THE DEVELOPMENT OF SUBJECTIVE BAYESIANISM

James M. Joyce

The Bayesian approach to inductive reasoning originated in two brilliant insights. In 1654 Blaise Pascal, while in the course of a correspondence with Fermat [1769], recognized that states of uncertainty can be quantified using probabilities and expectations. In the early 1760s Thomas Bayes [1763] first understood that learning can be represented probabilistically using what is now called Bayes's Theorem. These ideas serve as the basis for all Bayesian thought.

1.1 Pascal's Insights: Probability and Expectation

In modern terms, Pascal's insight is that uncertainty about the occurrence of an event can be expressed as a probability and, more generally, that uncertainty about the value of a quantity can be expressed as a mathematical expectation. The basic objects of uncertainty can be thought as propositions or events in a non-empty Boolean algebra Ω that is closed under negation and countable disjunction. A *probability function* on Ω is a mapping P of Ω into real numbers that obeys these laws:

> *Normality.* For any $A \in \Omega$, $P(A \vee \neg A) = 1$ and $P(A \wedge \neg A) = 0$.
>
> *Finite Additivity.* $P(A \vee B) + P(A \wedge B) = P(A) + P(B)$.
>
> *Continuity.* If $A_1 \subseteq A_2 \subseteq A_3, \ldots$ is a countable sequence of events with $A = \vee_n A_n$, then $P(A_n)$ converges to $P(A)$.

These laws make probabilities countably additive, so that $P(\vee_n A_n) = \sum_n P(A_n)$ for any countable set of contraries $\{A_1, A_2, \ldots\}$. They also ensure that probabilities respect logical relationships, so that $P(A) = 1$ when A is a logical truth, $P(A) = 0$ when A is a contradiction, and $P(A) \geq P(B)$ when A entails B.

A *random variable* is a function f that assigns a real number $f(A_n)$ to each element of a partition[1] $\{A_1, A_2, \ldots\}$ in Ω. f's *expected value* relative to P is $Exp_P(f) = \sum_n P(A_n) \cdot f(A_n)$. Pascal maintained that the expected value of a quantity provides the best estimate of its actual value. A useful example is provided the puzzle that inspired Pascal to invent the concept of an expectation.

[1] A *partition* is a set of contrary propositions whose disjunction is a logical truth.

Handbook of the History of Logic. Volume 10: Inductive Logic. Volume editors: Dov Gabbay, Stephan Hartmann and John Woods.
General editors: Dov M. Gabbay, Paul Thagard and John Woods.

Problem of the Points. Joe and Moe are tossing a fair coin until five heads or five tails come up. If it's heads, Joe wins a pot of \$100. If it's tails, Moe wins the pot. After tosses of $\langle h, t, h, t, h, h\rangle$ the game is interrupted, and Joe and Moe must split the pot. An even split is unfair to Joe since he was likely to win, but giving it all to Joe would be unfair to Moe, who still had a chance of winning. Pascal proposed that each player's fair share is his *expected payoff*. To obtain these values one finds probabilities for each endgame — $P(h) = 1/2$, $P(t, h) = 1/4$, $P(t, t, h) = 1/8$, $P(t, t, t) = 1/8$ — and computes $Exp(\$\text{ Joe}) = 1/2\$100 + 1/4\$100 + 1/8\$100 + 1/8\$0 = \$87.50$ and $Exp(\$\text{ Moe}) = 1/2\$0 + 1/4\$0 + 1/8\$0 + 1/8\$100 = \$12.50$.

Though not explicit about it, Pascal clearly understood that expectations satisfy the following laws:

Linearity. If $f(\bullet) = a \cdot g(\bullet) + b \cdot h(\bullet) + c$, then $Exp(f) = aExp(g) + bExp(h) + c$.
Dominance. If $f(A_n) \geq g(A_n)$ for all A_n then $Exp(f) \geq Exp(g)$. If, in addition, $f(A_n) > g(A_n)$ for some A_n with $P(A_n) > 0$, then $Exp(f) > Exp(g)$.
Continuity. If $A_1 \subseteq A_2 \subseteq A_3, \ldots$ is a countable sequence of events with $A = \vee_n A_n$, then $Exp(A_n)$ converges to $Exp(A)$.

These principles follow easily from the laws of probability. Conversely, they entail the laws of probability when applied to *indicator functions* or *truth-valuations* that assign elements of Ω truth-values in $\{0, 1\}$ in a consistent way, where $v(A) = 1$ signifies truth and $v(A) = 0$ indicates falsity. Dominance ensures $Exp(v(\top)) \geq Exp(v(A)) \geq Exp(v(\bot))$ for all A. Linearity entails $Exp(v(\top)) = 1$, $Exp(v(\bot)) = 0$, and $Exp(v(A \vee B)) + Exp(v(A \wedge B)) = Exp(v(A)) + Exp(v(B))$ since $v(A \vee B) + v(A \wedge B) = v(A) + v(B)$. $Exp(v(A_n))$ converges to $Exp(v(A))$ by Continuity. So, expectations of truth-value obey the laws of probability. Expectations of truth-value, however, just are probabilities since $Exp_P(v(A)) = P(A)1 + P(\neg A)0 = P(A)$. Thus we see that Pascal's view that uncertain events should be evaluated using probabilities and his view that random variables should be evaluated using expectations are at root one idea.

1.2 Bayes's Insights: Conditional Probability and Bayesian Updating

Thomas Bayes's insight was to recognize the central role that conditional probabilities play in learning. Unlike unconditional probabilities, which reflect all-things-considered uncertainties, conditional probabilities reflect uncertainties about one event on the supposition that another occurs. The probability of A conditional on C, written $P(A|C)$, is required to satisfy the following law:

Conditional Probability. $P(A \wedge C) = P(C) \cdot P(A|C)$

Clearly, $P(A|B)$ can be expressed as $P(A \wedge C)/P(C)$ when C has positive probability, but the law leaves $P(A|C)$ unspecified when $P(C) = 0$. Some theories allow for the assignment of probabilities conditional on the supposition of events of probability zero. See, e.g., [Popper, 1959; Rényi, 1955].

Bayesians use conditional probabilities both to capture evidential relationships and to describe the effects of learning. The following five principles are essential to understanding the role of conditional probabilities in Bayesian accounts of inductive reasoning.

