CHALLENGES TO BAYESIAN CONFIRMATION THEORY

John D. Norton

1 INTRODUCTION

Proponents of Bayesian confirmation theory believe that they have the solution to a significant, recalcitrant problem in philosophy of science. It is the identification of the logic that governs evidence and its inductive bearing in science. That is the logic that lets us say that our catalog of planetary observations strongly confirms Copernicus' heliocentric hypothesis; or that the fossil record is good evidence for the theory of evolution; or that the 3^oK cosmic background radiation supports big bang cosmology. The definitive solution to this problem would be a significant achievement. The problem is of central importance to philosophy of science, for, in the end, what distinguishes science from myth making is that we have good evidence for the content of science, or at least of mature sciences, whereas myths are evidentially ungrounded fictions.

The core ideas shared by all versions of Bayesian confirmation theory are, at a good first approximation, that a scientist's beliefs are or should conform to a probability measure; and that the incorporation of new evidence is through conditionalization using Bayes' theorem. While the burden of this chapter will be to inventory why critics believe this theory may not be the solution after all, it is worthwhile first to summarize here the most appealing virtues of this simple account. There are three. First, the theory reduces the often nebulous notion of a logic of induction to a single, unambiguous calculus, the probability calculus. Second, the theory has proven to be spacious, with a remarkable ability to absorb, systematize and vindicate what elsewhere appear as independent evidential truisms. Third is its most important virtue, an assurance of consistency. The larger our compass, the more we must digest evidence of diverse form and we must do it consistently. Most accounts of evidence provide no assurance of consistency in their treatment of larger bodies of evidence.[1] Bayesian confirmation theory affords us a simple picture: the entire bearing of evidence at any moment is captured by a probability distribution. No matter how large a body of evidence we contemplate,

[1]Some even fail in simple cases. Consider an enumerative induction on the white swans of Europe, which lets us conclude that all swans are white; and an enumerative induction on the black swans of Western Australia, which leads us to the contradictory conclusion that all swans are black.

Handbook of the Philosophy of Science. Volume 7: Philosophy of Statistics.
Volume editors: Prasanta S. Bandyopadhyay and Malcolm R. Forster. General Editors: Dov M. Gabbay, Paul Thagard and John Woods.

we will not be led to contradictions in our evidential judgments as long as we form and update our beliefs in conformity with the probability calculus.

Perhaps because of these virtues, Bayesian confirmation theory now enjoys the status of the leading account of induction and confirmation in the philosophy of science literature. Those who expect this circumstance to persist should recall that the present success of Bayesianism is relatively momentary. The probability calculus was born over 350 years ago in the seventeenth century in correspondence between Pascal and Fermat. One hundred years later, in the eighteenth century, the Reverend Thomas Bayes published his theorem as part of a proposal that probability theory be used to answer Hume's inductive skepticism. Nonetheless, the idea that the probability calculus afforded the right way to understand inductive inference remained a minority view. The dominant view of induction in the nineteenth century followed in the tradition of Bacon and his tables, with its most influential expression in John Stuart Mills' *System of Logic* (1891). This dominance persisted through to the mid twentieth century.[2] It competed with the hypothetico-deductive approach to inductive inference. That view's venerable history can be traced through Descartes' method of hypothesis to ancient Athens were astronomers were, as legend has it, asked by Plato to find the geometrical planetary constructions that would "save the phenomena" of astronomy [Duhem, 1969]. As recently as a few decades ago, the philosophy of science literature was filled with proposals for dealing with Hempel's [1945] notorious "paradox of the raven." That paradox arose within a development of Hempel's "satisfaction" criterion of confirmation, that was itself merely the ancient notion of enumerative induction transported from the context of Aristotle's syllogistic logic to modern, first order predicate logic. These differing approaches played out against persistent claims by Popper [1959] and his followers that the very notion of a logic of induction is mistaken. The Bayesian approach to induction and confirmation rose to dominance in philosophy of science only slowly in the course of the second half of the twentieth century,[3] with its importance codified by such works as Howson and Urbach [2006].

Given this history of competing traditions in induction and confirmation rising and falling over the centuries, it seems only prudent to expect that the Bayesian approach will recede from its present prominence and once again be merely one of several useful instruments for assessing inductive inference relations. The goal of this chapter is to review the weaknesses recounted in the literature that may drive this decline. It will draw on two sources. First are those outside the Bayesian fold who present direct challenges; second are shortcomings identified by Bayesians, who in turn have sought modifications within the Bayesian system. These two

[2]Mill's methods remain to this day the single most important methodological idea for computer repair people and auto mechanics seeking to diagnose problems. Their reach persists in the methodological literature at least to Skyrms [1975], whose Ch. IV gives a detailed development. See the historical survey of Blake, Ducasse and Madden [1960] for a broader sense of the minor role probability theory has played in the longer term history of ideas of scientific methodology.

[3]Glymour [1980, p. 64] identifies the influence of Carnap [1950] as decisive.

literatures overlap sufficiently for it to be impractical for me to disentangle them.[4]

The survey will proceed from larger to smaller issues. Section 2 will review global challenges to Bayesian confirmation theory. These challenges derive from differences in the basic conception of inductive inference and the principles that govern it. They include the view that notions of inductive inference are merely artifacts subservient to whatever methods may reliably get us to the truth; and the proposal that there is no universally valid logic of induction, but only localized logics adapted to individual domains. Sections 3, 4 and 5 are based on a decomposition of Bayesian confirmation theory into independent components, which are presented as intuitive notions pertaining to confirmation. Additivity amounts to a decision that degrees of probability span belief and disbelief, not belief and ignorance; the essence of Bayesian dynamics is glossed as the incorporation of new evidence by a simple rule of "refute and rescale" prior belief. This decomposition supports a catalog of challenges to the specific Bayesian commitment. Among them is "the problems of the priors," which proves to be multiple problems aggregated under one label. One — here called first problem — is that additivity precludes prior probabilities that represent a state of complete ignorance. The second problem derives from the simplicity of refute and rescale dynamics. It precludes specification of a neutral prior probability distribution that exerts little influence on the future course of conditionalization. Section 6 reports some further challenges and Section 7 is a brief conclusion.

Some topics lie largely outside the compass of this chapter. In statistical practice, as opposed to philosophy of science, an epochal battle continues to be waged between proponents of traditional Neyman-Pearson hypothesis testing and Bayesian approaches. That debate has been of far less importance in philosophy of science and will be addressed by Deborah Mayo and Aris Spanos (this volume).[5] Within the Bayesian literature itself, the dominant problem has been to find the correct interpretation of probability: subjective, objective or logical? To someone who is at a slight distance from Bayesianism, the debate seems misdirected in setting standards of precision for the interpretation of a central term of a theory that are rarely achieved elsewhere and are probably unachievable for probability. Its energy is reminiscent of an internecine feud among sect members who agree vastly more than they differ, but nonetheless tolerate no doctrinal deviations. This topic will be addressed only in so far as it arises in broader challenges to the Bayesian approach. These debates are addressed by Sandy Zabell (This Volume) and Philip Dawid (This Volume) as well as survey volumes [Galavotti, 2005; Gillies, 2000; Mellor, 2005].

[4]Hence I apologize in advance to Bayesians who feel that their work has been mischaracterized as a challenge to Bayesianism.

[5]For a brief account from the perspective of a philosopher sympathetic with the Bayesian approach, see [Howson, 1997]; and for a brief history from someone with opposing sympathies, see [Mayo, manuscript].

2 COMPETING ACCOUNTS OF THE NATURE OF INDUCTIVE INFERENCE

Any theory of inductive inference depends upon one or more principles or presumptions that distinguish the right inductive inference relations. These principles must be there, whether they are made explicit, or, as is the more usual case, left tacit. The most fundamental of the challenges to Bayesian confirmation theory come from differing views of these principles.

2.1 Alternatives

Given the large number of approaches to inductive inference, one might imagine that there is a plethora of competing principles. A recent survey [Norton, 2005], however, shows that most accounts of inductive inference can be grouped into one of three families, there called "inductive generalization," "hypothetical induction" and "probabilistic induction." Each family is based upon a distinct inductive principal and the different members emerge from efforts to remedy the deficiencies of the principle.

While it is impossible to give detailed descriptions of each family, it is useful here to indicate the breadth and longevity of the first two, which compete with the third probabilistic family. Accounts of induction belonging to inductive generalization depend on the principle that the instance of a hypothesis confirms the generalization. This ancient notion was implemented in syllogistic logic as enumerative induction: an A that is B confirms that all As are B. The principal weakness of this archetype of the family is that it may allow rather little to be inferred. We cannot use it to pass from the evidence of a 3^oK cosmic background radiation to the theory of big bang cosmology. The embellishments of the archetype are devoted to extending the reach of enumerative induction. Hempel's satisfaction criterion reformulates the basic notion within the richer, first order predicate logic. Mill's methods extend it by licensing the inference to a new notion, cause. In the Method of Agreement, if A is seen to follow a, then we infer this pattern will persist generally and label a as the cause of A. Glymour's "bootstrap" account of confirmation uses instance confirmation to enable evidence to confirm hypotheses that employ theoretical notions than are not explicitly mentioned in the evidence. For our purposes, what matters is that all these approaches to induction depend upon the basic principle that an instance confirms the generalization.

Accounts of induction belonging to hypothetical induction depend on the principle that the ability of an hypothesis deductively to entail the evidence is a mark of its truth. The archetype of accounts is this family is the saving of the phenomena in astronomy. That some model of planetary motions correctly fits and predicts the observed positions of the planets is a mark of its truth. The principal weakness of this principle is that it assigns the mark too indiscriminately. There are many planetary systems — geocentric, heliocentric and more — that can save the phenomena of astronomical appearances. They are all equally assigned the mark of

truth, even though there must be some sense in which one saves the appearances better than another. Embellishments of this basic notion are intended to rein in the reach of hypothetical induction by adding further conditions that must be met to earn the mark. The hypothesis that saves the phenomena must be simple, or simpler than its competitors; or we must in addition be assured of an exclusionary clause: that were the hypothesis that saves the phenomena false, then it would be unlikely for the phenomena to obtain.[6] Other versions require hypotheses not just deductively to entail the evidence, but also to explain it. Ptolemy's model of planetary motion entails that Venus and Mercury will always appear close to the sun. Copernicus' model explains why: these planets orbit the sun. Finally, another sort of embellishment enjoins us to take into account the historical process through which the hypothesis was generated. It must be generated by a method known to be reliable. Generating hypotheses *ad hoc* is not judged reliable. Parapsychologists hypothesize retrospectively that their experiments failed in the presence of skeptics because of the unintended obstructive influence emanating from the skeptics. We are free to discount this as an *ad hoc* hypothesis.

2.2 Bayesian Annexation

All these accounts represent challenges to Bayesian confirmation theory in that they are grounded in and can be developed in a framework that does not need the familiar structures of Bayesian confirmation theory. The natural Bayesian rejoinder is that the two basic principles and their embellishments can be absorbed and vindicated by the Bayesian system.

Whether one finds this possibility compelling depends on whether one finds this gain worth the cost of adopting the extra structures and commitments that Bayesian confirmation theory requires (such as will be discussed in Sections 3-5 below). The absorption and vindication can sometimes be quite successful. The clearest case is the success of the Bayesian analysis of when the success of an hypothesis H entailing true evidence E should count as inductive support for H [Earman and Salmon, 1992, §2.7]. A familiar application of Bayes' theorem[7]

$$\frac{P(H|E)}{P(\sim H|E)} = \frac{P(E|H)}{P(E|\sim H)}\frac{P(H)}{P(\sim H)}$$

shows that the degree of support is controlled essentially by the likelihood ratio $P(E|H)/P(E|\sim H)$. If it is high, so that the evidence is much less likely to come

[6] One important form of this is Mayo's [1996] error statistical approach to inductive inference in science. It requires evidence to provide a "severe test" of an hypothesis. The above exclusionary clause is realized in Mayo's [1996, p. 180] severity requirement: "There is a very low probability that test procedure T would yield such a passing result, if [the hypothesis] H is false." It of special interest here since its debate with Bayesian confirmation theory replicates in philosophy of inductive logic the tensions between Neyman-Pearson hypothesis testing and Bayesian approaches in statistics. For a Bayesian rejoinder, see [Howson, 1997, §4].

[7] The terms have their obvious meanings. For example, $P(H|E)$ is the (posterior) probability of H conditioned on E; $P(H)$ is the (prior) probability of H against the tacit background. See Section 5 below.

about if H is false than if it is true, then the posterior probability $P(H|E)$ increases correspondingly with respect to the prior probability $P(H)$. If the likelihood ratio is close to one so that E is pretty much as likely to come about whether or not H is true, then $P(H|E)$ enjoys no such increase. The analysis is appealing in that it vindicates the basic principle of hypothetical induction, as well as going beyond it in indicating when it will not work.

Other efforts at absorption have been less successful. In Section 5 below, we shall the difficulties facing Bayesian efforts to explicate notions like predictive power and simplicity. The problem is that the simple structures of the Bayesian theory do not have enough resources to distinguish the case of an hypothesis H merely entailing evidence E from it entailing it in a virtuous way that merits confirmatory rewards.

There also seems to be some troubling elasticity in the way that a Bayesian vindication of independent inductive norms is achieved. Sometimes it seems that almost any norm could be justified. The problem is illustrated by difficulties faced in a simple example due to Carnap, as summarized by Earman and Salmon [1992, pp. 85-89]. The traditional form of inductive inference called "example" is a weakening of enumerative induction that merely infers from instances to instances: these cases of x are X; so that new case of x will also be X. The illustration deals with finitely many individuals $a, b, c, \ldots$ that may carry a property F (written "Fa" etc.) or fail to carry it (written "$\sim Fa$" etc.) The individual possibilities are described by state descriptions, such as $Fa\&Fb\& \sim Fc\& \ldots$. Carnap initially assigned equal probability to each state description. The immediate outcome was that the argument form "example" failed. Learning that Fa and Fb are true leaves the probabilities of Fc and $\sim Fc$ unaffected. That could be offered as a proof of the failure of "example," if one was inclined against the argument form. Carnap, however, was not so inclined. He chose to readjust the probabilities on the state descriptions in a way that favored correlations between the properties, so that "example" ends up being vindicated.

Closer reflection on Carnap's illustration shows that the vindication (or otherwise) of the inductive argument form "example" does not depend on anything inherent in the probability calculus. Rather it depends upon additional assumptions we make. In this case it is our determination of whether the individuals under consideration are correlated in the loose sense that one individual carrying F tends to go with another also doing so. If we decide they are so correlated, then "example" is vindicated, whether we express the vindication formally within a probabilistic analysis or merely as a restatement of the informal notion of correlation.

This suggests that the Bayesian vindication of inductive norms is less a matter of extracting them from the probability calculus and more one of our introducing them as independent assumptions that can be expressed in probabilistic language.[8]

[8]Other examples are easy to find. The Bayesian analysis of hypothetical induction above depended essentially on our specifying as an independent assumption that the ratio $P(E|H)/P(E| \sim H)$ is large. That amounts to the external assumption that E would much

That makes them far less impressive than they initially seemed. One virtue does persist: since we are likely to want to combine many possibly competing inductive principles, if we do it within the framework of the probability calculus, we have some assurance that they will be combined consistently.

2.3 Bayesian Foundational Principles

The strongest response to these challenges of other principled accounts would be for Bayesian confirmation theory to display its own principled foundation, justify it and argue for its superiority. However prospects here are not strong. Bayesian confirmation theory is distinctive among accounts of confirmation theory in that a great deal of effort as been expended in seeking to justify it. These efforts are usually associated with different interpretations of probability and vary markedly in character according to the interpretation. While the literature on this problem is enormous, there seems to be no vindication that is widely agreed to be successful. Rather the agreement lies only with the conclusion that has to be reached — that the probability calculus is the logic of induction. Each Bayesian seems to find his or her own path to this conclusion, often from widely diverging starting points and with critical dismissal of other starting points.

A few examples will illustrate the divergences in the starting points and their difficulties. The most promising approach is associated with a relative frequency interpretation of probability, such as sought in [Salmon, 1966, pp. 83-96]. In it, the probability of a scientific hypothesis would simply be the relative frequency of true hypotheses in a reference class of hypotheses of the appropriate type. Since relative frequencies behave mostly like probabilities, it is easy to conclude that the probability calculus is the natural logic of induction, which is largely reduced to the simple task keeping count of truth in relevant reference classes. The proposal fails because of the insurmountable difficulty of defining which are the appropriate reference classes, let alone assuring that the appropriate relative frequencies are unique and well defined in the resulting infinite sets.

