *Philosophy of Science* 54:1–20.

[Skyrms(1993)] —. 1993. "A Mistake in Dynamic Coherence Arguments?" *Philosophy of Science* 60:320–328.

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

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

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

[Stigler(1978)] Stigler, Stephen M. 1978. "Mathematical Statistics in the Early States." *The Annals of Statistics* 6:239–265.

[Teller(1973)] Teller, Paul. 1973. "Conditionalization and Observation." *Synthese* 26:218–258.

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

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

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

[Van Fraassen(1981)] Van Fraassen, Bas C. 1981. "A Problem for Relative Information Minimizers in Probability Kinematics." *British Journal for the Philosophy of Science* 32:375–379.

[Van Fraassen(1989)] —. 1989. *Laws and Symmetry*. Oxford: Clarendon Press.

[von Neumann and Morgenstern(1944)] von Neumann, John and Morgenstern, Oskar. 1944. *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.

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

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

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

[Wible(1994)] Wible, James R. 1994. "Charles Sanders Peirce's Economy of Research." *Journal of Economic Methodology* 1:135–160.

[Wible(2008)] —. 2008. "The Economic Mind of Charles Sanders Peirce." *Contemporary Pragmatism* 5:39–67.

[Williams(1980)] Williams, Paul. 1980. "Bayesian Conditionalization and the Principle of Minimum Information." *British Journal for the Philosophy of Science* 31:131–144.

[Williamson(2010)] Williamson, Jon. 2010. *In Defense of Objective Bayesianism*. Oxford: Oxford University Press.

[Williamson(2000)] Williamson, Timothy. 2000. *Knowledge and its Limits*. Oxford: Oxford University Press.

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

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

[Zabell(1982)] Zabell, Sandy L. 1982. "W.E. Johnson's Sufficientness Postulate." *The Annals of Statistics* 10:1090–1099.

[Zabell(2005a)] —. 2005a. "Fingerprint Evidence." *Journal of Law and Policy* 13:143–179.

[Zabell(2005b)] —. 2005b. *Symmetry and its Discontents: Essays on the History of Inductive Probability*. Cambridge: Cambridge University Press.

[Zollman(2017)] Zollman, Kevin J.S. 2017. "The Credit Economy and the Economic Rationality of Science." *Journal of Philosophy* Forthcoming.