> *Conditional Probability.* $P(\dagger|C)$ is a probability on Ω for which $P(C|C) = 1$.
>
> *Total Probability.* If $\{C_1, C_2, \ldots\}$ is a partition, then $P(A) = \sum_n P(C_n) \cdot P(A|C_n)$.
>
> *Correlation.* A and C are positively correlated (i.e., $P(A \wedge C) > P(A) \cdot P(C)$) exactly if $P(A|C) > P(A)$, and they are uncorrelated exactly if $P(A|C) = P(A)$.
>
> *Preservation.* Conditioning on C does not disturb ratios of probabilities for events entailed by C, so that $P(A \wedge C|C)/P(B \wedge C|C) = P(A \wedge C)/P(B \wedge C)$ for all A, B.
>
> *Bayes's Theorem.* $P(A|C) = P(A) \cdot (P(C|A)/P(C))$

Conditional Probability tells us that conditioning always produces a new probability function that makes the condition certain. Total Probability expresses A's unconditional probability as a weighted average of its conditional probabilities. Correlation, a key element in Bayesian theories of evidence, captures the idea that one event is positively/negatively correlated with another to the extent that the occurrence of the first raises/lowers the second's probability. The "Bayes factor" $\beta_P(A, C) = P(A|C)/P(A)$, which provides one way of expressing the change that conditioning on C makes to A's probability, is a measure of this correlation. A and C are positively correlated when $\beta_P(A, C) > 1$, perfectly correlated when $\beta_P(A, C) = \beta_P(A, A) = 1/P(A)$. A and B are independent when $\beta_P(A, C) = 1$. They are anti-correlated when $\beta_P(A, C) < 1$, and perfectly so when $\beta_P(A, C) = \beta_P(A, \sim A) = 0$. Preservation ensures that conditioning on C produces a probability that "minimally departs" from P: if Q is a probability with $Q(C) = 1$ and $Q(A \wedge C)/Q(B \wedge C) = P(A \wedge C)/P(B \wedge C)$ for all A, B, then $Q(\bullet) = P(\bullet|C)$.

Bayes's Theorem sits at the heart of Bayesian approaches to inductive reasoning. Bayes is remembered not so much for discovering the theorem, a mathematical triviality, but for recognizing its significance. It relates the "direct" probability of one event condition on another to the unconditional probabilities of the two events and the "inverse probability" of the second event conditional on the first. As Bayes realized, there are many circumstances in which (a) one is interested in knowing the "direct" probability of some hypothesis conditional on certain data, (b) it is fairly easy to discover or deduce the "inverse" probability of the data

conditional on the hypothesis, and (c) one has "prior" information that allows one to estimate the probability of the hypothesis in the absence of the data. In such situations, Bayes's little theorem provides a way of arriving at the desired quantity in (a) from the information in (b) and (c).

More generally, imagine that one might receive an item of data x that is relevant to assessing the probability distribution over a partition of hypotheses $\mathcal{H}$. If one knows each of the "inverse probabilities" $P(x|h)$ and has a "prior" probability $P(h)$ for each for $h \in \mathcal{H}$, then Bayes's Theorem allows one to compute

$$P(h|x) = P(h)\cdot P(x|h)/[\sum_{g\in\mathcal{H}} P(g)\cdot P(x|g)]$$

In this way, "posterior" probabilities for hypotheses conditional on data are entirely determined by "prior" probabilities of hypotheses and "inverse probabilities" of data given hypotheses. Notice also that the expression $P(x|h)/[\sum_g P(g)\cdot P(x|g)]$ is just the Bayes factor $\beta(h,x)$. So, the theorem says $P(h|x) = P(h)\cdot\beta(h,x)$, which makes it clear that the Bayes factor measures the change that conditioning on x makes to h's probability.

Another illuminating form of the theorem reveals itself when we focus on *odds* and *likelihoods* rather than probabilities. The unconditional (conditional) odds of h to g is the ratio of $P(h)$ to $P(g)$ (or $P(h|x)$ to $P(g|x)$). Statisticians use the term "likelihood" to denote inverse probabilities. They call the map $L_x : H \to [0,1]$ defined by $L_x(\bullet) = P(x|\bullet)$ the *likelihood function* for x, and $L_x(h)/L_x(g) = P(x|h)/P(x|g)$ the *likelihood ratio* of h to g. In this jargon, Bayes's Theorem says that the ratio of the posterior odds to the prior odds is the likelihood ratio: $[P(h|x)/P(g|x)]/[P(h)/P(g)] = L_x(h)/L_x(g)$. The likelihood ratio is thus the factor by which we multiply unconditional odds to get conditional odds. In terms of Bayes factors, $L_x(h)/L_x(g) = \beta(h,x)/\beta(g,x)$. This formulation is noteworthy because it shows that, in contrast with probabilities, changes in *odds* among hypotheses produced by conditioning on x do not depend on prior probability over $\mathcal{H}$: ratios of likelihoods suffice.

> Example. Joe is being tested for the presence of a rare gene. We want to know how a positive result should affect our estimate of the chances that he has it. We know the test's true positive rate $L_+(gene) = P(+test \mid gene) = 0.9$, and its false positive rate $L_+(\neg gene) = P(+test \mid \neg gene) = 0.3$. We also know the gene occurs naturally in only one in a thousand cases, and have no reason to think Joe is special. So, $P(gene) = 0.001$. Under these circumstances, Bayes's theorem tells us that $P(gene \mid +test) = 0.0010.9 \;/\; [0.0010.9 + 0.9990.3] \approx 0.003$. We can also use the likelihood ratio to determine how a positive test will alter the odds of Joe having the gene. Since $L_+(gene)\;/L_+(gene) = 3$, the odds will triple. If our pre-test situation had been different and, say, we knew that Joe's mother has the gene, and so $P(\text{gene}) = 0.5$, then $P(gene \mid +test) = {}^3/_4$ owing to the higher unconditional probability. But, the likelihood ratio is still 3.

Bayes's other great insight was to recognize that the conditional probabilities governed by his theorem are closely tied to *learning*. In modern terms, we would express his idea like this:

> *Learning as Bayesian Updating.* Imagine a person's whose state of uncertainty is characterized by a "prior" probability P_0 on Ω, and who is not dogmatic about x, so that $1 > P_0(x) > 0$. If the person undergoes a learning experience in which the only new information she acquires is that x is certainly true, then her post-learning "posterior" probability P_1 should coincide with her pre-learning probability conditional on x, $P_1(\bullet) = P_0(\bullet|x)$.

Some people call this "conditioning," others "conditionalization," but the basic idea is the same: dogmatic learning, the kind in which one becomes certain of x (and this is all one learns), involves reapportioning probabilities so that all probability that was previously invested in $\neg x$ is shifted onto x in a way that preserves probability ratios among propositions that entail x.

The laws of conditional probability ensure that Bayesian updating has features that seem desirable in any dogmatic learning rule.

> *Dogmatism.* Updating on x makes x certain: $P_1(x) = 1$.
>
> *Preservation.* Updating on x leaves certainties intact: $P_1(y) = 1$ whenever $P_0(y) = 1$.
>
> *Coherence.* P_1 is a probability if P_0 is a probability.
>
> *Responsiveness to Evidence.* If x *is evidence* for (against)h according to the pre-learning probability, then updating on x raises (lowers) x's probability.
>
> *Minimal Change.* Updating on x does not disturb ratios of probabilities of events entailed by x, i.e., the *Bayes update factors* $P_1(y \wedge x)/P_0(y \wedge x) = 1/P_0(x)$ are constant.
>
> *Accumulation.* Learning is accumulative in the sense that updating on x_1 and then x_2 is equivalent to updating on their conjunction.[2]
>
> *Commutativity.* The temporal order in which data is acquired is irrelevant to its evidential import. If Q is obtained from P_0 by updating on x_1 and then on x_2, and if Q^* is obtained from P_0 by conditioning on x_2 and then on x_1, then $Q(\bullet) = Q^*(\bullet)$.

All these features seem like strengths given that Bayesian updating on x is only appropriate as a response to learning experiences whose entire content is that x, and nothing stronger, is certainly true, and where x does not contradict anything previously known. One might have worries about how common these sorts of

[2]This should not be confused with monotonicity, the idea that if learning x supports h then learning x and y will support h as well. Bayesian updating is highly non-monotonic.

experiences are (see §3.2), but Bayesian updating is the right way to model them when they do occur.