Another popular approach seeks to identify necessary conditions that any inductive logic must meet. A version has recently been developed by Jaynes [2003] following Cox [1961]. The attendant interpretation of probability is objective, in the sense that, in any one circumstance, there is one correct probability assignment, even though we may not know it; and is "logical" in the sense that the probabilities are treated as degrees of a logical relationship of support. The approach is strong in that, if there is a unique logic of induction to be found, circumscribing it by an ever-narrowing fence of necessary conditions will isolate it. Its difficulty is that it needs to convince the reader of the necessity of a list of conditions: that belief comes in numerical degrees that are always comparable; that the degree of belief assigned some proposition must be a function of those assigned to its disjunctive parts; and further, more specialized suppositions. We

more likely likely come about if H is the case than if it is not, which already pretty much gives us the principle wanted.

shall dissect these sorts of conditions in Sections 3, 4 and 5 below and find that they seem necessary only if one believes from the start that degrees of belief are probabilities.[9]

Finally, in association with a subjective approach to probability, there are arguments that arrive at the probability calculus by urging that any distribution of belief not in accord with the probability calculus exposes us to the danger of a combination of wagers that assure a loss. These are the Dutch book arguments, best known from the work of de Finetti [1937]. In a related approach, such as Savage [1972], natural conditions on our actions are translated into conditions on beliefs. The first and most evident weakness of these proposals is that they require a quite substantial body of presumptions on our actions and preferences and the rules that should be used to relate them to our beliefs. The arguments can be defeated by denying a major presumption. I may simply be unwilling to bet; or I may never find a bet such that I am indifferent to which side I take; or I may harbor intransitive preferences.[10]

However there is a deeper problem. These approaches fundamentally change the question asked. It was: when is this observation good evidence for that hypothesis? The presumption of the question is that evidential support obtains independently of our beliefs and actions. These arguments then change the question in two steps that introduce dependencies on our beliefs and actions. First, the question of evidential support (Does the fossil record support evolutionary theory?) is replaced by the question of distributions of belief (How do we distribute our belief over evolutionary theory and its competitors?). At this stage, prior to consideration of

[9]For example, the belief assigned to $A\&B$ is assumed to be a function of the belief assigned to A and to B given A. It is natural to assume this if one imagines all along that "belief" is just a code word for probabilities or frequencies. However, if one does not prejudge what "belief" must be, the assumption of this specific functional dependency is alien and arbitrary. For a natural counterexample that violates the rule, see [Norton, 2007, p. 144, fn. 6]. Similar difficulties arise for attempts to use "loss functions" or "scoring rules" as a means of vindicating the Bayesian system, such as in [Joyce, 1998]. They depend on using a scoring rule that picks out the subspace of additive measures as local optima in the larger space of all belief functions. Any such scoring rule must be selected carefully so as to enforce a coupling between belief in an outcome A and its negation $\sim A$; for an arbitrary change in the belief assigned to A must force a suitable change in the belief in $\sim A$ to ensure that optimal scoring occurs only in the subspace of additive measures. That coupling is just the sort of dependency noted in Section 4.1 and in (A') that is tantamount to choosing additivity. However natural they may seem, the specific conditions chosen to endow scoring rules with this particular property amount to a prior decision to interpret low belief values as disbelief rather than ignorance or some mix of disbelief and ignorance. In the latter case, belief in an outcome is no longer closely coupled to the belief in its negation.

[10]Hájek [2008]) argues that four of the principal means of vindicating Bayesianism fail, each in the same way. The means are the Dutch book arguments, representation theorem arguments, calibration arguments and gradational accuracy arguments. Each depends on displaying a theorem that is interpreted to show that rationality forces Bayesianism. Each vindication, Hájek urges, overlooks a "mirror-image" theorem that shows the reverse, that rationality requires degrees of belief *not* to be probabilities. For example, the Dutch book theorem assures the existence of a combination of fair bets that forces a sure loss for you if your numerical beliefs are not probabilities. Its mirror-image is the "Czech book theorem." It assures that a benevolent Czech bookie can find a combination of fair bets that assures a certain gain for you just if your numerical degrees of belief are not probabilities.

actions, we are to accept that there is no one right distribution of belief. Yours is presumed as good as mine. Second, rules are provided for translating beliefs into actions, so that the consequences of our actions can be assessed and used to restrict how we distribute our beliefs, while in general never eliminating the possibility of different belief distributions. These rules make most sense if we are at a racetrack or engaged in an engineering project where beliefs must be translated into design decisions. They are strained if our concern is the question originally asked: whether, for example, the observed motion of the galaxies provides good evidence for the hypothesized heat death of the universe, billions of years in the future — a circumstance independent of our beliefs and actions. The two-step process has produced an answer to a question, but it is not the question with which we started.[11]

2.4 Is There a Unique Logic of Inductive Inference?

So far, all the approaches considered — Bayesian and otherwise — have presumed that there is a unique logic of induction to be found. Another type of challenge to Bayesian confirmation theory denies this. Rather these challenges portray the conditions surrounding individual problems as determining which are the appropriate inductive strategies and, from this perspective, it turns out that there is no single logic of induction to be identified. Rather than having principled grounds for identifying the right inductive logic, these approaches offer principled grounds for denying that such a thing is possible.

A well articulated version of this approach is the learning theoretic paradigm, as developed in [Kelly, 1996]. It presumes that our goal is getting to the truth through some reliable method of discovery. At its simplest, an agent receives a potentially unlimited stream of data and must formulate hypotheses expressing the pertinent truths of the world, as the data accumulates. The central problem of the literature is to discern the sorts of strategies that will lead to convergence to the truth in the hypotheses formulated successively, with more rapid convergence prized. In certain simple worlds, these methods can be governed by familiar inductive notions. A data stream "day, night, day, night, day, ..." rapidly succumbs to the hypothesis that day follows night and night follows day. In this case, enumerative induction is all that an agent needs to formulate the true hypothesis quite early. Matters get more complicated as the problems get harder. Looking for regularities in what follows what, leads to limited statistical success if the problem is weather prediction at one location and the data stream is "rain, shine, shine, shine, rain, ..."; or stock market prediction with a data stream "up 3 points, up 2 points, down 20 points, up 3 points, ..." There is a small amount of inertia in both the weather and the stock market, so that conditions one day tend somewhat to be

[11] For a spirited assault on Dutch book arguments, see [Bacchus *et al.*, 1990]. They urge these arguments fail since an agent should simply refuse to accept a combination of bets that assures a loss; it is (pp. 504-505) "a matter of deductive logic and not of propriety of belief." Schick's [1986] view is similar.

replicated the next. These methods fail completely in inhospitable worlds. If the data stream is derived from a well-designed roulette wheel, "red, black, black, red, red, ...," no hypotheses on when red will follow black (other than chance) can succeed. And matters can be worse in a demonic universe, which delivers data maliciously tailored to deceive the particular agent in question.

In this reliabilist paradigm, inductive norms are parasitic on convergence to the truth. It is simply a matter of ingenuity to conceive universes in which a given inductive norm will be fruitful or harmful. For example, a standard supposition of scientific methodology is that we should only advance hypotheses that are consistent with the data at hand. However it turns out that, once we allow for the computational limitations of agents, a demand for consistency can conflict with reliability. So it may be efficient for us to offer an hypothesis inconsistent with the evidence, even though most accounts of induction, including the Bayesian, insists the hypothesis has been refuted. For a critique of Bayesian confirmation theory from the reliabilist perspective, see [Kelly and Glymour, 2004].

The idea that inductive strategies will vary according to the domain at hand is the central idea of the "material theory of induction" [Norton, 2003]. Its name derives from the idea that the license of an inductive inference does not come from the *form* of the sentences involved, as it does in deductive logic. Rather the license comes from the subject *matter* of the inference. Therefore the inductive logic applicable will vary from domain to domain as the licensing facts change; there is no universally applicable logic of induction.

In one restricted area, this notion of material facts fixing the inductive logic is already familiar to subjective Bayesians through Lewis' [1980] "principal principle." For processes for which objective chances are available, the principle enjoins subjective Bayesians to match their subjective degrees of belief to the physical chances. Quantum mechanical laws deliver chances for the timing of the decay of a radioactive atom. If we match our subjective degrees of belief concerning those timings to the quantum mechanical chances, then we are assured high belief in outcomes that are most likely to happen and low belief in those unlikely to happen. A by-product is that our degrees of belief automatically conform to the probability calculus.[12]

The principal principle naturally generalizes to the central claim of the material theory of induction: that the factual properties of the system under consideration will determine the appropriate inductive logic. It follows that there may be systems whose facts dictate that the applicable logic of induction is *not* probabilistic. Norton [2007, §8.3; forthcoming] describes certain physical systems whose governing laws are indeterministic in the sense that the complete specification of

[12]The facts that license probabilistic inferences need not be the sorts of physical chances recovered from stochastic theories like quantum mechanics. For example, imagine that one is a bookmaker accepting wagers on a definite matter of presently unknown fact: say, whether some recently discovered infectious illness is viral or bacterial in origin. If the factual conditions surrounding the bookmaking conform to all those assumed in Dutch book arguments, then those facts determine that one's degree of belief ought to conform to the probability calculus, on pain of exposure to a sure loss.

the present state of the system does not fix the future state; yet these governing laws provide no probabilities for the various possible futures. They simply tell us that, given such-and-such a present state, this or that future state is possible. In the corresponding inference problem, we are required to distribute beliefs over these possible futures, knowing the state of the present. If we distributed beliefs as probabilities rather than in some weaker form, we end up with beliefs that outstrip the full physical specification as given by the initial conditions and the applicable physical laws. We must assert that one possible outcome is twice as probable as another; or as probable as another; or half as probable. That is, we must pretend to know more than the physical laws, which are only able to assert that each outcome is possible, without any sense of "twice as possible" or "half as possible."

3 CHALLENGES TO FRAMEWORK ASSUMPTIONS

The challenges considered so far in Section 2 above have been global in the sense that they arise from tensions between Bayesian confirmation theory as a whole and other approaches to inductive inference. If we recall that Bayesian confirmation theory is a composite of many assumptions, then another class of challenges to Bayesian confirmation theory is identifiable. These are challenges to specific assumptions of Bayesian confirmation theory. In order to structure the inventory of these challenges, the assumptions comprising Bayesian confirmation theory will be divided into three parts, called: "Framework," to be discussed in this section; "Additivity," to be discussed in Section 4; and "Bayes' dynamics: to be discussed in Section 5.[13]

3.1 Framework Assumptions

What is common to all versions of Bayesian confirmation theory is the assumption that there is a real-valued magnitude, $P(H|E)$, that is one's subjective degree of belief or some objective measure of support for the hypothesis H from evidence E. This apparently simple assumption of a real valued magnitude can be decomposed further into a series of assumptions, each of which presumes the one before:

> *Precision.* There is a magnitude $P(A|B)$ that represents the degree of belief or support accrued to A given B.
>
> *Universal comparability.* It is always possible to compare two such degrees, finding one to be less than, equal to or greater than the other.
>
> *Partial order.* The comparison relation "no less than" $\leq$ is a partial

[13]This disassembly is similar to the one effected in [Norton, 2007], which in turn draws on an extensive literature in qualitative probability. For surveys see [Fine, 1973; Fishburn, 1986].

order: it is reflexive[14], antisymmetric[15] and transitive[16].

Real values. The degrees are real-valued.

Different challenges would require us to halt at different stages as we pass up this hierarchy. Lest that seem odd, the notion that a cautious ascent is needed may be more palatable if we consider an analogous problem of assigning degrees of simplicity to hypotheses. The same hierarchy can be formed, but few are likely to ascend all the way to the idea of single real value that measures the degree of simplicity of all hypotheses we encounter. We may stall at the start, judging the very idea of a degree of simplicity inadmissible. Or we may allow comparison among closely related hypotheses only. Jeffreys [1961, p.47], for example, wished to assign greater prior probability to simpler hypotheses. So he defined the complexity m of certain differential equations as the sum of their order, degree and the absolute values of their coefficients. Unit prior probability was then distributed over the resulting complexity classes as a decreasing function of m, such as 2^{-m} or $6/\pi^2 m^2$. Jeffreys' degree is defined only for a specific class of equations. It would not sustain universal comparability in that we could not compare the degree of simplicity of the hypothesis of a linear relationship between the velocity and distance to the galaxies with that of the hypothesis of the germ theory of disease.

3.2 The Very Idea.

The most fundamental objection is that there is something mistaken about the very idea of a degree of belief or support that evidence lends to a hypothesis. This objection can rest on the visceral sense that such assignments are simply spurious precision: that we can conjure up a number does not mean that we are measuring something. A colorful statement of the mismatch of Bayesian analysis with diagnostic clinical practice is Feinstein's [1977] "The haze of Bayes, the aerial palaces of decision analysis, and the computerized Ouija board."

In a different direction, Glymour's [1980, Ch. III] "problem of old evidence" speaks against the idea that the fundamental quantity $P(A|B)$ presumed in the framework is an adequate basis for an account of inductive inferences. The import of evidence E on hypothesis H is gauged by comparing the prior probability $P(H|B)$ with the posterior $P(H|E\&B)$, where B is the totality of background knowledge. It frequently happens that the evidence E is already part of our background knowledge. The celebrated case concerns the motion of Mercury, which, at the start of the 20th century, was know to deviate slightly from the predictions of Newtonian gravitation theory. In November 1915, Einstein found that his newborn general theory of relativity "H" predicted exactly the observed deviations

[14] For all admissible $A, B, P(A|B) \leq P(A|B)$.

[15] For all admissible A, B, C and D, if $P(A|B) \leq P(C|D)$ and $P(C|D) \leq P(A|B)$, then $P(A|B) = P(C|D)$.

[16] For all admissible A, B, C, D, E and F, if $P(A|B) \leq P(C|D)$ and $P(C|D) \leq P(E|F)$, then $P(A|B) \leq P(E|F)$.

"E". The universal agreement is that E provided strong support for H. However, since E is part of the background then known to Einstein, we have

$$E\&B = B$$

so that

$$P(H|E\&B) = P(H|B)$$

which shows that E is evidentially inert.

If this objection is sustained, it means that a confirmation theory employing only the quantities $P(H|E)$ and $P(H|E\&B)$ cannot support the judgment that E is good evidence for H.[17] The obvious escape is to replace the prior $P(H|B)$ with an adjusted prior $P(H|B')$, where B' is the background B with E somehow excised. Glymour's original presentation sought to block this escape by urging that it is quite unclear how this excision could be effected. Other, more elaborate escapes are reviewed in [Earman, 1992, Ch. 5], including the notion that we assign evidential value to learning the logical relation between some hypothesis H and the evidence E that is already, unbeknown to us, part of our background B.

3.3 Universal Comparability

Once we allow that the degree $P(A|B)$ is always defined, it may still not follow that all degrees are comparable. Keynes [1921, pp. 37-40] allows that all degrees will lie between those assigned to impossibility and certainty, but urges that not all intermediate degrees are mutually comparable.

The clearest way that comparability can fail is when we conditionalize on very different sorts of backgrounds so that we are distributing belief over incommensurable possibilities. For example, Norton [2007, pp. 147-48] expands a concern expressed by Humphreys [1985] in the context of propensity interpretations to suggest that the direct degrees $P(E|H)$ and the inverse $P(H|E)$ may measure very different things. $P(E|H)$ may be derived from the computing of a quantity in a stochastic theory: the chance of E, the decay of some particular radioactive atom in some time as deduced from H, the laws of quantum physics. The inverse $P(H|E)$ assigns degrees to different physical laws on the basis of the evidence of radioactive decay. The first deals with possibilities licensed by a known theory and is grounded in the science; the second deals with speculation on how the world might admit different laws whose range is delimited by our imagination. Even if we can find protocols that allow us to assign numbers in each case, it is not clear that there is anything meaningful in the comparison of those numbers — just as it is meaningless to compare degree of temperature and degrees Baumé of specific gravity, even though both are real numbers and arithmetic allows the

[17] Glymour's original presentation used Bayes' theorem to deduce $P(H|E\&B) = P(H|B)$, although the problem is independent of Bayes' theorem.

comparison.[18]

That there are different sorts of uncertainty even in decision theoretic contexts is illustrated by Ellsberg's [1961] urns. Consider two urns, each containing 100 balls. The first has some combination of red and black balls. Any ratio from 0-100% red is possible and we are completely uncertain over which. The second urn has exactly 50 red and 50 black balls. A ball will be drawn from each urn. What is our belief that it will be red? For both urns, the symmetry of the arrangements requires that we be equally certain of a red or a black, so, in drawing from both urns we should be indifferent to making bets on red or on black. Any protocol for converting these inclinations to bet into degrees of belief must respect that symmetry. Therefore, if the protocol generates probabilities solely on the basis of these inclinations, in each case it will return a probability of 0.5. However Ellsberg suggests that most of us prefer the bet on the urn of known composition, so that the degrees of belief assigned to red ought to differ in the two cases. For our purposes, this illustrates the possibility of different senses of uncertainty that may not be readily comparable if we try to characterize them in a single magnitude. These different senses reflect a distinction to be made in Section 4 below between disbelief and ignorance.

Jaynes [2003, pp. 658-59] regards such quibbles over different senses of uncertainty with characteristic, entertaining distain. He makes an analogy to a mineralogist who may classify rocks with different parameters but can still find a way to trade off changes in one parameter against the other to recover a single scale of comparison.

The idea that universal comparability must be discarded is a consequence of a significant development of the Bayesian approach that has adopted several forms. What they have in common is the notion that a single number may not be rich enough to capture the extent of belief. That concern is readily motivated whenever we are pressed to assign a definite probability to some outcome — say, whether it will rain tomorrow. Assigning a probability 0.6 is too precise. Our belief is better represented by something vaguer, a probability somewhere around 0.6. In one approach, we might replace definite values by intervals of values; the probability of rain is 0.5 to 0.7. In another we might take set of probability measures as representing our belief states — say the set of all probability measures that assign a probability of 0.5 to 0.7 to rain tomorrow. Developing these proposals into well functioning theories is not straightforward. For developments of these notions, see [Kaplan, 1998; Kyburg, 1959; Kyburg and Teng, 2001; Levi, 1974; 1980, Ch. 9; Walley, 1991].

For our purposes, what matters is that these proposals break with the framework sketched in that they deny universal comparability. Take the simplest case of interval-valued degrees of belief. The belief in some outcome represented by the

[18] On the basis of their being "sufficiently disparate," Fishburn [1986, p. 339] offers the following pair as candidates for failure of comparability: "A=Mexico's City's population will exceed 20,000,000 by 1994; B=The first card drawn from this old and probably incomplete bridge deck will be a heart."

interval [0.8, 0.9] is greater (in the sense of being closer to certainty) than the interval [0.1, 0.2]. But the intervals [0.3, 0.7] and [0.4, 0.6] are none of greater than, less than or equal to each other. Analogous effects arise for richer structures that dispense with single valued degrees of belief.

3.4 Transitivity

If we allow that the degrees are always comparable, it does not yet follow that they form a partial order, which is reflexive, antisymmetric and transitive. Of these three properties, transitivity is of the greatest concern. Norton [2007, §3.2.2] has suggested that transitivity may fail in certain complicated cases if common wisdoms on inductive support obtain. That is, if we have a case of evidence entailed by several different hypotheses, it is routine to say that the evidence supports one of these hypotheses more than another if the first hypothesis displays more of some relevant virtue, such as greater simplicity, greater explanatory power or greater fecundity. Using three hypotheses and three virtues, it is easy to envisage structures in which the first hypothesis is better confirmed than the second, the second better than the third and the third better than the first. This violates transitivity.[19]

3.5 Real Values

Once it is allowed that the degrees may be partially ordered, it is still not assured that they must be real valued, or isomorphic to the reals or some interval of real numbers. A great deal of effort has been expended in determining what additional assumptions are needed to assure that the degrees are real valued. These additional assumptions have been given many forms and are generally known as "Archimedean axioms." [Fishburn, 1986, pp. 341-42]. Their function is to block the possibility of value sets that extend beyond real values in admitting infinitesimally small or infinitely large values.

The familiar illustration in this literature of how partially ordered degrees may fail to be real valued is generate by degrees that are ordered pairs $\langle x, y\rangle$ of reals x and y [Jeffreys, 1961, pp. 19-20]. They are partially ordered by the rule

$$\begin{aligned}\langle X, Y\rangle > \langle x, y\rangle \quad & \text{when } X > x, \text{ no matter the values of } Y, y \\ & \text{or, if } X = x, \text{ when } Y > y\end{aligned}$$

There is no way to map these degrees $\langle x, y\rangle$ onto the reals so that the partial order is preserved.

We could imagine a two parameter family of hypotheses $H_{x,y}$ such that our degrees of belief in the various hypotheses may be ordered according to the rule

[19] For the example of a slightly bent coin, Fishburn [1986, p. 339] describes three propositions for which intransitivity "do[es] not seem unreasonable": "A=The next 101 flips will give at least 40 heads; B=The next 100 flips will give at least 40 heads; C=The next 1000 flips will give at least 460 heads." The suggestion is that we could have the same belief in A and in C; the same belief in C and in B, but that we must have more belief in A than B.

just sketched. One might doubt whether such a family could arise in realistic problems and take that doubt as a reason to discount the challenge. Or one might be concerned that adopting a Bayesian confirmation theory presumes in advance that such a distribution of belief is inadmissible. Such a presumption is not compatible with the universal applicability supposed for the theory.

4 ADDITIVITY

4.1 The Property

Probability measures are additive measures. That means that they conform to the condition:

$$\begin{array}{l}\text{If } A \text{ and } B \text{ are mutually exclusive then, for any } C,\\ P(A \vee B|C) = P(A|C) + P(B|C).\end{array} \qquad (A)$$

Additivity is the most distinctive feature of the probability calculus. In the context of confirmation theory, it amounts to assigning a particular character to the degrees of belief of the theory. The degrees of this additive measure will span from a maximum value for an assured outcome (conventionally chosen as one) to a minimum value for an impossible outcome (which must be zero[20]). If the high values are interpreted as belief with unit probability certainty, then it follows that the low values must represent disbelief, with zero probability complete disbelief. To see this, recall that near certain or certain belief in A corresponds to near complete or complete disbelief in its negation, $\sim A$.[21] If we assign high probability 0.99 or unit probability to some outcome A, then by additivity the probability assigned to the negation $\sim A$ is 0.01 or zero, so that these low or zero values correspond to near complete disbelief or complete disbelief.

This interpretation of low probability as disbelief is already expressed in the functional dependency between the probabilities of outcomes and those of their negations. We have $P(\sim A|C) = 1 - P(A|C)$, which entails

$$P(\sim A|C) \text{ is a strictly decreasing function of } P(A|C) \qquad (A')$$

More generally, taking relative negations, $P(\sim A\&B|C) = P(B|C) - P(A\&B|C)$, we recover the more functional dependency

$$P(\sim A\&B|C) \text{ is a strictly decreasing function of } P(A\&B|C) \qquad (A'')$$

and a strictly increasing function of $P(B|C)$.

These dependencies tell us that a high probability assigned to an outcome corresponds to a low probability assigned to its negation or relative negation, which is the characteristic property of a scale of degrees that spans from belief to disbelief.

[20] If Imp is an impossible outcome, we have Imp = Imp $\vee$ Imp. Since Imp and Imp are mutually exclusive in the sense that Imp & Imp = Imp, additivity applies so that P(Imp) = P(Imp) + P(Imp), from which we have P(Imp)=0.

[21] I take it as a definition that *dis*belief in A is the same thing as belief in *not-A*.

While additivity (A) entails weaker functional dependencies (A') and (A''), the gap between them and (A) is not great. If the functional dependencies (A') and (A'') are embedded in a natural context, it is a standard result in the literature that the resulting degrees can be rescaled to an additive measure satisfying (A). (See [Aczel, 1966, pp. 319-24; Norton, 2007, §7].)

4.2 Non-Additive Measures: Disbelief versus Ignorance

Once it is recognized that the additivity (A) of the probability calculus amounts to selecting a particular interpretation of the degrees, then the ensuing challenge becomes inevitable. In many epistemic situations, we may want low degrees to represent ignorance or some mix of ignorance and disbelief, where ignorance amounts to a failure to commit to belief or, more simply, an absence of belief or disbelief. Shafer [1976, pp. 22-25] considers how we might assign degrees of belief to the proposition that there are living beings in orbit around Sirius, an issue about which we should suppose we know nothing. No additive measure is appropriate. If we assign low probability to "life," additivity then requires us to assign high probability to "no-life." That asserts high certainty in there being no life, something we do know. The natural intermediate of probability 1/2 for each of the two outcomes "life" and "no-life" fails to be a usable ignorance value since it cannot be used if we are ignorant over more than two mutually exclusive outcomes.

This example makes clear that introducing an element of ignorance requires a relaxation of the functional dependencies (A') and (A''). It must be possible to assign a low degree of belief to some outcome without being thereby forced to assign a high degree to its negation. Measures that allow this sort of violation of additivity are "superadditive": if A and B are mutually exclusive then, for any $C, P(A \vee B|C) \geq P(A|C) + P(B|C)$. The extent to which the equality is replaced by an inequality is the extent to which the measure allows representation of ignorance.

The best-known superadditive calculus is the Shafer-Dempster calculus. In it, an additive measure m, called a " basic probability assignment," is defined over the power set of the "frame of discernment" Θ, so that $\Sigma_{A\subseteq\Theta} m(A) = 1$, where the summation extends overall all subsets of Θ, and $m(\{\}) = 0$. This basic measure probability assignment is used to generate the quantity of interest, the "belief function" Bel, which is defined as

$$\mathrm{Bel(A)} = \Sigma_{\mathrm{B}\subseteq\mathrm{A}}\mathrm{Bel(B)}$$

for any A in Θ, where the summation is taken over all subsets B of A. These belief functions allow blending of disbelief and ignorance. For example, that we are largely ignorant over the truth of "life" can be represented by

$$m(\text{life}) = 0.1 \quad m(\sim \text{life}) = 0.1 \quad m(\text{life} \vee \sim \text{life}) = 0.8$$

which induces the belief function

$$\mathrm{Bel(life)} = 0.1 \quad \mathrm{Bel}(\sim\mathrm{life}) = 0.1 \quad \mathrm{Bel(life} \vee \sim\mathrm{life}) = 1$$

4.3 Complete Ignorance: The First Problem of the Priors

In Bayesian confirmation theory, prior probability distributions are adjusted by Bayes' theorem to posterior probability distributions that incorporate new evidence learned. As we trace this chain back, the prior probability distributions represent states of greater ignorance. Bayesian confirmation theory can only give a complete account of this learning process if it admits an initial state representing complete ignorance. However representing ignorance is a long-standing difficulty for Bayesian confirmation theory and the case of complete ignorance has been especially recalcitrant. It is designated here as the first problem of the priors, to distinguish it from another problem with prior probability delineated below in Section 5.6 below.

There are two instruments already in the probability literature that are able to delimit the representation of this extreme case, the epistemic state of complete ignorance. Norton [2008] has given an extended analysis of how the two may be used to do this. The first instrument is the principle of indifference. It asserts that if we have no grounds for preferring one outcome to a second, then we should assign equal belief to both. This platitude of evidence is routinely used to ground the classical interpretation of probability and famously runs into trouble when we redescribe the outcome space. Complete ignorance in one description is equivalent to complete ignorance in another. That fact allows one to infer quite rapidly that the degree of belief assigned to some compound proposition $A \vee B$ should be the same as the degree of belief assigned to each of the disjunctive parts A and B, even though A and B may be mutually exclusively.[22] Since no probability distribution can have this property, it is generally concluded that there is something wrong with the principle of indifference.

The difficulty is that the principle of indifference is not so easily discarded. It is a platitude of evidence. If beliefs are grounded in reasons and we have no reasons to distinguish two outcomes, then we should have the same belief in each. The alternative is to retain the principle of indifference and discard the notion that a probability distribution can adequately represent complete ignorance. Instead we are led to a representation of complete ignorance by a non-probabilistic distribution with three values: *Max* and *Min* for the extreme values of certainty and complete disbelief and *Ig* ("ignorance") for everything in between

$$\begin{aligned} \text{Degree } (A) &= \mathit{Max}, \text{ for } A \text{ an assuredly true outcome} \\ &= \mathit{Min}, \text{ for } A \text{ an assuredly false outcome} \\ &= \mathit{Ig}, \text{ for } A \text{ any contingent}^{23} \text{ outcome} \end{aligned} \qquad (I)$$

[22] An example is von Mises' famous wine-water problem. We have a glass with a mixture of wine and water, knowing only that the ratio of wine to water lies in 1/2 to 2. So we are indifferent to each of the intervals wine to water: (a) 1/2 to 1, (b) 1 to 1 1/2, (c) 1 1/2 to 2; and assign equal probability of 1/3 to each. However if we redescribe the problem in terms of the ratio of water to wine, we end up assigning equal probability of 1/3 to the intervals water to wine (a′) 1/2 to 1, (b′) 1 to 1 1/2, (c′) 1 1/2 to 2. Now the interval (a) describes the same outcome as the disjunction (b′) ∨ (c′). So we assign probability 1/3 to the disjunction (b′) ∨ (c′) and also to each of its parts.

The essential property here is that we can assign the ignorance degree *Ig* to some contingent outcome $A \vee B$ and that same ignorance degree to each of its mutually exclusive, disjunctive parts, A and B. This is the only distribution for which this is true over all contingent outcomes.

The second instrument used to delineate the epistemic state of complete ignorance is the notion of invariance, used so effectively by objective Bayesians, but here used in a way that objective Bayesians may not endorse. The notion ultimately produces serious problems for Bayesian confirmation theory. The greater our ignorance, the more symmetries we have under which the epistemic state should be invariant. It is quite easy to accumulate so many symmetries that the epistemic state cannot be a probability distribution. For example, let us say we know only

$$x \text{ is a real number the interval (0,1).} \qquad \text{(DATUM)}$$

This information (DATUM) remains true if x is replaced by $x' = 1 - x$, where the function $x'(x)$ is self-inverting. It also remains true if x is replaced by $x'' = 1 - (1 - (1 - x)^2)^{1/2}$, where once again the function x"(x) is self-inverting. So our epistemic state must remain unchanged under each of these transformations. It is easy to show that no probability distribution can be invariant under both. (See [Norton, 2008, §3.2].)

Invariance requirements can be used to pick out the unique state of complete ignorance, which turns out to be (I) above. To see the relevant invariance, consider some proposition A over whose truth we may be completely ignorant. Our belief would be unchanged were A replaced by $\sim A$.[24] That is, the state of complete ignorance remains unchanged under a transformation that replaces every contingent proposition with its negation. It can readily be seen that the ignorance distribution (I) satisfies this invariance requirement. Each contingent proposition and its negation are assigned the same degree *Ig*.[25] That this ignorance distribution (I) is the only distribution satisfying this invariance that we are likely to encounter is made more precise by the demonstration that it is the only monotonic[26] distribution of belief with the requisite invariance [Norton, 2008, §6].

[23]Contingent propositions are defined here as propositions that may be either true of false.

[24]Describing this transformation more figuratively makes the invariance intuitive. Imagine that the content of proposition A has been written in a normal English sentence by a scribe on a slip of paper, folded before us on the table. We form our belief over the truth of the sentence on the paper: it is complete ignorance because we have no idea what the sentence says. We are now told that the scribe erred and mistakenly wrote the content of $\sim A$ on the paper instead. That new information would not change our belief state at all.

[25]It is assumed that the ignorance does not extend to logical truths, so that we know which propositions are assuredly true and false and we assign *Max* and *Min* to them. We could define a broader ignorance state in which logical truths are presumed unknown as well by assigning the same value *Ig* to all propositions.

[26]A distribution of belief "degree (.)" is monotonic if, whenever A logically entails B, degree(A) $\leq$ degree(B).

The negation map — the transformation that replaces each contingent proposition with its negation — is a little more complicated than it may initially seem. To see the difficulty, imagine that the outcome space is exhausted by n mutually exclusive atomic propositions, A_1, A_2, ... , A_n. The transformation replaces atomic propositions, such as A_1 by compound propositions, such as $A_2 \vee A_3 \vee \ldots \vee A_n$, so it may not be evident that the transformation is a symmetry of the outcome space. It is, in the sense that it maps the outcome space back to itself and is self-inverting; $A_2 \vee A_3 \vee \ldots \vee A_n$ is mapped to A_1. However, it does not preserve additive measures. The transformation takes an additive measure m to what Norton [2007a] describes as a "dual additive measure" M. These dual additive measures have properties that are, on first acquaintance, odd looking. They are additive, but their additivity is attached to conjunctions. If we have propositions A and B such that $A \vee B$ is always true, then we can add their measures as $M(A\&B) = M(A) + M(B)$. The notion of a dual measure allows a simple characterization of the ignorance distribution (I): it is the unique, monotonic measure that is self-dual.

If one does not see that an epistemic state of complete ignorance is represented by the non-probabilistic (I), one is susceptible to the "inductive disjunctive fallacy" [Norton, forthcoming a]. Let $a_1, a_2, a_3, \ldots$ be a large number of mutually exclusive outcomes over which we are in complete ignorance. According to (I), we remain in that state of complete ignorance for any contingent disjunction of these outcomes, $a_1 \vee a_2 \vee a_3 \vee \ldots$ If one applies probabilities thoughtlessly, one might try to represent the state of complete ignorance by a broadly spread probability distribution over the outcomes. Then the probability of the disjunction can be brought close to unity merely by adding more outcomes. Hence one would infer fallaciously to near certainty for a sufficiently large contingent disjunction of outcomes over which we are individually in complete ignorance. The fallacy is surprisingly widespread. A striking example is supplied by van Inwagen [1996] in answer to the cosmic question "Why is there anything at all?" There is, he asserts, one way for no thing to be, but infinitely many ways for different things to be. Distributing probabilities over these outcomes fairly uniformly, we infer that the disjunction representing the infinitely many ways things can be must attract all the probability mass so that we assign probability one to it.