1.3 The Basic Bayesian Apparatus for Inductive Reasoning

As we have seen, Pascal and Bayes had four big ideas:

- Uncertainty is best represented using probabilities.
- Estimates of uncertain quantities are expectations.
- Conditional probabilities are fruitfully interpreted using Bayes's Theorem.
- Dogmatic learning experiences involve conditioning on the data received.

Let's call this *the basic Bayesian apparatus for inductive reasoning*.

To see how it works, consider an abstract model of inductive problems. A *Bayesian experimental setup* is a triple $\langle \mathcal{H}, \mathcal{X}, P\rangle$ where $\mathcal{H}$ is a partition of hypotheses, $\mathcal{X}$ is a partition of *potential data*, and P is a probability defined over the Boolean algebra $\mathcal{H}\wedge\mathcal{X}$ generated by all conjunctions $h\wedge x$ with $h\in\mathcal{H}$ and $x\in\mathcal{X}$. A *learning experience* is an exogenous change in the probability distribution on $\mathcal{X}$ whose direct effect is to replace each "prior probability" $P(x)$ by a "posterior probability" $Q(x)$.[3] In some experiences a single data item is learned for certain, in which case $Q(x)=1$ for some x, but the model allows for experiences that readjust probabilities over $\mathcal{X}$ in other ways as well. It will, however, be required that learning never "wakes the dead" by raising an event of prior probability zero to positive probability.

The goal of Bayesian inference is to explain how learning-induced changes in the probabilities over $\mathcal{X}$ ramify through the rest of $\mathcal{H}\wedge\mathcal{X}$, in particular how they alter probabilities on $\mathcal{H}$. Given a "prior" P defined over $\mathcal{H}\wedge\mathcal{X}$ and a partial posterior $Q^{\mathcal{X}}$ defined over $\mathcal{X}$ alone, the goal is to find the extension Q of $Q^{\mathcal{X}}$ over all of $\mathcal{H}\wedge\mathcal{X}$ that is best justified in light of both the information encoded in the prior and the new evidence. Bayesian updating rules amalgamate this evidence into a single posterior that agrees with the new observations about $\mathcal{X}$ and reflects the input of prior information.

Different rules are appropriate for different evidential inputs. In simplest learning experiences the experiment will show only that the truth lies within some subset X of $\mathcal{X}$. Learning as Conditioning then requires that $Q(\bullet)=P(\bullet|\mathcal{X})$ and Bayes's Theorem tells us that, for each $h\in\mathcal{H}$,

$$\begin{aligned} Q(h)=P(h)\cdot[P(X|h)/P(X)] &= P(h)\cdot[\textstyle\sum_x P(x|X)\cdot(P(x|h)/P(x))] \\ &= P(h)\cdot[\textstyle\sum_x (Q(x)/P(x))\cdot P(x|h)] \end{aligned}$$

[3] I am slurring over the distinction between probability functions and probability densities. One needs to be careful about this when either $\mathcal{H}$ or $\mathcal{X}$ is uncountable.

If we wanted to estimate the value of some random variable f defined on $\mathcal{H}$ in light of this information we would use

$$\begin{aligned} Exp_Q(f) &= \textstyle\sum_h f(h)\cdot Q(h) = \sum_h f(h)\cdot P(h)\cdot[P(X|h)/P(X)] \\ &= \textstyle\sum_h f(h)\cdot P(h)\cdot[\sum_x (Q(x)/P(x))\cdot P(x|h)] \end{aligned}$$

Example. Three balls are about to be drawn, with replacement, from an urn that contains black and white balls. You know that the urn can be of two types: Type-1 urns contain 20% black balls; Type-2 urns contain 60% black balls. You want to know the urn's type, and are also interested in estimating the number of black balls that will appear among the last two balls on the basis of information about the first ball. Here $\mathcal{H} = \{Type_1,\ Type_2\}$ and $\mathcal{X} = \{Black,\ White\}$. Let's also suppose that you have information that leads you to think that $P(Type_1) = 0.25$,$P(Type_2) = 0.75$.

The prior probability then looks like this:

$\wedge$	$Type_1$ (0.3)	$Type_2$ (0.7)
Black (0.5)	0.05	0.45
White (0.5)	0.20	0.30

Suppose the first ball drawn is black. Conditioning on this information requires you to move all the posterior probability on to Black and to preserve ratios among events of the form $Type_m \wedge Black$, so that the posterior looks like this:

$\wedge$	$Type_1$ (0.1)	$Type_2$ (0.9)
Black (1)	0.1	0.9
White (0)	0	0

To estimate the number of blacks that will appear in the next two draws, first use the law of total probability to compute the probabilities: $Q(BB) \approx 0.33$, $Q(BW) = Q(WB) \approx 0.235$, $Q(WW) \approx 0.2$, and then the expectation is given by

$$Exp_Q(\#B) = 2\cdot Q(BB) + 1\cdot Q(BW) + 1\cdot Q(WB) + 0\cdot Q(WW) = 1.6$$

1.4 Comparison with Frequentist Approaches

It may be instructive to contrast the Bayesian approach to inductive inference with the *likelihood based* approaches of "frequentist" statisticians. One can think of the probability that appears in a Bayesian experimental setup as being factorable into

a prior probability P^H over $\mathcal{H}$ and a family of normalized[4] *likelihood functions* $L_x : \mathcal{H} \to [0, \infty)$, one for each $x \in \mathcal{X}$. The likelihoods must agree with P^H in the sense that $\sum_x \sum_h P^H(h) \cdot L_x(h) = 1$. Together the prior and likelihoods determine an unconditional probability for each atomic element of $\mathcal{H} \wedge \mathcal{X}$ via the rule $P(x \wedge h) = P^H(h) \cdot L_x(h)$, and this fixes P over all of $\mathcal{H} \wedge \mathcal{X}$. For example, the probability of a data item $x \in \mathcal{X}$ is the expected value of its likelihood function $P(x) = \sum_n P^H(h_n) \cdot L_x(h_n)$.

Frequentist statisticians also draw inductive inferences using likelihoods, but they eschew priors. In frequentist experiments the probability P is replaced by a family of general *likelihood functions* $l_x : \mathcal{H} \to [0, \infty)$ that are only determined up to multiplication by a positive constant, so that $l_x(h) = \lambda_x \cdot P(x|h)$ where $\lambda_x > 0$ is a constant that can depend on x but not h. The values of l_x have no meaning taken individually. In particular, they cannot be directly identified with inverse probabilities in a Bayesian model. They can, however, be used to express facts about the *relative* degree to which a datum x is predicted by hypotheses in $\mathcal{H}$. For example, one can say that a given l_x assumes its maximum at h^x or one can compute *likelihood ratios*, $l_x(h)/l_x(g) = P(x|h)/P(x|g)$, which compare x's expectedness given distinct hypotheses. Since $l_x(h)/l_x(g) = \beta(h, x)/\beta(g, x)$, this means that classical statisticians can, if they want, describe certain kinds of *changes* in the probabilities of hypotheses. They can say, e.g., that learning x increases the probability of h as a proportion of its prior more than it increases the probability of g as a proportion of its prior. But, whatever happens, they will *not* say anything about the absolute probabilities of hypotheses in light of the data since this information cannot be extracted from the likelihood function without invoking unconditional probabilities for elements of $\mathcal{H}$. On a classical picture, *all* inductive reasoning boils down to drawing inferences from observed data based on facts about likelihoods.