4.4 Bayesian Responses

The literature in Bayesian confirmation theory has long grappled with this problem of representing ignorance. That is especially so for prior probability distributions, where the presumption that ignorance must be representable is most pressing. Perhaps the most satisfactory response comes through the basic supposition of subjective Bayesians that the probabilities are subjective and may vary from person to person as long as the axioms of the probability calculus are respected. So the necessary deviation from the non-probabilistic ignorance distribution (I) in some agent's prior probability distribution is discounted as an individual aberration not reflecting the true evidential situation. The price paid in adopting this

response, the injection of subjectivity and necessity of every prior being aberrant, is too high a price for objective Bayesians, who are committed to there being one probability distribution appropriate for each circumstance.

However objective Bayesian methods have not proven able to deliver true "ignorance priors," even though the term does appear over-optimistically in the objective Bayesian literature [Jaynes, 2003, Ch.12]. One approach is to identify the ignorance priors by invariance properties. That meets only with limited success, since greater ignorance generates more invariances and, as we saw in Section 4.3 above, eventually there are so many invariances that no probability measure is admissible. An alternative approach is to seek ignorance priors in distributions of maximum entropy.[27] Maximum entropy distributions do supply what are, in an intuitive sense, the most uniform distributions admissible. If, for example, we have an outcome space comprising n atomic propositions, without further constraints, the maximum entropy distribution is the one uniform distribution that assigns probability 1/n to each atomic proposition. However, if there is sufficient ignorance, there will be invariances under which the property of having attainted maximum entropy will not be preserved. In the end, it is inevitable that these methods cannot deliver the ignorance distribution (I), for (I) is not a probability distribution. So the best that can be expected is that they will deliver a distribution that captures ignorance over one aspect of the problem, but not all. The tendency in the literature now is to replace the misleading terminology of "ignorance prior" by more neutral terms such as "noninformative priors," "reference priors" or, most clearly "priors constructed by some formal rule" [Kass and Wasserman, 1996].

Another popular approach to representing ignorance with Bayesian confirmation theory is to allow that an agent's epistemic state is not given by any one probability measure, but by a set of them, possibly convex. (See for example [Levi, 1980, Ch. 9].) The deepest concern with this strategy is that it amounts to an attempt to simulate the failure of additivity that is associated with the representation of ignorance. Something like the complete ignorance state (I) can be simulated, for example, by taking the set of all probability measures over the same outcome space. The resulting structure is vastly more complicated than (I), the state it tries to simulate. It has become non-local in the sense that a single ignorance value is no longer attributed to an outcome. Each of the many probabilities assigned to some outcome must be interpreted in the context of the other values assigned to other propositions and in cognizance that there are many other distributions in the set.[28]

Finally, as pointed out in Norton [2007a; 2008], a set of probability measures necessarily falls short of simulating (I). For no set of additive measures can have the requisite invariance property of a complete ignorance state — invariance under the negation map. For a set of additive measures is transformed by the negation

[27] For further objections to maximum entropy methods, see [Seidenfeld, 1979].

[28] For a discussion of convex sets of probability measures and the how they contain the Shafer Dempster belief functions as a special case, see [Kyburg, 1987].

map into a set of dual additive measures. In informal terms, any set of additive measures on an outcome space preserves a directedness. For each measure in the set, as one proceeds from the assuredly false proposition to the assuredly true by taking disjunctions, the measures assigned are non-decreasing and must, at some point, increase strictly. Invariance under the negation map precludes such directedness.

4.5 Ignorance over a Countable Infinity of Outcomes

The difficulties of representing ignorance have been explored in the literature in some detail in the particular problem of identifying a state of ignorance over a countable infinity of outcomes. It has driven Bayesians to some extreme proposals none of which appear able to handle the problem in its most intractable form.[29]

The traditional starting place — already "well-known" when de Finetti [1972, p.86] outlined it– is to determine our beliefs concerning a natural number "chosen at random." Its more figurative version is "de Finetti's Lottery" [Bartha, 2004] in which a lottery ticket is picked at random from a countable infinity of tickets. If we write the prior probability for numbers $1, 2, 3, \ldots$ as $p_1, p_2, p_3, \ldots$, we cannot reconcile two conditions. First, since we have no preference for any number over any other, we assign the same probability to each number

$$p_i = p_k \text{ all } i, k$$

Second, the sum of all probabilities must be unity

$$p_1 + p_2 + p_3 + \ldots = 1 \qquad (CA)$$

No set of values for p_i can satisfy both conditions.[30]

De Finetti's own solution was to note that the condition (CA), "countable additivity" (as applied to this example), is logically stronger than finite additivity (A). The latter applies only to a finite set of outcomes — say, that the number chosen is a number in $\{1, 2, \ldots, n\}$. It asserts that the probability of this finite set is the finite sum $p_1 + p_2 + p_3 + \ldots + p_n$. Condition (CA) adds the requirement that this relation continues to hold in the limit of n$\rightarrow \infty$. De Finetti asserted that the

[29]The problem of ignorance over a countable infinity of outcomes is actually no worse than the corresponding problem with a continuum outcome space. The latter problem contains the former in that a continuum outcome space can be partitioned into a countable infinity of subsets. Perhaps the countable case is deemed more problematic since outcomes can still be counted so it appears (incorrectly) that there will be a unique, natural probability measure recoverable from ratios of counts. In the continuum case, no such illusions are possible. The continuum case is more problematic in the sense that in it non-zero probabilities cannot be assigned to individual outcomes. Instead a probability density is used to assign probabilities to certain sets of outcomes. That presumes the outcome space has a topology that admits an additive measure. This device of a probability density becomes harder to use as the outcome space becomes larger. If the space consists of all real numbers, then the only uniform probability density is an improper density that cannot be normalized to unity.

[30]Let $p = p_i = p_k$, then $p_1 + p_2 + p_3 + \ldots = 0$ or ∞, according to whether p is zero or non-zero, both of which contradict (CA).

probability distribution representing our epistemic state in this problem need only be finitely additive, not countably additive. That allows us to set $p_i = 0$ for all i, without forcing the probability of infinite sets of outcomes to be zero. So if *odd* $= \{1, 3, 5, \ldots\}$ and *even* $= \{2, 4, 6, \ldots\}$ we can still set $P(odd) = P(even) = 0.5$.

Solving the problem by dropping countable additivity has proven to be a popular and well-understood solution. While proponents of the restriction to finite additivity are typically not frequentists (who identify probabilities with relative frequencies), the connection to frequentism is natural. The frequency of any particular natural number among all is zero and the frequency of even numbers is 1/2, in a naturally defined limit. Kadane and O'Hagan [1995] have mapped out which uniform, finitely additive probability distributions are possible over the natural numbers, noting how these are delimited by natural conditions such as agreement with limiting frequencies and invariance of probability under translation of a set of numbers. There are also variant forms of the proposal, such as the use of Popper functions and the notion of "relative probability" [Bartha, and Johns, 2001, Bartha, 2004]. However dropping countable additivity is not without disadvantages and enthusiasm for it is not universal. There are Dutch book arguments that favor countable additivity; and important limit theorems, including Bayesian convergence of opinion theorems, depend upon countable additivity. See [Williamson, 1999; Howson and Urbach, 2006, pp. 26-29; Kelly, 1996, Ch. 13].[31] Other approaches explore the possibility of assigning infinitesimally small probabilities to what would otherwise be zero probability outcomes [McGee, 1994].

While all these solutions come at some cost, they are eventually unavailing. For they have not addressed the problem in its most acute form. They deal only with the case of ignorance over natural numbers, where this set of a countable infinity of outcomes has a natural order. If we presume that we know of no such natural order, then all these solutions fail, as has been shown in a paradox reported by Bartha [2004, §5.1] and in the work of Arntzenius [manuscript].

Imagine that we have a countable infinity of outcomes with no way to order them. If we have some labeling of the outcomes by natural numbers, that numbering is completely arbitrary.[32] Let us pick some arbitrary labeling of the outcomes, 1, 2, 3, ..., and seek an ignorance distribution over these labels. That ignorance distribution should be unaffected by any one-to-one relabeling of the outcomes; that is, the ignorance distribution is invariant under a permutation of the labels, for permutations are a symmetry of this system. Consider the outcomes *odd* $= \{1, 3, 5, \ldots\}$ and *even* $= \{2, 4, 6, \ldots\}$. There is a permutation that simply

[31] To someone with an interest in physics, where probabilities are routinely summed over infinitely many outcomes, the restriction to finite additivity appears ruinous to ordinary physical theorizing. Countable additivity is hidden in many places. It is presumed whenever we normalize a probability distribution $p(x)$ over some real-valued parameter x. For example $1 = \int_0^1 p(x)dx = \int_0^{1/2} p(x)dx + \int_{1/2}^{3/4} p(x)dx + \int_{3/4}^{7/8} p(x)dx + \ldots$

[32] For a concrete example, imagine that an infinite space is partitioned into a countable infinity of geometrically identical cubes and that our cosmology tells us that a single hydrogen atom will appear in one of them without favoring any one of them. We arbitrarily labeling the cubes as 1, 2, 3, ...

switches the labels of the two sets ($1 \leftrightarrow 2, 3 \leftrightarrow 4, 5 \leftrightarrow 6, \ldots$), so that outcomes in each set are exchanged. Since our belief distribution is invariant under such a permutation, it follows that the permutation does not alter our belief and we must have the same belief in each outcome set *odd* and *even*. Now consider the four outcome sets

$$\begin{aligned} \textit{one} &= \{1, 5, 9, 13, \ldots\} = \{4i + 1 : i = 0, 1, 2, 3, \ldots\} \\ \textit{two} &= \{2, 6, 10, 14, \ldots\} = \{4i + 2 : i = 0, 1, 2, 3, \ldots\} \\ \textit{three} &= \{3, 7, 11, 15, \ldots\} = \{4i + 3 : i = 0, 1, 2, 3, \ldots\} \\ \textit{three} &= \{4, 8, 12, 16, \ldots\} = \{4i + 4 : i = 0, 1, 2, 3, \ldots\} \end{aligned}$$

There is a permutation that switches labels of *one* and *two*; so we have equal belief in *one* and *two*. Proceeding pairwise through the sets, we find we must have equal belief in each of *one*, *two*, *three* and *four*. Now there is also a pairwise permutation that switches the labels of *one* with those of *two* $\cup$ *three* $\cup$ *four*, where

$$\textit{two} \cup \textit{three} \cup \textit{four} = \{2, 3, 4, 6, 7, 8, 10, 11, 12, 14, 15, 16, \ldots\}$$

It is just the obvious permutation read off the above sets

$$1 \leftrightarrow 2, 5 \leftrightarrow 3, 9 \leftrightarrow 4, 13 \leftrightarrow 6, 17 \leftrightarrow 7, 21 \leftrightarrow 8, \ldots$$

So we now infer that we must have the same belief in *one* as in *two* $\cup$ *three* $\cup$ *four*. Combining we find: we must have the same belief in *two* $\cup$ *three* $\cup$ *four* and in each of its disjunctive parts *two*, *three* and *four*. This requirement cannot be met by a probability distribution for it contradicts (finite) additivity.[33] It is however compatible with the ignorance distribution (I).

Finally, it is sometimes remarked that the very idea of uniform ignorance over a countable set is somehow illicit for there is no mechanical contrivance that could select outcomes so that they have equal chances. No lottery commission could build a device that would implement the de Finetti lottery. For references to these concerns, see [Bartha, 2004, pp. 304-305], who correctly objects that the inevitable non-unformity of probabilities of the lottery machine does not force non-uniformity of beliefs. What needs to be added is that the entire objection is based on a circularity. In it, the notion of a mechanical contrivance tacitly supposes a contrivance whose outcomes are governed by a probability distribution. So it amounts to saying that no machine whose outcomes are governed by a probability distribution can generate outcomes governed by a non-probabilistic distribution. Recent work in philosophy of physics has identified idealized physical mechanisms that produce indeterministic outcomes that are *not* governed by a probability distribution. An example is the "dome," described in [Norton, 2007, §8.3; forthcoming], where it is shown that the dome's outcomes are governed by a non-probabilistic distribution

[33] Another possibility is the improper probability distribution that assigns some small probability ε to each, individual outcome and infinite probability to the total outcome space. This improper distribution is badly behaved under conditionalization. For example $P(2|\textit{even}) = 0$, not $\varepsilon/2$; and $P(\textit{two}|\textit{even}) = \infty/\infty =$ undefined, not $1/2$.

with the same structure as (I). There are many more examples of these sorts of indeterministic systems in the "supertask" literature. (For a survey, see [Laraudogoitia, 2004].) Many of these mechanisms conform with Newtonian mechanics, but depend on idealizations some find "unphysical" for their distinctive behavior.[34] These processes could be the physical basis of an idealized contrivance that implements something like the de Finetti lottery.

5 BAYESIAN DYNAMICS

5.1 The Property: Refute and Rescale

The properties investigated so far — framework and additivity — are essential to Bayesian confirmation theory. However we are still missing the essential part of the theory from which its name is derived. For so far, we have no means to relate probability measures conditioned on different propositions, so that we cannot yet relate the posterior probability $P(H|E)$ of some hypothesis H conditioned on evidence E with its prior probability $P(H) = P(H|B)$, for some background B. (Here and henceforth, for notational convenience, "$P(A)$" will be written as shorthand for "$P(A|B)$", for some assumed background B.)

These means for relating prior and posterior probabilities are supplied by two properties. The first property is "narrowness" and asserts for any A and C that

$$P(A\&C|C) = P(A|C) \qquad (N)$$

The second property is "multiplication," which asserts for any A and C that

$$P(A\&C) = P(A\&C|C).P(C) = P(C\&A|A).P(C) \qquad (M)$$

These properties combined entail[35]

$$P(A|C) = P(A\&C)/P(C)$$

when $P(C)$ is not zero. That this formula arises through compounding the two properties (N) and (M) is not generally reported. That compounding is important in the discussion that follows, since these two properties make distinct assertions in the context of confirmation theory and require separate analysis.

These two properties (N) and (M) can also be combined to yield Bayes' theorem. For hypothesis H and evidence E:

$$\begin{aligned} &P(H\&E) = P(H\&E|E).P(E) = P(H|E).P(E) \\ &= P(E\&H) = P(E\&H|H).P(H) = P(E|H).P(H) \end{aligned}$$

[34] For an analysis of the notion of "physical" in this context and the suggestion that we discount these concerns, see [Norton, 2008a].

[35] The combined formula is sometimes regarded as the definition of conditional probability in terms of unconditional probability. That view is not taken here since $P(A)$ is not unconditional, but shorthand for $P(A|B)$. Hájek [2003] has objected that this candidate definition fails to yield an everywhere serviceable notion of conditional probability.

so that when $P(E)$ is not zero we have Bayes' theorem

$$P(H|E) = \frac{P(E|H)}{P(E)} P(H) \qquad (B)$$

It tells us how learning evidence E should lead us to update our beliefs, as expressed in the transition from the prior $P(H)$ to the posterior probability $P(H|E)$.

Bayes' theorem embodies a very simple model of belief dynamics whose two steps essentially correspond to the two properties (N) and (M) above.

> *Refute.* When evidence E is learned, those parts of the hypothesis H logically incompatible with the evidence are discarded as irrelevant to the bearing of evidence. That is, the hypothesis H has two parts in relation to evidence E : $H = (H\&E) \vee (H\& \sim E)$. The second part $(H\& \sim E)$ is that part which is refuted by E. The first part $(H\&E)$ is the part of H that entails E. H can only accrue support through this part, since we have from (N) that $P(H|E) = P(H\&E|E)$.
>
> *Rescale.* The import of evidence E on competing hypotheses $H, H', \ldots$ is then simply the prior probability $P(H\&E), P(H'\&E), \ldots$ of those parts of the hypotheses logically compatible with the evidence, linearly rescaled so as to preserve normalization to unity; for, we have from (B) and (N) that
>
> $$\frac{P(H|E)}{P(H'|E)} = \frac{P(H\&E|E)}{P(H'\&E|E)} = \frac{P(E|H\&E).P(H\&E)}{P(E|H'\&E).P(H'\&E)} = \frac{P(H\&E)}{P(H'\&E)}$$
>
> since $P(E|H\&E) = 1 = P(E|H'\&E)$.