This suggests a very different picture of inductive inference from what one finds in the Bayesian approach. For example, to estimate the value of a random variable $f : \mathcal{H} \to \Re$ in light of datum x the classical statistician cannot calculate f's expected value conditional on x since that invokes a prior over $\mathcal{H}$. Instead, she might use maximum likelihood estimation and estimate f's value as $f(h^x)$ where h^x is the hypothesis in $\mathcal{H}$ for which $l_x(h)$ attains its maximum. Likewise, the classical statistician cannot adopt a broadly Bayesian policy of assessing the acceptability of hypotheses on the basis of their posterior probabilities, since these depend on priors. She might, instead, decide to reject a hypothesis if, in the event it were true, the probability of observing the data actually observed or even more unlikely data falls below some threshold value.

By adopting a model that permits both likelihoods and priors, Bayesians have been able to secure a far richer and more coherent theory of inductive inference than anything to which frequentist statisticians might aspire. As frequentists like

[4] L is *normalized* when there are non-negative constants μ_x, one for each $x \in \mathcal{X}$, with $\sum_x \mu_x = 1$ such that the numbers $p_x = \sum_n \mu_x \cdot L_x(h_n)$ sum to one. This ensures that the $L_x(h_n)$ can be consistently thought of as being equal to the probability of x conditional on h_n.

Fisher [1959] and Neyman [1950] have been quick to reply, however, there are substantial costs associated with these benefits. The use of a prior in drawing inductive inferences requires one to trust its probabilities. Sometimes this is fine. In situations where there are determinate, objectively measurable and agreed upon probabilities for hypotheses and potential data items, everyone will agree that there is no better way to draw conclusions than by using the Bayesian apparatus. Everybody should be a Bayesian in the casino! Even so, there is no getting around the fact that Bayesianism is a garbage-in-garbage out enterprise: if one applies the apparatus using a prior that is accurate and well unjustified, the conclusions derived will be accurate and well justified as well; if one applies the apparatus using a prior that is inaccurate or unjustified, the conclusions will also inaccurate or unjustified. This is the heart of frequentist misgivings. From the perspective of frequentist statisticians, Bayesian methods carry a massive uncollateralized risk or error.

By analogy, suppose that a rogue group of logicians proposed to redefine the notion of logical consequence as follows: p counts as a logical consequence of q not only when p can be deduced from q by the laws of logic, but also when p can be deduced form q together with $r, s, t, \ldots$, where $r, s, t, \ldots$ are "prior" premises the rogue logicians find reasonable. While the rogues will be able to deduce far more that any classical logician can, their reliance on "priors" introduces an objectionable new source of error. Indeed, it would seem that part of the reason to have a logical consequence relation is to avoid risking such errors. Frequentist statisticians see Bayesians as rogues of the same sort. They are skeptical of Bayesian methods because they doubt that prior probabilities can be made epistemologically respectable, and feel that their introduction threatens to undermine the accuracy and objectivity of our inductive reasoning.

2 THE PROBLEM OF THE PRIORS

Bayesians must address this criticism head-on by offering some rationale for the use of priors in inductive reasoning. Three broad sorts of rationales have been proposed. *Objective Bayesians* maintain that certain priors can be justified *a priori* as the uniquely correct way to represent uncertainty. *Subjectivists* argue, in contrast, that priors reflect the subjective degrees of belief, or *credences*, of agents. They are bound by no requirements, save the laws of probability. *Tempered Bayesians* believe that priors reflect subjective credences, but suggest that agents who update in light of evidence will, at least eventually, end up using "priors" that are well justified and likely to be accurate. Let's consider these three approaches in turn.

2.1 Objective Bayesianism and Ignorance Priors

The first objective Bayesian was Bayes himself, but the driving force behind the approach was unquestionably Pierre-Simon Laplace whose *Principle of Insufficient*

Reason[5] was meant to provide an *a priori* rationale for imposing uniform prior distributions in contexts where the data provides no basis for distinguishing among rival hypotheses. In Laplace's terminology, the hypotheses in $\mathcal{H}$ are "equipossible" when nothing in the available evidence favors any one over any other. When the evidence is symmetrical in this way, Laplace reasoned, the probability assignment that best reflects this evidence is symmetrical as well. He codified this insight in the following principle:

> *PIR*. If the available evidence provides no reasons to favor any hypothesis in $\mathcal{H}$ over any other, then the uniquely correct prior to assign is the uniform distribution in which $P(h) = P(g)$ for all $h, g \in \mathcal{H}$.

Notice that the Principle does not say anything about the quantity or quality of the "available evidence". This is one of its most controversial features since the main purpose of *PIR* is to allow for the assignment of prior probabilities in contexts where little or no evidence exists. These assignments — which have been called "ignorance priors," "uninformative priors," "informationless priors" or "reference priors" — are supposed to provide the input for the Bayesian apparatus in contexts where there is not much hard data to be had.

> Example. An urn was selected from a population $\{\mathrm{U}_0, \mathrm{U}_1, \ldots, \mathrm{U}_{10}\}$ where U_i contain i black balls and $10 - i$ white balls. What is the prior probability that a ball randomly drawn from the urn will be black? Before answering, consider three ways of fleshing out the story.
>
> $Case_1$ We know that our urn is U_5.
>
> $Case_2$ We know that our urn was selected from via a random process in which each U_i had an equal objective chance of being chosen.
>
> $Case_3$ We know nothing about the identity of the urn or about the process by which it was selected.
>
> Since our evidence for *Black* and $\neg$*Black* is symmetrical in all three cases, *PIR* tells is that the right probability is $P(Black) = P(\neg Black) = 1/2$ in all three cases.

$Case_3$ is the controversial one. In $Case_1$ and $Case_2$ *PIR* gets things right, but the answer can be derived independently from the plausible requirement that priors should line up with known objective chances (see § 4.2). In $Case_3$, however, we know almost nothing about the chances of *Black*. How are we supposed to go from this sparse evidential basis to the same probability assignment that is warranted in the other two cases?

The standard rationale goes like this: Consider all possible chance hypotheses about the way the urn was selected, i.e., the set Π of probability distributions on

[5]The name comes from von Kries [1871]. The name *Principle of Indifference* is used by Keynes [1921].

$\{0, 1, \ldots, 10\}$. Since our evidence provides no grounds for preferring any $\pi \in \Pi$ to any other, our prior should not play favorites among Π's elements. The only way to avoid playing favorites is by assigning each $\pi \in \Pi$ the same probability. Thus, the right prior for this problem is uniformly distributed over Π, which forces $P(black)$ to be $1/2$. In effect, $Case_3$ is reduced to case $Case_2$.

This reasoning offers the glittering prospect of a Bayesian inductive logic in which prior probabilities are justified *a priori* on the basis of the sound epistemological principle that one should not include information in one's prior that is not found in the data. Once such a uniquely correct "ignorance prior" is in place, the Bayesian apparatus tells us everything there is to know about inductive reasoning. To obtain the posterior probability distribution that is best justified in light of one's data one should update the ignorance prior by conditioning on that data. Here is a noteworthy application of this method.

> Example. You are presented with a coin of unknown bias. You toss it once and it comes up heads. What is the probability that it will land heads on the second toss? Laplace [1774] argued that this probability should be exactly $2/3$. Since you know nothing about the coin's bias p, he reasoned, you should invoke *PIR* and adopt a prior with the uniform density dp over the unit interval. The probability of a head on the first toss is then $\int_0^1 p\ dp = 1/2$. If you observe a head and update via Bayes's theorem, you obtain a posterior density of $(2pdp)$, and the probability of getting a second head is $\int_0^1 p \cdot (2pdp) = 2/3$. If you observe a second head and update, you get a density of $(3p^2dp)$ and the probability of a third head is $\int_0^1 p \cdot (3p^2dp) = 3/4$. More generally, if you keep tossing and conditioning you will emulate Laplace's *rule of succession*, which says that the probability of observing a head on the $N+1^{st}$ trial given s heads and $N-s$ tails on previous trial is $\frac{s+1}{N+2}$.

If this broad approach to inductive reasoning is correct, then objective Bayesians have an answer for frequentist critics. It is legitimate to invoke ignorance priors when assessing the impact of data on hypotheses, they will argue, because the "added premises" can be justified *a priori*. Among all the probability distributions that could be applied to a given inductive problem, the ignorance prior is the one that introduces the least amount of additional information: any other way of proceeding would require drawing distinctions among hypotheses that are not justified by the data.

This Laplacian picture of inductive logic, which was greatly advanced by Jeffreys [1939], finds it fullest expression in the work E. T. Jaynes [2003]. Jaynes, a militant objectivist about priors, writes:

> Consistency demands that two persons with the same relevant prior information should assign the same prior probabilities... Objectivity requires that a statistical analysis should make use, not of anybody's personal opinions, but rather the specific factual data on which those opinions are based. [1968, p. 53]

Jaynes seeks to secure consistency and objectivity by using information theory to generalize *PIR*. In Jaynes's [1973] picture, an inductive problem is defined by a set of *objective constraints* that stipulate expected values[6] $c_1, \ldots, c_K$ for random variables $f_1, \ldots, f_K$ defined on $\mathcal{H}$. An ignorance prior P^* for such a problem must satisfy two conditions: it must yield the required expectations, so that $Exp_{P*}(f_m) = c_m$ for all m; and it must maximize *entropy* $Ent(P) = \Sigma_m\ P(h_m) \cdot ln(P(h_m))$ across all probabilities P that yield the required expectations. Under broad conditions, there will always be a unique such $P*$. Moreover, since *Ent* measures the amount of *information* that a probability convey about elements of $\mathcal{H}$, Jaynes argues that using P^* as one's prior introduces the least amount of additional information into the problem consistent with constraints. The proposal, then, is this:

> *MaxEnt*. If the available evidence specifies that $Exp_P(f_k) = c_k$ for all k, then the uniquely correct prior P^* is such that $Ent(P^*) > Ent(P)$ for all P that produce the required expectations.

This generalizes *PIR* since, in the absence of constraints, the uniform distribution uniquely maximizes entropy. But, MaxEnt has wider application.

> Example. In the strange country of Bulmania the expected number of boys among families with two children is 1.5. To find the MaxEnt prior over sex-pairs $\langle BB, BG, GB, GG\rangle$, use Lagrange multipliers to obtain $P^* \approx \langle 0.564, 0.186, 0.186, 0.064\rangle$.

Should we embrace this "objective" method for selecting priors? Many philosophers and statisticians reject *PIR* and MaxEnt on the grounds that they produce inconsistent results when applied to single situation under different descriptions. This objection was first raised by John Venn [1866] and has been recapitulated many times since, perhaps most forcefully in Keynes [1921]. Here are two famous versions of it:

> Example (Venn's Paradox). Your car's gas tank is a cube with sides between 20 and 40 centimeters in length. This is *all* you know. Partitioning the possibilities by side-length and distributing your prior probability uniformly over [20cm, 40cm] yields an expected side-length of 30cm and an expected volume of 27 liters. But, partitioning by volume, and distributing your prior uniformly over [8 liters, 64 liters] produces an expected volume of 36 liters and an expected side length of 33cm.

> Example. (Bertrand's Paradox). Imagine an equilateral triangle of side length $\sqrt{3}$ that is inscribed within a circle of radius 1. What is the probability p that a chord drawn at random across the circle will have length greater than $\sqrt{3}$? Here are three plausible answers:

[6] More generally, the constraints might specify an allowable range of expected values for each variable.

- Chords with midpoints equidistant from the circle's center are the same length, and those for which this distance d is more than $1/2$ are longer than $\sqrt{3}$. Thus, if we apply *PIR* to the possible values of $d \in [0, 1]$, we get $p = 1/2$.
- Each chord splits the circle into two arcs, a short one of length $a \in [0, \pi]$ and a long one of length $2\pi - a$. Chords with the same short arc length have the same length, and those for which $2\pi/3 < a \leq \pi$ are longer than $\sqrt{3}$. If we apply *PIR* to the possible values of a, the interval $[0, \pi]$, we get $p = 1/3$.
- Chords with midpoints that fall inside the circle of radius $1/2$ that is inscribed within the triangle have lengths that exceed $\sqrt{3}$. Imposing a uniform density over all possible midpoints for the chord (points in the circumscribing circle) produces a probability equal to the ratio of the area of the smaller circle, $\pi/4$, to the area of the larger circle, π. Thus, $p = 1/4$.

While opponents of *PIR* often regard these objections as dispositive, its proponents argue that the principle is being misapplied. In response to Venn's objection, Jeffreys [1939] argued that an ignorance prior should not be uniform over *either* length in centimeters or volume in liters since these involve arbitrarily chosen measuring units, and it is clear *a priori* that policies for selecting priors should not rely on arbitrary choices. An acceptable prior should be *scale-invariant*: if $T(x) = u{\cdot}x$ (for $u > 0$) is a transformation that alters the unit of distance from centimeters x into, say, inches ($u = 0.3937$), then the rule for assigning priors should yield the same results whether applied to x or to $T(x)$. More precisely, the prior should be defined by a density $p(\cdot)$ such that $P(a < x < b) = \int_a^b p(x)dx = \int_a^b p(T(x))dx$ for all a and b. It can be shown, see [Lee, 1997, p. 101], that the only density that fits the bill for all $u > 0$ is $p(u{\cdot}x) = (u{\cdot}x)^{-1} / \int_{20}^{40} (u{\cdot}x)^{-1}\ dx$. This makes $P(u{\cdot}a < u{\cdot}x < u{\cdot}b) = ln(b/a)/ln(2)$ the unique scale invariant prior for lengths. In effect, Jeffreys applies *PIR* not to length itself, but to its logarithm. By similar reasoning, $P(u{\cdot}c < u{\cdot}v < u{\cdot}d) = ln(c/d)/ln(8)$ is the unique scale-invariant prior for volume. With these priors the contradiction vanishes since for any side lengths $a < b$ and any unit of length $u > 0, P(u^3{\cdot}a^3 < u^3{\cdot}v < u^3{\cdot}b^3) = ln(b/a)/ln(2) = P(u{\cdot}a < u{\cdot}x < u{\cdot}b)$.