A figurative model of this dynamics underscore its very great simplicity. In his "muddy Venn diagram," Van Fraassen's [1990. p. 161-62] pictures the total outcome space as a surface, such as a table top, and the degrees of belief assigned to outcome sets are represented by volumes of mud piled over the areas corresponding to each outcome. Conditionalization on evidence E occurs in two steps, as shown in Figure 1. First ("*refute*"), all the mud piled outside E is carefully swept away. Second ("*rescale*"), that swept-away mud is redeposited on E in such a way that the proportions in the contours over E are retained.[36]

It is informally evident that this qualitative notion of refute and rescale dynamics captures the essence of Bayesian dynamics. That informal idea has been made more precise in [Norton, 2007, §4], where the dependencies of refute and rescale dynamics are used to generate an axiom system for degrees, which turn out to be ordinary probabilities up to monotonic rescalings.

[36]For example, if the mud is twice as high at one spot in relation to another prior to the redepositing of the mud, it will be piled twice as high after, as well.

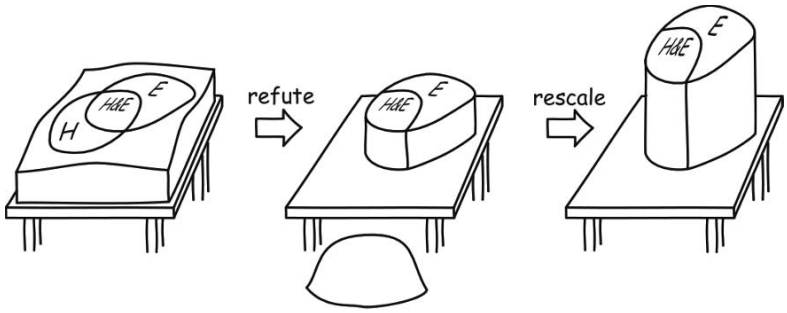

Figure 1. Muddy Venn diagram illustrates rescale and refute dynamics

5.2 Non- Bayesian Shifts

Under refute and rescale dynamics, the import of evidence is always to modify prior beliefs. Earman [1992, pp. 195-198] and Brown [1996] state the obvious difficulty. The dynamics cannot apply if there is no prior belief since the hypothesis or outcome in question is not in the outcome space. Just such a circumstance occurs at decisive moments in the history of science. At moments of rapid change — best known through Thomas Kuhn's notion of a scientific revolution — the inconceivable becomes conceivable and, we are told, that evidence compels us to believe it. When Einstein proposed his special theory of relativity in 1905, he urged that there is no absolute fact as to whether spatially separated events are simultaneous; the simultaneity of events depends upon the inertial reference frame from which they are assessed. That notion was regarded as bizarre by Einstein's critics, many of whom continued to refuse to take it seriously long after 1905. Another example is Bohr's theory of the atom of 1913. That theory seemed to require commitment to an inconsistency: in different processes, accelerated electrons bound in an atom would be supposed to radiate, as classical electrodynamics required, or would be supposed not to radiate, in contradiction with classical electrodynamics, according to which supposition was expedient for the recovery of the results Bohr needed.

Closely connected to the problem of outcomes not in the space are those to which it is natural to assign zero probability. A logical inconsistency, such as Bohr's theory, is a natural candidate. It follows immediately from Bayes' theorem that, once we have assigned a zero prior probability to any hypothesis, conditionalization cannot change its probability from zero. The only recovery is to reassign a non-zero probability to the hypothesis by a process that contradicts Bayes' theorem.

We might discount as rarities these inversions of what is conceivable and the oddity of accepting inconsistencies as established law. However even if they are rarities, they are important rarities that reappear throughout the history of our science and cannot be neglected by a comprehensive account of scientific rationality. The natural Bayesian rejoinder is to protest that too much is being asked of it. The formation of new outcome spaces belongs to Reichenbach's context of discov-

ery. Bayesian confirmation theory, indeed any logic at all, cannot be responsible both for logical relations among propositions and also for the discovery methods used to create new outcome spaces.

5.3 Is Bayes' Theorem Sensitive Enough to the Connections between Evidence and Theory?

The simplicity of refute and rescale dynamics invites the next challenge: the dynamics is too simple and too insensitive to the ways that evidence and theory can relate.

5.3.1 Problems Arising from "Refute. . . ."

The first "refute" step of Bayesian conditionalization depends essentially on the idea that the disjunctive part of $H = (H\&E) \vee (H\& \sim E)$ that extends beyond the total evidence E, that is $H\& \sim E$, does not affect the bearing of evidence E on H. That is, "narrowness" directly asserts that $P(H|E) = P(H\&E|E)$. While this idea is so familiar as generally to pass without comment, it does represent a blindness in Bayesian confirmation theory. As an illustration of it, imagine that we seek to identify some animal. Our evidence is that it is a bird. What support does that evidence lend to the possibility that the animal is a canary or, alternatively, a canary or a whale? Narrowness asserts that it lends the same support:

$$P(\text{canary or whale} \mid \text{bird}) = P(\text{canary} \mid \text{bird})$$

In some scenarios, this can make sense. As we check instances of birds, we will find the frequency of "canary or whale" to arise exactly as often as "canary." If these frequencies are all that matters, then we would say the two outcomes are equally supported. However few would halt there. We would discount the "whale" disjunct as a nuisance embellishment that should be dismissed as a distraction precisely because it is deductively refuted by the evidence "bird." The frequency count just mentioned is blind to its nuisance character in so far as it assigns the same frequency of success to "canary" as to "canary or whale."

It is possible to devise a logic that is sensitive to this problem and punishes an hypothesis in the degree that it extends beyond the evidence. An example[37] is the "specific conditioning" logic defined in [Norton, forthcoming b, Section 10.2]. In a formulation compatible with developments here, the logic amounts to a new definition for the conditional probability $P(H|E)$. It starts with an additive measure

[37]Specific conditioning resembles some of the measures of incremental support that have been investigated in the Bayesian literature, such as the ratio measure $r(H, E) = P(H|E)/P(H) = P(H\&E)/(P(E).P(H))$. (See [Eels and Fitelson, 2002].) The two differ essentially. The incremental measures are functions of three arguments, H, E and B, that represent the incremental support accorded H by evidence E alone with respect to a tacit background B. $P_{SC}(H|E)$ is a function of two arguments, H and $E\&B$, that measures the total support accorded to H by evidence E conjoined with the background B, that is, by the total evidence $E\&B$.

$P(.)$ on the background and defines the alternative

$$P_{SC}(H|E) = P(H\&E)^2/(P(E).P(H)) = P(H\&E)/(P(E).P(H\&E)/(P(H) \quad (SC)$$

This alternative rule of conditionalization rewards hypotheses H with more support the closer they come to the total evidence E = (E and background) — hence the term "specific conditioning." $P_{SC}(H|E) = 1$ only when $H = E$ excepting measure zero differences. H can differ from E in two ways and the logic of (SC) penalizes it for both. In the first way, H can fail to exhaust E in measure. Then the first factor of (SC), that is, $P(H\&E)/(P(E)$, is less than one (which alone would correspond to the usual case of Bayesian conditionalization). In the second way, H can extend beyond E in measure. Then the second factor of (SC), that is $P(H\&E)/(P(H)$, is less than one. Deviations in both ways are penalized equally and it turns out that this is expressed in a striking symmetry in P_{SC}. That is $P_{SC}(H|E) = P_{SC}(E|H)$.

Returning to the bird example, with plausible values for $P_{SC}(.)$, we will have

$$P_{SC}(\text{canary or whale} \mid \text{bird}) < P_{SC}(\text{canary} \mid \text{bird})$$

Indeed we will have

$$P_{SC}(\text{canary or whale} \mid \text{canary}) < P_{SC}(\text{canary} \mid \text{canary}) = 1$$

That is, even though "canary" deductively entails "canary or whale," the latter is not accorded full support from "canary." The logic of (SC) judges "canary or whale" to be supported less specifically by the evidence "canary." In the case of Bayesian confirmation theory, evidence accords unit support to any of its deductive consequences, no matter how much they are weakened logically. P_{SC} reduces the support accorded to these deductive consequences according to how much they are weakened logically.

5.3.2 Problems Arising from "... Rescale"

In the special case in which they hypothesis H deductively entails the evidence E, the first "refute" step of Bayesian conditionalization is not invoked and the process reduces to rescaling only by means of the property "multiplication" (M). This special case provides a means of isolating the second step and plumbing its problems.

Consider two hypotheses H and H' both of which entail the evidence E. In that case, it follows from Bayes' theorem that the ratio of the posterior probabilities $P(H|E)/P(H'|E)$ is the same as the ratio of the priors. It now follows that the incremental confirmation, as measured by the ratio of posterior and prior $P(H|E)/P(H)$, is the same in both cases. That is, when two hypotheses both entail the same true evidence, they get the same confirmatory boost. In particular, if H is more probable than, less probable than or just as probable as H' prior to conditionalization, it will remain so afterwards.

This failure to separate H and H' evidentially, according to critics, is to overlook that their entailments of the evidence E can differ in more subtle qualities that have epistemic import. For example, recall efforts in the late 19th century to detect the earth's motion through the luminiferous ether. These experiments yielded a null result (E). We may entertain two hypotheses:

> H: There is no ether state of rest against which the earth can move.
>
> H': There is an ether state of rest but we happen to be motionless with respect to it each time the experiments are performed.

Each of H and H' entail the evidence E, but it seems wrong that both should receive the same support from E. H succeeds, as it were, by honest toil; but H' by theft. There are many ways that this metaphor of honest toil and theft is explicated.

One way notes that H succeeds because it explains why the experiments yielded null results, where as H' does not. Considerations like this lead Achinstein [2001, Ch. 7] to argue that it is not sufficient to count as evidence that some datum incrementally increases the probability of an hypothesis or that the probability of the hypothesis conditioned on the datum is high.[38] In addition there must be an explanatory connection.

Other approaches are variants of this idea that some successful entailments are virtuous and others not. The hypothesis H makes the prediction that all ether current experiments will fail and that prediction is verified by the evidence. Since H' supposes that it is by happenstance that we were at rest in the ether just at the moment of the experiments, H' simply accommodates the evidence in the sense that it has been adjusted to accommodate experimental reports at hand. Predictions, for example, are virtuous and their success merits an increase in belief; accommodations are not virtuous and are epistemically inert. See [Horwich, 1982, Ch. 5] for discussion and an argument that we should not differentially reward prediction and accommodation. Another way of expressing these concerns is to note that H' (but not H) was cooked up "ad hoc" specifically to match the evidence. A classic example is the creationist hypothesis that the world was created in 4004 BC, complete with a fossil record that perfectly simulates a more ancient, evolutionary past. It is an evidential truism that such ad hoc hypotheses gain no support from the observations they accommodate. For an account of how ad hoc hypotheses can be treated in a Bayesian context, see [Howson and Urbach, 2006, pp. 121-126], who also suggest that sometimes there are "good" ad hoc hypotheses that deserve support.

The literature on these evidential virtues in Bayesian confirmation theory is large and convoluted. In general, however, there are two broad strategies that Bayesian confirmation theory can employ when faced with a virtuous and non-virtuous pair of hypotheses. The first is to expand the outcome space so that there

[38] See [Achinstein, 2001, Ch.7]and *passim*) for numerous examples. For example, learning that most lottery tickets were unsold may increase greatly my belief that mine is the winning ticket, even though the absolute probability may remain small. See also [Laudan, 1997].

are more resources available to pick the hypotheses apart.[39] That risks merely postponing the problem since it may now return for another pair of hypotheses in the larger space. The second is to use prior probabilities to reward virtuous hypotheses and punish non-virtuous ones. That is, since the ratio of the posterior probabilities $P(H|E)/P(H'|E)$ equals the ratio of the priors $P(H)/P(H')$, we may reward the explanatory, predictive or non-ad hoc H by making the ratio $P(H)/P(H')$ very large, so that there is a corresponding advantage for H in the posterior probabilities. As observed in [Norton, 2007, §5.3.3], this use of prior probabilities requires a curious prescience: we are to penalize the prior probability of an ad hoc hypothesis in advance in just the right degree so that the punishment perfectly cancels the as yet unknown confirmatory reward that will be accorded to the hypotheses by Bayes' theorem. Further, prior probabilities for hypotheses or theories are assigned once, globally, whereas virtues are manifested locally and may differ from domain to domain. So we may find a theory explanatory in one domain and want to reward it with a high prior probability; but may find it explanatorily deficient in another, and may want to punish it with a low prior probability.

5.4 The "Likelihood Theory of Evidence" and "Likelihoodism"

The above discussion has dealt with the case of hypotheses that entail the evidence. Matters are slightly more complicated when the hypothesis only makes the evidence more or less probable. In that case, Bayes' theorem allows us to compare the import of evidence E for two hypotheses H and H' as

$$\frac{P(H|E)}{P(H'|E)} = \frac{P(E|H)}{P(E|H')}\frac{P(H)}{P(H')} \qquad (B')$$

It follows that the relative import of the evidence E is fully determined by the two likelihoods, $P(E|H)$ and $P(E|H')$. For the ratio of the prior probabilities $P(H)/P(H')$ reflects our comparative belief in H and H' prior to any consideration of evidence E; the ratio of the posteriors $P(H|E)/P(H|E')$ reflects our comparative belief in H and H' after the import of evidence E has been accommodated. The ratio of the likelihoods $P(E|H)/P(E|H')$ is what takes us from the ratio of the prior to the ratio of the posteriors. Therefore, it expresses the relative import of evidence E on these two hypotheses.

That the likelihoods capture all needed information for evaluating the bearing of evidence is the core notion of a "likelihood theory of evidence." (In the special case in which the hypotheses entail the evidence, $P(E|H) = P(E|H') = 1$. Then the ratio of the posteriors equals the ratio of the priors and E cannot discriminate evidentially between H and H'.)

Since so many of the problems of Bayesian confirmation theory focus on prior probabilities, an attractive modification to the theory is a view that excises prior

[39] For example Maher treats prediction and accommodation by supposing that hypotheses are generated by methods that may or may not have certain items of evidence available to them when they generate the hypotheses. He then shifts assessment to the credibility of the methods. See [Achinstein, 2001, pp. 215-221] for discussion.

probabilities but leaves the rest intact. The view, popularly known as "likelihoodism," has been advocated by Edwards [1972] and Royall [1997]. As we have just seen, in Bayes' theorem (B'), the evidence E favors the hypothesis H over H' just in the ratio of their likelihoods, $P(E|H)/P(E|H')$. The proposal is that this consequence of Bayes' theorem be extracted and elevated to an independent principle, the "law of likelihood." The judgment of evidential import is always comparative — evidence favors this hypothesis more or less than that — and the troublesome prior probabilities never need enter.

The principal difficulty for likelihoodism is that it faces all the problems of the Bayesian refute and rescale dynamics recounted here, but without the resources of prior probabilities to ameliorate them. In particular, likelihoodists must say that evidence entailed by two hypotheses is equally favorable to both hypotheses no matter what their virtues or vices. For likelihoodists have lost the Bayesian mechanism of using prior probabilities to reward simpler hypotheses or ones with anticipated predictive power and to penalize ad hoc or contrived hypotheses. Royall [1997, §1.7] considers and offers likelihoodist rejoinders to such concerns.

5.5 Model Selection, Prediction and Simplicity

While the notion that the likelihoods capture the incremental import of evidence has enjoyed notable successes, its overall history has been troubled. There are recalcitrant cases in which the likelihoods alone are clearly too coarse a measure of the import of evidence. Likelihood rewards only accuracy in fitting data at hand. It has difficulty accommodating our common preference for simpler hypotheses that may be less accurate. In rewarding fit to the data at hand, likelihood may be a weaker guide to the greater truth sought that extends beyond these data; that is, likelihood tends to favor accommodation to present data as opposed to predictive success with future data.

The familiar and long-standing illustration of these difficulties comes in the problem of curve fitting. Let us say that we wish to find the relationship between variables x and y:[40]

$$\langle x_1, y_1\rangle = \langle 0.7, 1.0\rangle, \langle x_2, y_2\rangle = \langle 1.5, 1.8\rangle, \langle x_3, y_3\rangle = \langle 2.1, 2.0\rangle, \ldots, \qquad (DATA)$$

which are shown on the graph of Figure 2 below. Our presumption is that these data were generated by the relation

$$y_i = f(x_i) + \text{ error}_i$$

where error_i is a random error term affecting the i-th pair of $x - y$ values. (It is common to assume that these error terms are normally distributed and independent of one another.) The goal of our analysis is identification of the function

[40] The full data set is $\langle 0.7, 1.0\rangle, \langle 1.5, 1.8\rangle, \langle 2.1, 2.0\rangle, \langle 2.3, 0.6\rangle, \langle 2.6, 1.0\rangle, \langle 3.8, 2.1\rangle, \langle 4.5, 3.1\rangle, \langle 4.7, 6.0\rangle, \langle 5.6, 6.9\rangle, \langle 5.6, 7.7\rangle, \langle 5.8, 4.9\rangle, \langle 6.2, 4.4\rangle, \langle 7.1, 7.7\rangle, \langle 7.6, 6.7\rangle, \langle 8.8, 10.1\rangle, \langle 8.9, 8.2\rangle, \langle 9.1, 8.1\rangle, \langle 9.3, 7.4\rangle$.