Jeffreys and his followers have generalized this sort of maneuver to cover a variety of applications. A particularly beautiful example is Jaynes's [1973] solution to Bertrand's paradox. Jaynes agues that, in addition to rotational symmetry (which all three proposed solutions have), an adequate rule for choosing priors should not vary with changes in the size of the circle or its position in space. Thus, we should look for a prior density that is invariant under rotations, under variation in scales for measuring lengths, and under translations of the midpoint of the circle in space. Remarkably, this suffices to fix $p = 1/4$ as the uniquely correct answer! The moral is supposed to be that *PIR*/MAXENT can only be applied after one has identified all relevant symmetries that apply in a situation. A "well

posed" inductive problem will include evidential constraints that express all of these various symmetries and will treat symmetrical alternatives as "equipossible". The paradoxical character of *PIR*/MAXENT then disappears, says Jaynes.

The Jeffreys/Jaynes approach is subject to a number of criticisms. First, there is the technical worry that ignorance priors are often "improper" in the sense that their probability densities go infinite. For example, the Jeffreys density for the Venn problem blows up both when zero is in x's range and when the range is unbounded.[7] A more serious issue concerns the status of symmetry principles. While objective Bayesians portray these as *a priori* constraints on priors, they clearly import *a posteriori* information into the situation. There is no reason, in principle, why the units in which length or volume is measured *could* not matter to the probabilities of various results.

> Example. You know that all Bulmanian cars have cubical gas tanks with a capacity of B buliliters, but you have no idea whether the buliliter is a unit of length or volume. When you ask how much gas is your rental car you are told only that the attendant who filled it always chooses some number $b \in [0, B]$ that strikes her as lucky and puts b buliliters in the tank. It seems in the spirit of *PIR* to impose a uniform distribution over b, rather than $ln(b)$. The units matter since the attendant makes her choice on the basis of the unitless *number*.

The imposition of rotational and translational symmetries seems even less *a priori*. Take the Bertrand paradox. Jaynes suggests that one can deduce the physically relevant symmetries *a priori*, and even argues on this basis that the observed frequencies must conform to $p = 1/4$. He reports an experiment that confirms this: "The Bertrand experiment has, in fact, been performed by... tossing broom straws from a standing position onto a 5-in.-diameter circle drawn on the floor.... 128 successful tosses confirmed [p = 1/4] with an embarrassingly low value of chi-squared." The key word in this quote is the "the". While $p = 1/4$ is a natural solution when we think of the chords being generated by throwing straws at random across a circle on the floor, there is no reason to think of this is *the* Bertrand experiment. It is *one* chance setup that fits with Bertrand's story, but there are physically possible random experiments for which the other two solutions make sense. Suppose that, instead of tossing straws, we think of the circle as a rapidly spinning wheel of fortune that repeatedly carries a distinguished point past a stationary pointer. If the wheel spins at a constant rate until, at some random time, it suddenly stops and the chord is identified with the line segment from the fixed point to the pointer, then $p = 1/3$ is right since the fixed point spends a third

[7] There are ways to finesse this. For instance, the uncertainty inherent in a problem can often be confined to a finite interval. It is also possible to show that certain improper priors generate *proper* posteriors when updated. For example, if you begin being absolutely uncertain where a length falls in the interval $[0, \infty)$ your Jeffreys prior will be improper, but conditioning on any item of data of the form "$x \in [a, b]$", for $0 < a < b < \infty$, yields a proper posterior. It is controversial whether these maneuvers succeed, see Howson [2002, 53-56] for relevant discussion.

of its time farther than $\sqrt{3}$ away from the pointer. Alternatively, if we imagine ourselves tossing circles of fixed size onto a single line painted on the floor, $p = 1/2$ is correct. Since nothing precludes these scenarios *a priori*, symmetry conditions cannot be *a priori* either. All Bayesians are, of course, happy to impose empirically motivated symmetry conditions when appropriate, but most do not believe that they can be deduced *a priori*.

A deeper problem with *PIR*/MaxEnt, whether augmented with symmetry principles or not, is that they seek to capture states of ambiguous or incomplete evidence using a *single* probability function. Many people see this as an illegitimate way of smuggling in information. Recall Jaynes's assertion that "consistency demands that two persons with the same relevant prior information should assign the same prior probabilities." As we shall see, some Bayesians reject the idea that believers with the same objective evidence should end up in the same epistemic state. But, even if one grants that perfect symmetry in one's evidence requires symmetry in one's epistemic attitudes, it is a further step to say that the best way to capture these attitudes is by assigning equal prior probabilities. When your evidence is highly *unspecific*, see Joyce [2005], it might be better not to assign any determinate prior probabilities at all. Recall the example of the ten urns. Suppose the urn in front of you is painted red, and that you started out knowing that U_5 is red and being confident (but mistaken) that is the only red urn. Imagine that you undergo a series of experiences in which you become increasingly uncertain about the number of red urns. First, you learn that U_4 and U_6 are red. Then, you learn that U_3 and U_7 are red, and so on until you end up knowing that all eleven urns are red. *PIR* and MaxEnt say that all these symmetrical states of evidence should be represented by a $P(\mathit{Black}) = 1/2$ prior. While this is proper in the first case, it seems increasingly inappropriate as we move down the line. As the number of red urns grows you steadily lose information about the proportion of balls in the urn in front of you, but this loss is nowhere reflected in your unconditional probabilities.[8]

The problem with this, as R. A. Fisher 1922, p. 326] forcefully argued, is that it extracts "a vitally important piece of knowledge, that of the exact form of the distribution... out of complete ignorance." Fisher's point is that using a single probability function to represent your ignorance in all these different evidential situations requires smuggling in information that is not found in the data. This

[8]There will be differences in the resilience of *conditional* probabilities in light of various potential data items. When one knows that U_5 is the only red urn, $P(\mathit{black} \mid \mathit{data})$ will remain fixed at $1/2$ for every pattern of black and white balls might be drawn (with replacement). For any other case, *black*'s probability will vary with changes in the initial data, and greater variation will occur for less specific data. This difference, though important for other purposes, does not answer the objection being pressed here. If the evidence is that n black and $N - n$ white balls are observed in the first N draws, then the conditional probability in the least specific case ($= 11$ red urns) is given by $P(U_k|N,n) = P(N,n)^{-1} \cdot (1/11) \cdot (k/10)^n \cdot ((10-k)/10)^{N-n}$, where $P(N,n) = \frac{N!}{n!(N-n)!} \Sigma_k (1/11) \cdot (k/10)^n \cdot ((10-k)/10)^{N-n}$ is the prior probability of receiving that particular sequence of data. $P(N,n)$ reflects the prior information contained in the uniform distribution. In this context, Fisher's concern resurfaces as an objection to using the prior $P(N,n)$ over the data sequences to compute conditional probabilities.

information is vividly revealed when we focuses on the prior distribution over possible data sequences. If $P(N, n)$ denotes the prior probability an arbitrary sequence of N draws, with replacement, in which n black balls are observed, then $P(N, n) = {}^{N!}/_{n!(N-n)!} \cdot {}^{1}/_{2^N}$ when we know that U_5 is the only red urn. In the least specific case where we know all the urns are red, however, $P(N, n) = {}^{N!}/_{n!(N-n)!}$ $\Sigma_k \, ({}^{1}/_{11}) \cdot ({}^{k}/_{10})^n \cdot ({}^{(10-k)}/_{10})^{N-n}$, for $k \in \{0, 1, \ldots 10\}$. This is a *very* specific set of numbers.[9] The reason we get such a specific distribution, of course, is that we are calculating it exactly as we would if we knew that each urn had an equal objective chance of being selected. But, since we know no such thing, we have no right to such specific numbers. Objective Bayesianism is just bad epistemology from Fisher's perspective. It may be that the amount of added information encoded in the *PIR*/MaxEnt prior is, in some sense (e.g., in terms of entropy) the minimum that can be achieved using a single probability function, but this does not change the fact that the decision to represent uncertainty using a single probably function often involves adding information. In the urn case, for example, you smuggle in information in every case except the first, and the more red urns there are the more information you smuggle in.

Bayesians who share Fisher's worries can go a number of ways. Pure subjectivists feel that any probability not directly contradicted by the constraints of an inductive problem may legitimately be used as a prior, so that, e.g., if you know the urn is in $\{U_3\ U_5,\ U_7,\ U_8\}$ then it is permissible to set $P(Black)$ to 0.3, 0.5, 0.7, 0.8 or to any mixture of these values. Subjectists thus agree with objective Bayesians that it is permissible to introduce information not included in the prior data to for purposes of making inductive inferences, but they deny that we should attach any special importance to any one way of adding information as opposed to any other. Other Bayesians — see [Levi, 1980; Jeffrey, 1983; Walley, 1991; Kaplan, 1996; Joyce, 1999] – represent symmetrical but incomplete states of evidence not with symmetrical probability values, but by symmetrical *sets* of probability functions. Instead of using one probability function to capture prior uncertainty this approach uses the *family* of all probability functions that the evidence does not explicitly exclude. When one knows only that the urn is $\{U_3,\ U_5,\ U_7,\ U_8\}$, for example, one's "prior" would be the set of all probability functions defined over {3, 5, 7, 8} and the event of a black ball being draws would have the imprecise, "interval-valued" probability $P(black) \in [0.3, 0.8]$.[10] Proponents of such "imprecise" probabilities argue that they are both more psychologically realistic than sharp probabilities for representing degrees of belief, and that they are also the proper response to incomplete, ambiguous, or unspecific evidence.

[9] For $N = 5$ it generates the distribution $\langle 0.20075, 0.14775, 0.1515, 0.1515, 0.14775, 0.20075 \rangle$.

[10] Not all imprecise probabilities will be interval-valued, and not all families of probability functions that may be used to represent uncertainty will be convex. While some have required this, see [Levi, 1980, p. 402], it is not plausible in the (fairly common) case in which the prior information specifies that two events are independent without specifying any definite probability for either one. For relevant discussion see [Jeffrey, 1987].

Even if one adopts one of these "non-objectivist" approaches, however, there is no denying that *PIR*, MAXENT and other methods for assigning priors have often been successful in practical applications. As a result, many Bayeisans who reject the idea that "ignorance priors" can be justified *a priori* will still agree with Gillies [2000, p. 48] that *PIR* and MAXENT are very fruitful *heuristic* devices even if it is not valid as *logical* principles. Indeed, when approaching an inductive problem from a Bayesian perspective, it is often useful to start with an ignorance prior, and then to be willing to modify one's thinking in light of empirical information as well as expert opinion. There is nothing wrong with this, anti-objectivist Bayesians will say, provided that one always keeps in mind that "ignorance priors" are empirical hypothesis like any other.

2.2 Subjective Bayesianism and the Requirement of Coherence

Subjective Bayesianism is the work of many hands, but key contributions have been made by Frank Ramsey, Bruno de Finetti, Leonard Savage, I. J. Good, David Lindley and Richard Jeffrey. The unifying ideas of the subjectivist approach are these:

- Beliefs come in varying *gradations of strength*. Instead of asking whether a person accepts or rejects a proposition outright, we must speak of her *level of confidence* in it.
- A person's level of confidence in a proposition corresponds to the extent to which she is disposed to presuppose its truth in her theoretical and practical reasoning.
- The goal of an account of inductive reasoning is to explain how the gradational beliefs, or *credences*, of rational agents change in light of changes in their evidence.
- In a Bayesian experimental setup, the prior distribution of credences over $\mathcal{H}$ reflects the initial amount of confidence that an agent invests in the various hypotheses in $\mathcal{H}$. The likelihood function represents her subjective probabilistic predictions about what data from $\mathcal{X}$ she is likely to receive conditional on various hypotheses obtaining.
- In idealized cases, a person's credences can be represented by a single function b from $\mathcal{H} \wedge \mathcal{X}$ into $[0, 1]$. (In more realistic cases, sets of functions will be employed.)
- *The Requirement of Probabilistic Coherence*. A rationally permissible credence assignment must conform to the laws of probability.[11]

[11] In the idealized case, b must be a probability. In less ideal cases, every credence function consistent with an agent's attitudes must be a probability. For example, the agent cannot judge that A is more probable than B and that $A \vee C$ is less probable than $B \vee C$ when C is incompatible with A and B since no probability function condones these judgments.

- *The Requirement of Conditioning.* A rationally permissible method of belief revision must involve conditioning on data using a probabilistically coherent prior.
- *Radical Subjectivism.* Any prior credences that satisfy the laws of probability may be permissibly held, and any posterior credences that are arrived at by conditioning on data using a probabilistically coherent prior may be permissibly held.

This last principle repudiates any "objectivist" element in inductive logic. It says that a person's reasoning cannot be criticized as long as her prior credences present a probabilistically coherent picture of the world and she updates using Bayes's Theorem. This rejects demands, by frequentists and objective Bayesians, to "solve" the problem of the priors by identifying criteria that portray certain coherent credence assignments as better than others. According to the radical subjectivist, there is no *problem* of the priors. It is simply wrong to think, as Jaynes does, that "consistency demands that two persons with the same relevant prior information should assign the same prior probabilities." Consistency only demands that priors obey the laws of probability: everything else is a matter of "inductive taste".

> Example. Pierre and Bruno have seen a coin tossed ten times and have observed seven heads. Neither has any other information about the coin's bias. Pierre begins with a uniform prior over the possible biases, and so concludes that the probability of a head on the next toss is $^2/_3$ (by the rule of succession). Bruno, for no particular reason, is certain that the bias is either $^1/_{10}$, $1/2$ or $^9/_{10}$, and, on a hunch, assigns these priors of 0.01, 0.01 and 0.98, respectively. He deduces that the probability of a head on the next toss is about 0.82.

For the radical subjectivist, there is no disparaging either Pierre or Bruno. Each began with a probabilistically coherent prior, and each arrived at his estimate by conditioning on the evidence received. So, each has reasoned perfectly. Bruno, of course, plays favorites while Pierre is evenhanded, but these are merely inductive predilections.