$f(x)$. The standard method is to find that function that generates a curve of best fit to the data. For data such as in Figure 2, we would normally seek a straight line. That is, we would assume that the unknown function is linear, so that the data was generated by

$$y_i = A + Bx_i + \text{ error}_i \qquad (LIN)$$

That straight line turns out to be

$$y_i = -0.332 + 0.997x_i + \text{ error}_i \qquad (\text{LIN}_{best})$$

The straight line LIN_{best} is the best fit to the data in the sense that it is the straight line drawn from LIN that makes the DATA most probable. That is, its values of A and B maximize the likelihood $P(\text{DATA} \mid \text{LIN}_{best})$; and $A = -0.332$ and $B = 0.997$ are the "maximum likelihood estimators" of A and B.

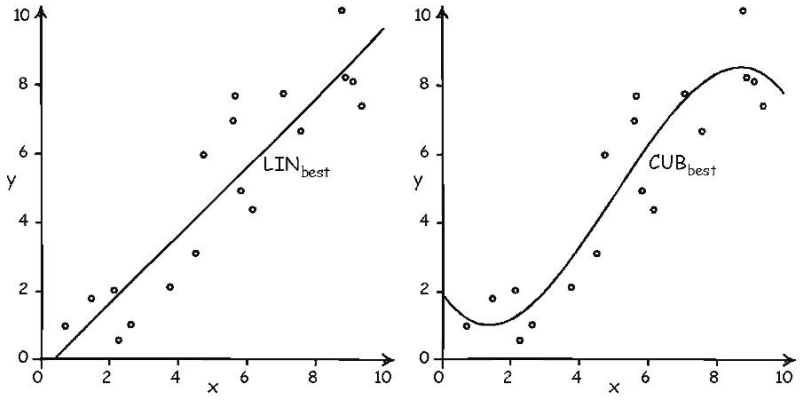

Figure 2. Linear and cubic curves of best fit

The complication is that we can find other functions $f(x)$ that fit the data more closely and make it even more probable. Consider, for example, cubic functions

$$y_i = A + Bx_i + Cx_i^2 + Dx_i^3 + \text{ error}_i \qquad (CUB)$$

The particular cubic plotted in Figure 2 against the background of the same data,

$$y_i = 1.952 - 1.377x_i + 0.581x_i^2 - 0.0389x_i^3 + \text{ error}_i \qquad (\text{CUB}_{best})$$

maximizes the likelihood $P(\text{DATA} \mid \text{CUB}_{best})$.

As one can see from comparing the two curves visually, the cubic CUB_{best} fits the data more closely that the linear LIN_{best} and it turns out that

$$P(\text{DATA}| \text{ CUB}_{best}) > P(\text{DATA} \mid \text{LIN}_{best})$$

Therefore, according to the likelihood theory of evidence, we should conclude that the data better supports the cubic CUB_{best} over the linear LIN_{best}.

It is evident from a cursory scan of the data that this is most likely a mistake. Whatever trend $f(x)$ may be embodied within the data, it is heavily confounded by noise from the error term. While the data definitely gives us warrant to believe that $f(x)$ is an increasing function of x, we have no warrant for the specifics of the cubic CUB_{best}.[41] We should be much more modest in what we infer from data this noisy. The better fit of the cubic CUB_{best} has been achieved by the greater ability of a cubic to conform to random fluctuations from the trend induced by noise; that is, the curve is "overfitted." This overfitting is most evident around $x = 1$, where the curve reverses direction; and again around $x = 9$, where the curve also reverses direction. Given the sparsity and scattering of the data, cubic CUB_{best} is merely tracking noise at these places, whereas the linear LIN_{best} is responding to a real increase with x in the trend.

The informal discussion of the last paragraph captures the practical thinking of curve fitting. It is grounded tacitly in the considerations of simplicity and prediction at issue here. We know that we can always find a curve that fits the data better by looking to more complicated functional forms, such as higher order polynomials. However, at some stage, we judge that we should prefer the simpler curve over the more complicated one that fits the data better. The reason is that our interests extend beyond the narrow problem of finding a curve that merely fits this particular data set well. Our data is only a sample of many more values of x and y; and we are interested in finding a function $f(x)$ that will also fit these other as yet unseen values. That is, our goal is to find a function $f(x)$ that will support prediction. We expect that we can best achieve that goal by forgoing some accuracy of fit with the present data of an overfitted curve in favor of a simpler functional form for $f(x)$ that will enable future data to be fitted better. For an overfitted curve responds to random noise whose specific patterns of deviations will probably not reappear in future data; however a curve drawn from the simpler model responds more to the real relation that drives present and, we suppose, future data as well.

The challenge to a Bayesian analysis is to find a way of capturing these last informal thoughts in a more precise analysis. The natural way to do this is to break up the inference problem into two parts.[42] The first part requires us to consider only models. These are sets of hypotheses indexed by parameters. For example, LIN above is the set of linear functions with parameters A and B taking all real values. CUB above is the set of all cubic functions with the parameters A, B, C and D taking all real values. In the first part we decide which model is

[41] Perhaps it is unfair to introduce a more elevated perspective that affirms this judgment against the cubic CUB_{best}. The data plotted was generated artificially from the model $y_i = x_i + \text{error}_i$, so the linear LIN_{best} has actually done a very good job of identifying the trend $y = x$.

[42] Traditional non-Bayesian statistical methodology implements this two part procedure by testing whether deviations in the coefficients B and C of CUB from zero are statistically significant. If they are not, the null hypothesis of B=C=0 is accepted and estimation proceeds with the simple model LIN.

appropriate to the data. For DATA above, we presume that would turn out to be LIN and not CUB. In the second part of the inference problem, we then find the curve that fits the data best just within that selected model.

Implementing this program within Bayesianism runs into several difficulties. The functions of LIN are a subset of those of CUB, for CUB reverts to LIN if we set $C = D = 0$. So if we compute the posterior probabilities of the models on the DATA, we will always find

$$\text{P(LIN} \mid \text{DATA)} \leq \text{P(CUB} \mid \text{DATA)}$$

from which it follows that that the model LIN can never be more probable on the DATA than the model CUB. So it seems that we cannot have a straightforward Bayesian justification for preferring LIN over CUB.

Slightly less than straightforward grounds can be found. We compare not the posterior probabilities of models, but the boosts in probability each model sustains upon conditionalizing on the data. That is, in this modified approach, we would prefer LIN over CUB if the ratio P(LIN | DATA) / P(LIN) exceeds P(CUB | DATA) / P(CUB). The ratio of these two ratios is itself called the "Bayes' factor."

This modified criterion of the Bayes' factor has some plausibility since it conforms to the spirit of the likelihood theory in that we consider changes in probability under conditionalization, not absolute values of the posterior probability. However the modified criterion rapidly runs into difficulties with prior probabilities. To compute P(LIN | DATA), for example, we would use Bayes' theorem, in which the likelihood P(DATA | LIN) appears. This likelihood is really a compound quantity. It is a summation over the infinitely many curves that appear in the model LIN, corresponding to all possible values of A and B. That is,

$$P(DATA|LIN) = \int_{all A,B} p(DATA|A,B)p(A,B)dAdB$$

The problematic term is $p(A, B)$, which is the prior probability density for the parameters A and B given that LIN is the correct model. Since it expresses our distribution of belief over different values of A and B prior to evidence, $p(A, B)$ should favor no particular values of A or B. However it is familiar problem that no probability distribution can do this. If $p(A, B)$ is any non-zero constant, then it cannot be normalized to unity when we integrate over all possible values of A and B.

While no unqualified solution to these problems has emerged, it has proven possible to show that, for large data sets, the Bayes factor depends only weakly on the choice of the prior probability density. In that case, the model favored by the Bayes factor is the one that maximizes what is known as the Bayes Information Criterion (BIC):

$$\text{BIC} = \log \text{L}_{best} - (\text{k}/2) \log \text{n}$$

where the maximum likelihood L_{best} is the likelihood of the data when conditionalized on the best fitting curve in the model, k is the number of parameters in the model and n is the size of the data set. (For LIN, k is 2; for CUB, k is 4.) For further discussion, see [Wasserman, 2000].

In maximizing BIC, we tradeoff accuracy of fit (expressed in the maximizing of likelihood) against the simplicity of the model (expressed in the number of parameters k). This was not the goal in constructing BIC; it was generated from an analysis that seeks the most probable model on the evidence. An analogous criterion can be generated by directly seeking the model with the greatest predictive power, as opposed to the one that is most probable. This criterion, the Akaike Information Criterion (AIC), advocated by Forster and Sober [1994], is given by

$$\text{AIC} = \log L_{best} - k$$

The rationale is that overtly seeking predictive power is a surer way to get to a truth that extends beyond the particular data at hand. Seeking the most probable model risks favoring the vagaries of the particular data at hand.

The connection to prediction is achieved through the notion of "cross-validation." While we cannot now test the predictive powers of a model against presently unknown data, we can simulate such a test by dividing the data at hand into two parts. We use the first part to generate a best fitting hypothesis in the model and then we check how well the resulting hypothesis fits the remaining data. In a large set of N data, we leave out one datum and use the remaining $N-1$ to generate hypotheses from the model. We then average the errors in fitting the original datum. If we allow each datum successively to be left out in the procedure and average the resulting scores, we recover an estimate of the predictive powers of the model. That estimate turns out to approach the AIC statistic for large N. [Forster, 2002, p. S128; Browne, 2000]. The presumption is that predictive powers manifested within a known data set in this computation will persist beyond the confines of that data set.

It remains an open question whether likelihoods may be used as the basis of assessment of models when we need to find an appropriate accommodation of accuracy, simplicity and predictive power. That all the proposals above fail has been suggested recently by Forster [2006; 2007]. He has devised examples of pairs of models with the same number of parameters k that also deliver the same likelihood and maximum likelihood for a specified set of data. As a result, considerations of likelihood alone or the more refined BIC or AIC are unable to discriminate between the models. However it is evident from informal considerations that one model is predictively superior to the other.

Forster's simplest example[43] draws on a set of data with three points $\langle x_i, y_i \rangle$:

$$\langle 1,1\rangle, \langle 2,2\rangle, \langle 3,3\rangle \qquad \text{(DATA}')$$

[43]This example was presented in a talk by Malcolm Forster, "Is Scientific Reasoning Really that Simple?" in "Confirmation, Induction and Science" London School of Economics, March 8 – March 10, 2007, on March 8.

The first model, $H_a : y_i = ax_i$, comprises four functions

$$H_{1/2} : y_i = (1/2)x_i \quad H_1 : y_i = x_i \quad H_2 : y_i = 2x_i \quad H_3 : y_i = 3x_i$$

corresponding to the parameter values $a = 1/2, 1, 2, 3$. The relation of this model to DATA' is shown in Figure 3.

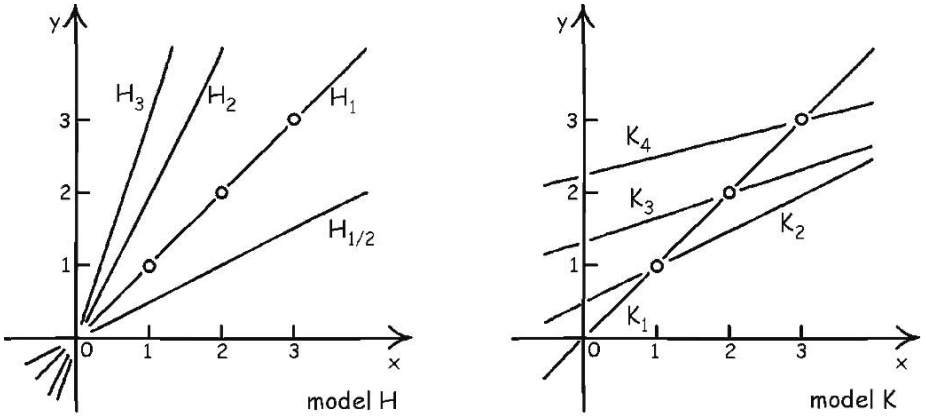

Figure 3. Models H and K

The second model, $K_a : y_i = x_i/a + (a-1)^2/a$, comprises four functions

$$K_1 : y_i = x_i \quad K_2 : y_i = x_i/2 + 1/2 \quad K_3 : y_i = x_i/3 + 4/3 \quad K_4 : y_i = x_i/4 + 9/4$$

corresponding to parameter values $a = 1, 2, 3, 4$. The relation of this model to DATA' is shown in Figure 3.

Which model does DATA' favor? The computation of the likelihoods of the individual functions is easy since they are all zero or one. The only non-zero likelihoods are

$$\mathrm{P(DATA' \mid H_1) = 1 \; P(DATA' \mid K_1) = 1}$$

We assume that the prior probabilities of functions within each model is the same. That is[44]

$$P(H_a|H) = 1/4 \text{ for } a = 1/2, 1, 2, 3$$
$$P(K_a|K) = 1/4 \text{ for } a = 1, 2, 3, 4$$

Thus the likelihood associated with each model is the same

$$P(DATA'|H) = P(DATA'|H_1) \times P(H_1|H) = 1/4$$
$$P(DATA'|K) = P(DATA'|K_1) \times P(H_1|K) = 1/4$$

[44]In these formulae, read $H = H_{1/2} \vee H_1 \vee H_2 \vee H_3$ and $K = K_1 \vee K_2 \vee K_3 \vee K_4$.

Hence, we cannot use DATA' to discriminate between the two models H and K by means of their maximum likelihoods or the likelihoods of the models. Moreover, we cannot discriminate between them by means of BIC or AIC; for both models agree on the maximum likelihood and the number of parameters, $k = 1$.

While all the likelihood based instruments fail to discriminate between the models, informal considerations favor the model H. For it requires only one datum from DATA', such as $\langle 1, 1\rangle$, to fix the function $y = x$ in H that fits the remaining data. The remaining two data points act as tests of the model; or, we may view them as successful predictions. However it takes two data points from DATA' to fix the function in K that fits the remaining data. The datum $\langle 1, 1\rangle$, for example, is compatible with both K_1 and K_2. A second datum is needed to decide between them. Hence only one datum remains to test the function fitted; or the model needs more data before it can make successful predictions. As a result we informally judge that the model H is better tested and predictively stronger and thus better warranted by the evidence. (For discussion of this aspect of curve fitting, see [Glymour, 1980, Ch. VIII.])[45]

For further discussion of Bayesian and Likelihood based analyses of this problem, see [Howson and Urbach, 2006, 288-96] and [Forster and Sober, 2004] (which includes commentary by Michael Kruse and Robert J. Boik).

5.6 A Neutral Starting Point: The Second Problem of the Priors

Bayesian confirmation theory depicts learning from evidence as the successive conditionalization of a prior probability distribution on evidence so that the resulting posterior probability distributions increasingly reflect just the evidence learned. As we trace back along this chain of conditionalization, we expect to come to epistemically more neutral probability distributions, reflecting our lesser knowledge earlier in the process. The problem of the priors in both forms developed here is that the structure of Bayeisan confirmation theory essentially precludes neutral probability distributions. In the first problem of the priors (Section 4.3 above), we saw that the additivity of probability measures precludes their representing epistemic states of complete ignorance. Here we shall see that the Bayesian dynamics of refute and rescale also precludes a different sort of neutrality of the probability distribution. Bayesian dynamics makes it impossible to select a prior probability distribution that does not exercise a controlling influence on subsequent course of conditionalization.[46]

The essential problem is suggested by the notion of refute and rescale dynamics. The results of conditionalization must already be present in the prior probability; all that conditionalization does is to remove those parts of the prior probability

[45]The informal considerations invoked here may appear more familiar if we vary the example slightly. Imagine that we are to choose between the model $H : y = ax$ and $L : y = ax + b$ for DATA$'$. Then we would allow that H is better warranted since a single datum is needed to fix the parameter a in H; whereas two data are needed to fix the two parameters a and b in L.

[46]The phrase "problem of the priors" seems to mean somewhat different to things to different authors. For further discussion, see [Earman, 1992, pp. 57-59; Mellor, 2005, pp. 94-95].

distribution attached to outcomes refuted by the evidence to reveal the pre-existing distribution of belief between the unrefuted outcomes. This same point is evident if we recall the formula governing conditional probabilities

$$P(H|E) = P(H\&E)/P(H)$$

This formula tells us that $P(H|E)$, the degree of belief we should have in H given that we have learned E, is fixed by the two prior probabilities $P(H\&E)$ and $P(H)$. That means that if we fully specify our prior probabilities over some outcome space, we have also delivered an exhaustive catalog of what our belief in each outcome would be, given that we have learned some other outcome. For from being neutral, the prior probability distribution anticipates how our beliefs will evolve as we learn any sequence of admissible outcomes and different prior probability distributions can yield courses that differ greatly. This consideration gives further grounds for the abandoning of such terms as "ignorance priors" or "informationless priors" in favor of "priors constructed by some formal rule" as reported in Section 4.4 above.

As with the first problem, objectivists have the most acute difficulty with this second problem of the priors. For they are committed to there being one, appropriate prior probability distribution in any given epistemic situation. Yet none can supply the neutral starting point appropriate in the absence of all evidence. So the objectivist must either accept the unpalatable conclusion that the initial prior probability contains information without evidential warrant; or that Bayesian dynamics can only be used once we have learned enough by other means to make a properly grounded selection of the prior probability.

Subjectivists at first seem to have a better response. For them, a prior probability is simply a statement of personal opinion with no pretensions of evidential warrant. As evidence accumulates and is incorporated into the probability distributions by conditionalization, limit theorems assure us of a merging of opinion onto a unique distribution that properly represents the bearing of evidence. (For discussion of these theorems, see [Earman, 1992, Ch. 6].) However this solution comes at some cost. At any finite stage of the process, the posterior probability distribution is an unknown mix of unwarranted opinion and warranted support. While the limit theorems may assure us that, in particular circumstances, the mix will eventually converge onto warranted support, at any definite state we may be arbitrarily far from it.[47] Imagine that we have collected some very large set of evidence E_{large} and we are interested in its bearing on some hypothesis H. It will always be possible for a determined mischief maker to identify some prior probabilities for $P(H\&E_{large})$ and $P(E_{large})$ so that $P(H|E_{large})$ is as close to one as you like or as close to zero as you like. The two differ essentially. The incremental measures are functions of three arguments, H, E and B, that represent the incremental support accorded H by evidence E alone with respect to a tacit background B. $P_{SC}(H|E)$ is a function of two arguments, H and $E\&B$, that measures

[47] Here is the obligatory reporting of Keynes' quip: "In the long run we are all dead."

the total support accorded to H by evidence E conjoined with the background B, that is, by the total evidence $E\&B$.

5.7 The Prior Probability of Universal Generalizations

Neither objectivist nor subjectivist fares well if zero or unit prior probability has been assigned injudiciously to some outcome. Then Bayesian dynamics becomes completely dogmatic. For once they have been assigned zero or unit probability, it follows from Bayes' theorem that these assignments are unrevisable by conditionalization.[48] The complete disbelief of a zero prior probability and the complete belief of a unit probability will persist indefinitely. In concrete terms, this sort of dogmatism is a familiar if unwelcome phenomenon. Consider the dogmatic conspiracy theorist who discounts the benign explanation of a catastrophe in favor of an elaborate conspiracy committed by some secret agency. Once a zero prior probability has been assigned to the benign explanation, no failure to reveal the secret agency's intervention or even existence can restore belief in the benign explanation. Rather Bayes' theorem will provide mounting assurance that each failure is further evidence of the perfection of the secret agency's cover-up.

While the best Bayesian advice would seem to be very cautious before assigning a zero prior probability, there are suggestions, reviewed in more detail in [Earman, 1992, §4.2, 4.3], that hypotheses with the power of universal generalizations ought to be assigned a zero prior probability. These concerns arise for any hypotheses H that entails a countable infinity of consequences $E_1, E_2, \ldots$ [49] Considering only the first n consequences we have

$$\begin{aligned} P(H) &= P(H|E_1\&\ldots\&E_n).P(E_n|E_1\&\ldots\&E_{n-1}).P(E_{n-1}|E_1\&\ldots\&E_{n-2}). \\ &\quad \ldots P(E_2|E_1).P(E_1) \end{aligned}$$

Popper [1959, Appendix VII] urged that we should expect the instances $E_1, E_2, \ldots$ of a universal generalization H to be equiprobable and independent of one another. That is,

$$\begin{aligned} &P(E_n|E_1\&\ldots\&E_{n-1}).P(E_{n-1}|E_1\&\ldots\&E_{n-2})\ldots.P(E_2|E_1).P(E_1) \\ &= P(E_n).P(E_{n-1})\ldots.P(E_2).P(E_1) = P(E_1)^n \end{aligned}$$

But since this must hold for arbitrarily large n and with $P(E_1) < 1$, it follows that the prior $P(H) = 0$. The obvious weakness of Popper's proposal is that the very fact that E_1, E_2, ... are instances of a universal generalization H suggests that they are *not* independent. However, just a little dependence between $E_1, E_2, \ldots$ is not enough to allow a non-zero prior for H. Inspection of the above expression for P(H) shows that $P(H) > 0$ entails

[48] Once $P(H) = 0$, Bayes' theorem (B) requires $P(H|E) = 0$ for all admissible E. (Conditionalizing on evidence E for which $P(E) = 0$ leads to an undefined posterior $P(H|E)$.) It follows that for any $H' = \sim H$ for which $P(H') = 1, P(H'|E) = 1$ as well.

[49] The simplest example is a universal generalization that asserts, "For all $x, Q(x)$." where x ranges over a countably infinite set of individuals $a, b, c, \ldots$ Its consequences are $Q(a), Q(b), \ldots$

$$\mathrm{Lim}_{n\to\infty} P(E_n|E_1\&\ldots\&E_{n-1}) = 1$$

That is, finitely many favorable instances of H eventually make us arbitrarily sure that the next instance will be favorable. That was an outcome that Jeffrey [1983, p. 194] felt sufficiently akin to "jumping to conclusions" to want to renounce non-zero prior probabilities on universal hypotheses.

Lest this commitment to the projectability of the hypothesis not seem immodest, the analysis needs only a slight modification to make it more striking. Group the instances $E_1, E_2, \ldots$ into sets that grow very rapidly in size. For example

$$F_1 = E_1 \quad F_2 = E_2\&\ldots\&E_{10} \quad F_3 = E_{11}\&\ldots\&E_{100}\ldots$$

so that

$$F_1\&\ldots\&F_n = E_1\&\ldots\&E_{(10^{n-1})}$$

Each of the F_i is a consequence of H, so the above analysis can be repeated with F_i substituted for E_i to infer that $p(H) > 0$ entails

$$\mathrm{Lim}_{n\to\infty} P(E_1\&\ldots\&E_{(10^{n+1})}|E_1\&\ldots\&E_{(10^n)}) = 1$$

That is, if we assign non-zero prior probability to an hypotheses with the power of a universal generalization, eventually we are willing to project from some finite set of its positive instances to arbitrarily many more at arbitrarily high probability.

The point is *not* that there is something intrinsically wrong with such immodesty. Indeed for the right hypothesis, this may be just the appropriate epistemic attitude. The point is that assigning a zero or a non-zero prior probability to a hypothesis with the power of a universal generalization is never neutral. The former commits us dogmatically never to learning the hypothesis; the second, no matter how small the non-zero prior, commits us to its arbitrary projectability on finite favorable evidence. There is no intermediate value to assign that leaves our options open. Our inductive course is set once the prior is assigned.

Finally, one sees that some significant prior knowledge is needed if we are to assign the non-zero prior probabilities prudently. For a countably infinite set of outcomes $E_1, E_2, \ldots$ corresponds to an uncountable infinity of universal hypotheses: each hypothesis corresponds to one of the uncountably many combinations of the E_i and their negations $\sim E_i$. However we can assign non-zero prior probability to at most countably many of these hypotheses. That is, loosely speaking, when we assign our priors, we are selecting in advance the measure zero subset that we will be prepared to learn and dogmatically banishing the rest to disbelief no matter what evidence may accrue.

5.8 Uncertain Evidence

The versions of Bayesian confirmation theory considered so far depict learning as an absolute: if we are given evidence E, it cannot be doubted. That is quite unrealistic. No datum is ever certain and it is not uncommon in science that the accumulation of later evidence leads one to doubt the correctness of evidence collected

earlier. Something like this could be incorporated into traditional Bayesianism if we separate, say, the inerrant sense datum D: "my retina senses a plesiosaur like shape on the surface of Loch Ness" from the fallible evidence E: "I saw a plesiosaur on Loch Ness." That fallibility could then enter the analysis through the conditional probability $P(E|D)$. This solution complicates the analysis without ultimately allowing that all evidence is fallible, for it merely pushes the inerrancy deeper into an elaborated outcome space.

A direct solution has been proposed by Jeffrey [1983, Ch. 11]. Standard Bayesianism introduced evidence E by by conditionalizing, which forces unit probability onto E. Jeffrey instead proposed that we merely shift a lesser amount of probability to E, commensurate with our confidence in it. This non-Bayesian shift from initial probability P_i to final probability P_f is then propagated through the outcome space by Jeffrey's rule. For this simplest case of evidence that partitions the outcome space into $\{E, \sim E\}$, the rule asserts that, for each proposition A,

$$P_f(H) = P_i(H|E).P_f(E) + P_i(H| \sim E).P_f(\sim E)$$

The distinctive property of this rule is that it leaves unaffected our judgments what the probability of H would be were E definitely true or definitely false. That is, for all H, it satisfies $P_i(H|E) = P_f(H|E)$ and $P_i(H| \sim E) = P_f(H| \sim E)$. The uniqueness of Jeffrey's rule follows in that it proves to be the only rule that satisfies these two conditions [Jeffrey, 1983, p. 169; Howson and Urbach, 2006, p. 85]. Jeffrey's rule reduces to ordinary conditionalization if the evidence E is certain, $P_f(E) = 1$.

One unappealing aspect of Jeffrey's rule is that the order in which it is applied in successive applications matters. That is, if we apply it first on a partition $\{E, \sim E\}$ and then $\{F, \sim F\}$, in general we do not arrive at the same final probability distribution as when the order is reversed. (See [Diaconis and Zabell, 1982].)

In the context of the Shafer-Dempster theory (Section 4.2 above), Dempster's rule of combination allows uncertain evidence to be combined with other belief distributions in an analogous way, but one that is independent of the order of combining. With appropriate restrictions, Dempster's rule of combination can yield ordinary Bayesian conditionalization as a special case. See [Diaconis and Zabell, 1986] for discussion of the relationship between Jeffrey's rule and Dempster's rules.

6 FURTHER CHALLENGES

This concluding section collects mention of some of the many further challenges to Bayesian confirmation theory that can be found in the literature.

In a letter to *Nature* and a subsequent article, Popper and Miller [1983, 1987] urged that the probabilistic support evidence E provides an hypothesis H in Bayesian confirmation theory is not inductive support. Their argument depended upon decomposing H into conjunctive parts by the logical identity

$$H = (H \leftarrow E)\&(H \vee E)$$

where the first conjunct, "H, if E," is $(H \leftarrow E) = (H \vee \sim E)$. Since the second conjunct $(H \vee E)$ is deductively entailed by E, they identify the first, $(H \leftarrow E)$, as "containing all of [H] that goes beyond [E]" [1983, p. 687] and require that E supplies inductive support for H only in so far as it supports this part that is not deductively entailed by E. However it is easy to show that $P(H \leftarrow E|E) \leq P(H \leftarrow E)$ so that $(H \leftarrow E)$ never accrues probabilistic support from E.

While Popper and Miller's argument drew much discussion (for a survey see [Earman, 1992, pp. 95-98]), the decisive flaw was already identified immediately by Jeffrey [1984], who objected that $(H \leftarrow E)$ is not the part of H that goes beyond E. It rests on a confusion of "part" with "deductive consequence." While we might think of theorems as logical parts of the axioms that entail them, that sense of part no longer works when inductive relations are the concern. For example, let H be that my lost keys are in my office ("office") or in my bedroom ("bedroom") and E is the evidence that they are in my house. Then $(H \leftarrow E)$ is the enormous disjunction

$$\text{office} \vee \text{bedroom} \vee \text{Mars} \vee \text{Jupiter} \vee \text{Saturn} \vee \ldots$$

where the ellipses include all places not in my house. It is as peculiar to call this disjunction a part of the original hypothesis as it is unsurprising that it is not probabilistically confirmed by the evidence.

Another challenge to Bayesian confirmation theory comes from what first appears as an unlikely source. Over recent decades, there has been a mounting body of evidence that people do not assess uncertainty in accord with the strictures of probability theory. The most famous example comes from Kahneman and Tversky's [1982] experiments that showed that people are quite ready to contradict the monotonicity of a probability measure — the property that an outcome is never less probable that each of its logical consequences. In the case of "Linda," they found that people would affirm that a suitably described character Linda is more likely to be a bank teller and feminist than a consequence of this possibility, that she is a bank teller. One response to this body of work is to lament the unreliability of people's intuitions about uncertainty and the need for them to receive training in the probability calculus. Another response notes that we humans have been dealing more of less successfully with uncertainty for a long time; this success extends to much of science, where uncertainties are rarely treated by systematic Bayesian methods. So perhaps we should seek another approach based a little more closely on what people naturally do. Cases like "Linda" suggest that we naturally try to select the one conclusion that best balances the risk of error against informativeness and detach it from the evidence. That is quite compatible with preferring to detach less probable outcomes. When told a coin was tossed, most people would infer to the outcome "heads or tails" as opposed to "heads or tails or on edge" even though the latter is more probable and a certainty.

In philosophy of science, if a numerical calculus is used to assess the bearing of evidence, it is most often the probability calculus. In other domains, that is less so. In the artificial intelligence literature, alternative calculi proliferate. They include the Shafer-Dempster theory, possibility theory [Zadeh, 1978] as well as numerical systems generated for particular contexts, such as the MYCIN system applied in medical diagnostics [Shortliffe and Buchanan, 1985]. One motivating factor has been pragmatic. A probabilistic analysis requires a prohibitively large number of quantities to be fixed for the general case. In an outcome space with n atomic propositions, there are of the order of n! conditional probabilities. Seen from that perspective, possibility theory's simple rule for disjunction is appealing: the possibility of AvB is just the larger of the possibilities of A or B individually. That same rule would seem disastrous to the systematic philosopher of science schooled with Bayesian intuitions, who will wonder why the rule does not distinguish the cases of A and B coincident or mutually exclusive and all those in between. The debate between proponents of the different calculi, including probabilistic systems, has been fierce. The principle goal seems to be to establish that each proponent's system has the same or greater expressive power than another's. (See for example [Cheeseman, 1986; Zadeh, 1986].)

The issue of "stopping rules" has entered the philosophy of science literature from the debate between Bayesian and Neyman-Pearson statisticians. It pertains to a deceptively simple plan for assuring belief in a false hypothesis. Assume we have a fair coin. On average, on repeated tosses, it will show heads with a frequency of about 0.5. The actual frequency will fluctuate from this expected 0.5 and there will always be some definite but small probability that the observed frequency is arbitrarily far removed from the expected. In order to advance the misperception that the coin is not fair, we toss it repeatedly, waiting mischievously for chance to realize one of these rare removals that is possible but very unlikely for a fair coin. That is our "stopping rule." When the rare removal happens we report the resulting count of heads and tails as evidence that the coin is not fair. The plan can be generalized to seek disconfirmation of any statistical hypothesis. We accumulate observations, waiting for chance fluctuations to deliver results that can only arise with low probability if the hypothesis is true.

The result is troubling for Bayesians. For them, the import of evidence is fully determined by the likelihood ratio and that likelihood ratio has been contrived to favor arbitrarily strongly the falsity of the hypothesis. That the experimenter intended to generate a deceptive result is irrelevant; the intentions of the experimenter collecting the data do not appear in Bayes' theorem. Neyman-Pearson statisticians have a ready escape that is not available to the Bayeisan: the strategy of "try and try again until you get a statistically significant result" is not a good test of the hypothesis. Correspondingly, the procedure that uses this stopping rule does not count as a severe test of the hypothesis in the sense of Mayo [1996, §11.3, 11.4].

The Bayesian response has been to reaffirm that the likelihoods are all that is relevant to discerning the bearing of evidence and also to point out that the

stopping rule may never be satisfied, so that the planned disconfirmation can fail. That is, assume that a Bayesian agent assigns a prior probability p to some hypothesis and decides to accumulate evidence concerning the hypothesis until the posterior probability has risen to $q > p$. Kadane, Schervisch and Seidenfeld [1996] show that, if the hypothesis is false, the agent's probability that the stopping rule is eventually satisfied is given by

$$\text{Probability of stopping} \leq \frac{p(1-q)}{q(1-p)}$$

Therefore, the closer the desired posterior probability q is set to one, the closer the probability of stopping comes to zero; that is, the probability as assessed by the agent approaches one that the experiment will continue indefinitely without returning the result sought in the stopping rule.[50] For more discussion see see [Savage *et al.*, 1962, pp. 17-20 and subsequence discussion; Mayo, 1996, §11.3, 11.4; and Kadane *et al.*, 1996].

7 CONCLUSION

This chapter has reviewed many challenges to Bayesian confirmation theory. Some are global, reflecting differences in the very conception of inductive inference and the inductive enterprise in science. Others are more narrowly targeted to specific commitments of Bayesian confirmation theory. That there are so many challenges is a reflection of the dominance that Bayesian confirmation theory now enjoys in philosophy of science. It offers a well-articulated account of inductive inference in science that is of unrivalled precision, scope and power. In this regard it greatly outshines its competitors. Abductive accounts — inference to the best explanation — are still stalled by the demand for a precise account of their central notion, explanation, whereas Bayesian confirmation theory suffers from the reverse: excessive elucidation of its central notion, probability. This precision and depth of articulation is part of what enables the challenges to Bayesian confirmation theory to be mounted. For it is only after a position is clearly articulated that its weaknesses may also be discerned. When a new Goliath enters the arena, many would be Davids come with their pebbles.

ACKNOWLEDGEMENTS

I am grateful to Malcolm Forster for helpful discussion on an earlier draft.

[50] If a tenfold boost in probability is sought, so $q/p = 10$, we must have $p \leq 0.1$ to start with and the formula assures us that the probability of stopping is less than $(p/q - p)/(1 - p) = (0.1 - p)/(1 - p) < 0.1$.

BIBLIOGRAPHY

[Achinstein, 2001] P. Achinstein. *The Book of Evidence.* New York: Oxford University Press, 2001.

[Aczel, 1966] J. Aczel. *Lectures on Functional Equations and their Applications.* New York: Academic Press, 1966.

[Artzenius, manuscript] F. Arntzenius. Goodman's and Dorr's Riddles of Induction.

[Bacchus *et al.*, 1990] F. Bacchus, H. E. Kyburg, Jr., and M. Thalos. Against Conditionalization, *Synthese*, **85**, pp. 475-506, 1990.

[Bartha, 2004] P. Bartha. Countable Additivity and the de Finetti Lottery, *British Journal for the Philosophy of Science*, **55**, pp. 301-321, 2004.

[Bartha and Johns, 2001] P. Bartha and R. Johns. Probability and Symmetry, *Philosophy of Science*, **68** (Proceedings), pp. S109-S122, 2001.

[Blake *et al.*, 1960] R. M. Blake, C. J. Ducasse, and E. H. Madden. *Theories of Scientific Method: The Renaissance Through the Nineteenth Century.* E. H. Madden, ed. Seattle: University of Washington Press, 1960.

[Brown, 1994] H. Brown. Reason, Judgment and Bayes's Law, *Philosophy of Science*, **61**, pp. 351-69, 1994.

[Browne, 2000] M. W. Browne. Cross-Validation Methods, *Journal of Mathematical Psychology* **44**, pp. 108-132, 2000.

[Carnap, 1950] R. Carnap. *Logical Foundations of Probability.* Chicago: University of Chicago Press, 1950.

[Cheesman, 1986] P. Cheeseman. Probabilistic versus Fuzzy Reasoning, pp. 85-102 in L. N. Kanal and J. F. Lemmer, eds., *Uncertainty in Artificial Intelligence: Machine Intelligence and Pattern Recognition. Volume 4.* Amsterdam: North Holland, 1986.

[Cox, 1961] R. T. Cox. *The Algebra of Probable Inference,* Baltimore: Johns Hopkins Press, 1961.

[Diaconis and Zabell, 1982] P. Diaconis and S. Zabell. Updating Subjective Probability *Journal of the American Statistical Association.* **77**, pp. 822-30, 1982.

[Diaconis and Zabell, 1986] P. Diaconis and S. Zabell. Some Alternatives to Bayes's Rule, pp. 25-38 in *Information and Group Decision Making: Proceedings of the Second University of California, Irvine, Conference on Political Economy.* B. Grofman and G. Owen, eds. Greenwich, Conn,: JAI Press, 1986.

[Duhem, 1969] P. Duhem. *To save the phenomena, an essay on the idea of physical theory from Plato to Galileo.* Chicago, University of Chicago Press, 1969.

[Earman, 1992] J. Earman. *Bayes or Bust.* Cambridge, MA: Bradford-MIT, 1992.

[Earman and Salmon, 1992] J. Earman and W. Salmon. The Confirmation of Scientific Hypotheses, ch. 2 in *Introduction to the Philosophy of Science.* M. H. Salmon, et al. Prentice-Hall, 1992; repr. Indianapoli: Hackett, 1999.

[Eells and Fitelson, 2002] E. Eells and B. Fitelson. Symmetries and Asymmetries in Evidential Support, *Philosophical Studies,* **107**, pp. 129-142, 2002.

[Edwards, 1972] A. W. F. Edwards. *Likelihood.* London: Cambridge University Press, 1972.

[Ellsberg, 1961] D. Ellsberg. Risk, Ambiguity, and the Savage Axioms, *Quarterly Journal of Economics*, **75**, pp. 643-69, 1961.

[Feinstein, 1977] A. R. Feinstein. Clinical Biostatistics XXXIX. The Haze of Bayes, the Aerial Palaces of Decision Analysis, and the Computerized Ouija Board, *Clinical Pharmacology and Therapeutics* **21**, No. 4, pp. 482-496, 1977.

[de Finetti, 1937] B. de Finetti. Foresight: Its Logical Laws, Its Subjective Sources, trans. H. E. Kyburg, in H. E. Kyburg and H. Smokler (*eds*), *Studies in Subjective Probability*, New York: John Wiley & Sons, 1964, pp. 93-158.

[Fine, 1973] T. L. Fine. *Theories of Probability.* New York: Academic Press, 1973.

[Fishburn, 1986] P. C. Fishburn. The Axioms of Subjective Probability, *Statistical Science*, **1**, pp. 335-58, 1986.

[Forster, 2002] M. Forster. Predictive Accuracy as Achievable Goal of Science, *Philosophy of Science*, **69**, pp. S124-34, 2002.

[Forster, 2006] M. Forster. Counterexamples to a Likelihood Theory of Evidence, *Minds and Machines*, **16**, 319–338, 2006.

[Forster, 2007] M. Forster. A Philosopher's Guide to Empirical Success, *Philosophy of Science*, 74, pp. 588–600, 2007.

[Forster and Sober, 1994] M. Forster and E. Sober. How to Tell when Simpler, More Unified, or Less *Ad Hoc* Theories will Provide, *British Journal for the Philosophy of Science*, **45**, pp. 1-35, 1994.

[Forster and Sober, 2004] M. Forster and E. Sober. Why Likelihood? pp. 153-190 in M. L. Taper and S. R Lee, *The Nature of Scientific Evidence: Statistical, Philosophical, and Empirical Considerations*. Chicago and London: University of Chicago Press, 2004.

[Galavotti, 2005] M. C. Galavotti. *Philosophical Introduction to Probability*. Stanford: CSLI Publications, 2005.

[Gillies, 2000] D. Gillies. *Philosophical Theories of Probability*. London: Routledge, 2000.

[Glymour, 1980] C. Glymour. *Theory and Evidence*. Princeton: Princeton University Press, 1980.

[Hájek, 2003] A. Hájek. What Conditional Probability Could Not Be, *Synthese*, **137**, pp. 273-323, 2003.

[Hájek, 2008] A. Hájek. Arguments for — or against — Probabilism? *British Journal for the Philosophy of Science*, **59**, pp. 793-819, 2008.

[Hempel, 1945] C. G. Hempel. Studies in the Logic of Confirmation, *Mind*, 54, pp. 1-26, 97-121, 1945; reprinted with changes, comments and Postscript (1964) in Carl G. Hempel, *Aspects of Scientific Explanation*. New York: Free Press, 1965, Ch.1.

[Horwich, 1982] P. Horwich. *Probability and Evidence*. Cambridge: Cambridge University Press, 1982.

[Howson, 1997] C. Howson. A Logic of Induction, *Philosophy of Science*, **64**, pp.268-90, 1997.

[Howson and Urbach, 2006] C. Howson and P. Urbach. *Scientific Reasoning: The Bayesian Approach*. 3rd ed. La Salle, Illinois: Open Court, 2006.

[Humphreys, 1985] P. Humphreys. Why Propensities Cannot be Probabilities, *Philosophical Review*, 94, pp. 557-570, 1985.

[Jaynes, 2003] E. T. Jaynes. *Probability Theory: The Logic of Science,* Cambridge: Cambridge University Press, 2003.

[Jeffrey, 1983] R. Jeffrey. *The Logic of Decision*. 2nd ed. Chicago: University of Chicago Press, 1983.

[Jeffreys, 1961] H. Jeffreys. *Theory of Probability*. 3rd ed. Oxford: Clarendon Press, 1961.

[Joyce, 1998] J. Joyce. A Nonpragmatic Vindication of Probabilism, *Philosophy of Science,* **65**, pp. 575-603, 1998.

[Kadane and O'Hagan, 1995] J. B. Kadane and A. O'Hagan. Using Finitely Additive Probability: Uniform Distributions on Natural Numbers, *Journal of the American Statistical Association*. **90**, pp. 626-31, 1995.

[Kadane *et al.*, 1996] J. B. Kadane, M. J. Schervish, and T. Seidenfeld. Reasoning to a Foregone Conclusion, *Journal of the American Statistical Association*, **91**, pp. 1228-36, 196.

[Kass and Wasserman, 1996] R. E. Kass and L. Wasserman. The Selection of Prior Distributions by Formal Rules, *Journal of the American Statistical Association*, **91**, pp. 1343-70, 1996.

[Kaplan, 1998] M. Kaplan. *Decision Theory as Philosophy*. Cambridge: Cambridge University Press, 1998.

[Kelly, 1996] K. T. Kelly. *The Logic of Reliable Inquiry*. New York: Oxford, 1996.

[Kelly and Glymour, 2004] K. T. Kelly and C. Glymour. Why Probability Does not Capture the Logic of Justification, in C. Hitchcock, ed., *Contemporary Debates in Philosophy of Science*.Malden, MA: Blackwell, 2004.

[Keynes, 1921] J. M. Keynes. *A Treatise of Probability*. London: Macmillan, 1921; repr. New York, AMS Press, 1979.

[Kyburg, 1959] H. E. Kyburg. Probability and Randomness I and II, *Journal of Symbolic Logic*, **24**, pp. 316-18, 1949.

[Kyburg, 1987] H. E. Kyburg. Bayesisan and non-Bayesisan Evidential Updating, *Artificial Intelligence*, **31**, pp. 271-93, 1987.

[Kyburg and Teng, 2001] H. E. Kyburg and C. M. Teng. *Uncertain Inference*. New York: Cambridge University Press, 2001.

[Laraudogoitia, 2004] J. P. Laraudogoitia. Supertasks, *The Stanford Encyclopedia of Philosophy* (Winter 2004 Edition), Edward N. Zalta (ed.),2004. `$http://plato.stanford.edu/archives/win2004/entries/spacetime-supertasks/`.

[Laudan, 1997] L. Laudan. How About Bust? Factoring Explanatory Power Back into Theory Evaluation, *Philosophy of Science*, **64**, pp. 306-16, 1997.
[Levi, 1974] I. Levi. On Indeterminate Probabilities, *Journal of Philosophy*, **71,** pp. 391-418, 1974.
[Levi, 1980] I. Levi. *The Enterprise of Knowledge.* Cambridge, MA. MIT Press, 1980.
[Lewis, 1980] D. Lewis. A Subjectivist's Guide to Objective Chance, in R.C. Jeffrey (*ed.*), *Studies in Inductive Logic and Probability.* Berkeley: University of California Press, pp. 263-93, 1980.
[Mayo, 1996] D. Mayo. *Error and the Growth of Knowledge.* Chicago: University of Chicago Press, 1996.
[Mayo, manuscript] D. Mayo. Three Battle Waves in this Philosophy of Statistics.
[McGee, 1994] V. Mc Gee. Learning the Impossible, pp. 179-199 in E. Eels and B. Skyrms, eds., *Probability and Conditionals: Belief Revision and Probability Conditions.* Cambridge: Cambridge University Press, 1994.
[Mellor, 2005] D. H. Mellor. *Probability: A Philosophical Introduction.* Abingdon: Routledge, 2005.
[Mill, 1891] J. S. Mill. *A System of Logic: Ratiocinative and Inductive.* 8th ed. Honolulu, Hawaii: University Press of the Pacific, 1892/2002.
[Norton, 2003] J. D. Norton. A Material Theory of Induction *Philosophy of Science*, **70**, pp. 647-70, 2003.
[Norton, 2005] J. D. Norton. A Little Survey of Induction, in P. Achinstein, ed., *Scientific Evidence: Philosophical Theories and Applications.* Johns Hopkins University Press. pp. 9-34, 2005.
[Norton, 2007] J. D. Norton. Probability Disassembled, *British Journal for the Philosophy of Science*, **58**, pp. 141-171, 2007.
[Norton, 2007 a] J. D. Norton. Disbelief as the Dual of Belief, *International Studies in the Philosophy of Science*, 21, pp. 231–252, 2007.
[Norton, 2008] J. D. Norton. Ignorance and Indifference. *Philosophy of Science*, 75, pp. 45–68, 2008.
[Norton, 2008 a] J. D. Norton. The Dome: An Unexpectedly Simple Failure of Determinism, *Philosophy of Science*, 74, pp. 786–98, 2008.
[Norton, forthcoming] J. D. Norton. There are No Universal Rules for Induction, *Philosophy of Science*, forthcoming.
[Norton, forthcoming a] J. D. Norton. Cosmic Confusions: Not Supporting versus Supporting Not-, *Philosophy of Science*, forthcoming.
[Norton, forthcoming b] J. D. Norton. Deductively Definable Logics of Induction, *Journal of Philosophical Logic*, forthcoming.
[Popper, 1959] K. R. Popper. *Logic of Scientific Discovery.* London: Hutchinson, 1959.
[Popper and Miller, 1983] K. R. Popper and D. Miller, David (1983) A Proof of the Impossibility of Inductive Logic, *Nature*, **302**, pp. 687-88, 1983.
[Popper and Miller, 1987] K. R. Popper and D. Miller. Why Probabilistic Support is not Inductive, *Philosophical Transactions of the Royal Society of London, Series A, Mathematical and Physical Sciences.* **321**, pp. 569-91, 1987.
[Royall, 1997] R. M. Royall. *Statistical Evidence: A Likelihood Paradigm.* London: Chapman & Hall, 1997.
[Salmon, 1966] W. Salmon. *The Foundation of Scientific Inference.* Pittsburgh: University of Pittsburgh Press, 1966.
[Savage, 1972] L. J. Savage. *The Foundations of Statistics,* Second revised ed. New York: Dover, 1972.
[Savage *et al.*, 1962] L. J. Savage *et al. The Foundations of Statistical Inference.* London: Methuen, 1962.
[Schick, 1986] F. Schick. Dutch Bookies and Money Pumps,' *Journal of Philosophy*, **83**, pp. 112-119, 1986.
[Seidenfeld, 1979] T. Seidenfeld. Why I am not an Objective Bayesian: Some Reflections Prompted by Rosenkrantz, *Theory and Decision*, **11**, pp.413-440, 1979.
[Shortliffe and Buchanan, 1985] E. H. Shortliffe and B. G. Buchanan. A Model of Inexact Reasoning in Medicine, Ch. 11 in Buchanan, B. G., and Shortliffe, E. H., eds., *Rule-Based Expert Systems: The MYCIN Experiments and the Stanford Heuristic Programming Project.* Reading, MA: Addison-Wesley, 1985.

[Skyrms, 1975] B. Skyrms. *Choice and Chance: An Introduction to Inductive Logic*. 2nd ed. Encino, CA: Dickenson, 1975.

[Tversky and Kahneman, 1982] A. Tversky and D. Kahneman. Judgments of and by representativeness, pp. 84-98 in D. Kahneman, P. Slovic & A. Tversky (Eds.), Judgment under uncertainty: Heuristics and biases. Cambridge, UK: Cambridge University Press, 1982.

[van Fraassen, 1990] B. Van Fraassen. *Laws and Symmetry*. Oxford: Clarendon, 1990.

[van Inwagen, 1996] P. van Inwagen. Why is There Anything at All? *Proceedings of the Aristotelian Society*, Supplementary Volumes, 70, pp. 95–120, 1996.

[Walley, 1991] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, 1991.

[Wasserman, 2000] L. Wasserman. Bayesian Model Selection and Model Averaging, *Journal of Mathematical Psychology* **44**, pp. 92-107, 2000.

[Williamson, 1999] J. Williamson. Countable Additivity and Subjective Probability, *British Journal for the Philosophy of Science,* **50**, pp. 401-16, 1999.

[Zadeh, 1978] L. Zadeh. Fuzzy Sets as the Basis for a Theory of Possibility, *Fuzzy Sets and Systems,* **1**, pp. 3-28, 1978.

[Zadeh, 1986] L. Zadeh. Is Probability Theory Sufficient for Dealing with Uncertainty in AI: A Negative View, pp. 103-116 in L. N. Kanal and J. F. Lemmer, eds., *Uncertainty in Artificial Intelligence: Machine Intelligence and Pattern Recognition. Volume 4*. Amsterdam: North Holland, 1986.