Since the only substantive constraints radical subjectivists impose on believers are those of probabilistic coherence and Bayesian updating, these requirements bear a lot of normative weight. Why should we accept them? If the assignment of priors is a matter of taste, why isn't it also a matter of taste, say, that one's credences for A and $\neg A$ cannot sum to more than one, or that one cannot update except by conditioning? Bayesians have offered a variety of justifications for both the probabilistic coherence and updating requirements. In the rest of this section we will discuss four rationales for coherence: Dutch book arguments, R. T. Cox's *a priori* derivation, rationales generated from qualitative constraints on credences, and accuracy based justifications. Reasons for updating using Bayes's Theorem will be discussed in §3.

2.2.1 Dutch Book Arguments

Dutch book arguments purport to show that having probabilistically incoherent credences will, of necessity, lead believers to make unwise decisions. This approach was pioneered by Ramsey in his [1931] and has been developed by many authors. Perhaps the most sophisticated presentation is found in De Finetti [1974]. De Finetti imagines an agent who announces a real number p_k for each member of an arbitrary set of propositions $A_1, A_2, \ldots, A_K$. For each k, the agent receives a prize of $S(p_k, v(A_k)) = 1 - (p_k - v(A_k))^2$ in units of something she values, where $v(A_k)$ is one or zero depending upon whether A_k is true or false. In effect, the agent is offered a choice among all wagers of the form

$$\text{Win } S(p_k, 1) = 2{\cdot}p_k - p_k^2 \text{ if } A_k; \text{ Win } S(p_k, 0) = 1 - p_k^2 \text{ if } \neg A_k$$

where she selects the value of p_k. Since any $p_k \in (0, 1)$ assures a positive outcome whether A_k is true or false, it will always be in the agent's interest to specify a vector $\langle p_k \rangle = \langle p_1, \ldots, p_K \rangle$. De Finetti calls these numbers her *previsions* for the A_k.

In conjunction with any logically consistent truth-value assignment v to the A_k, the act of specifying a prevision vector will produce a total prize of $\sum_k S(p_k, v(A_k)) = K - \sum_k (p_k - v(A_k))^2$. Notice that this prize decreases as a function of the distance between previsions and truth-values. Of course, some acts produce larger prizes than others, and the agent will try to secure the best prize possible. This is easy if she knows the truth-values, for she can simply announce these as her previsions and claim the maximum prize of K. When she is unsure about the truth-values it will typically not be in her interest to set each p_k to zero or one. While such a strategy offers the possibility of obtaining the maximum payoff, it also threatens to yield the minimum prize of 0. Depending on the agent's views about the chances that various truth-value assignments have of being actualized, she will usually be better off "hedging her bets" by selecting intermediate previsions that have high *estimated* payoffs. So, if A is the proposition that it will rain in Ann Arbor on June 11, 2020 then, depending on what the agent knows about the summer weather in Michigan, it might be best for her to announce previsions of $\langle 0.3, 0.7 \rangle$ for $\langle A, \neg A \rangle$. It is no sin against practical rationality, of course, if she does not end up receiving the largest possible prize or if someone else secures a larger prize because they happen to know more; one does the best one can given the information one has.

It would be a sin against rationality, however, if someone who knew no more than the believer could identify an act that was sure to secure a larger prize *no matter what truth-value assignment is actual*.

> Example. Suppose a believer chooses previsions of $\langle 0.3, 0.6 \rangle$ for $\langle A, \neg A \rangle$. Someone can then do better, *whether A is true or false*, by choosing $\langle 0.35, 0.65 \rangle$.

	A true, $v = \langle 1, 0 \rangle$	A false, $v = \langle 0, 1 \rangle$
Choose $\langle 0.35, 0.65 \rangle$	Prize = 1.155	Prize = 1.755
Choose $\langle 0.3, 0.6 \rangle$	Prize = 1.15	Prize = 1.75

Notice how the upper act *dominates* the lower act.

De Finitti, like Ramsey before him, saw a general principle in this. Say that a set of previsions $\langle p_k \rangle$ is *practically incoherent*[12] if there are alternative previsions $\langle q_k \rangle$ that dominate it the sense that $\text{Prize}(\langle p_k \rangle, v) < \text{Prize}(\langle q_k \rangle, v)$ for all logically consistent assignments v of truth-values to the propositions $A_1, \ldots, A_K$. De Finetti maintained that practical incoherence is a defect of rationality, and so imposed the following:

> *Requirement of Practical Coherence.* A rational agent will report practically coherent previsions. If she reports $\langle p_k \rangle$, then for each alternative set of previsions $\langle q_k \rangle$ there is a logically consistent truth-value assignment v with $\text{Prize}(\langle p_k \rangle, v) > \text{Prize}(\langle q_k \rangle, v)$.

The challenge is to determine the conditions under which previsions are practically rational.

The celebrated Dutch book theorem provides the answer. As in our example, it turns out that probabilistic incoherence is the hallmark of practical incoherence.

> Dutch Book Theorem (De Finetti's version): A set of previsions $\langle p_k \rangle$ for $\langle A_k \rangle$ is practically coherent if and only if there is a Boolean algebra Ω containing all the A_k and a finitely additive probability function P on Ω such that $P(A_k) = p_k$ for all k.[13]

In other words, announcing dominated previsions is equivalent to announcing previsions that violate the laws of probability.

This lovely piece of mathematics does not yet establish the requirement of probabilistic coherence since we do not yet know how an agent's previsions relate to her credences. To close the loop, defenders of Dutch book arguments maintain that previsions reveal credences directly.

> *Elicitation.* A practically rational agent will report previsions that coincide with her credences, so that $p_k = b(A_k)$ for all k.

Some early Bayesians, including Ramsey [1931] and de Finetti [1937], sought to justify this principle by arguing that talk of credences can only be scientifically respectable if degrees of belief are *operationally defined* as the previsions that believers in fact announce.[14] Later on, many Bayesians, including Savage [1971] and de Finetti [1974], sought to justify Elicitation by invoking the idea that practically

[12] Most authors just use the term "coherence" for both the practical and probabilistic requirements, but they are conceptually distinct ideas.

[13] For a version of the Dutch book theorem that yields countable additivity see [Skyrms, 1984].

[14] There are many ways of eliciting credences. Here are two: (i) $b(A)$ is the agent's fair price for a wager that pays one unit of utility if A and zero units if $\neg A$; (ii) $b(A) = l/(w + l)$ where $l, w > 0$ are any numbers such that the agent is indifferent between owning or covering a bet that pays w utiles if A and costs l utiles if $\neg A$. The Bayesian picture assumes that all of these methods will elicit identical credences. This is tantamount to assuming that rational agents act to maximize expected payoffs.

rational agents will always seek the highest *expected* payoffs by reporting previsions that maximize $Exp(\text{Prize}\langle p_k\rangle) = \sum_v b(v)\cdot[\sum_k S(p_k, v(A_k))]$, where v ranges over truth-value assignments. De Finetti was careful to choose S to be a *strictly proper scoring rule*, a rule for which the choice of $p_k = b(A_k)$ uniquely maximizes $Exp(\text{Prize}\langle p_k\rangle)$. So, on the assumption that practical rationality involves maximizing ones expected prize, it follows that rational believers will announce previsions that reveal their degrees of belief, and the Dutch book theorem then shows that these degrees of belief must be probabilistically coherent on pain of practical incoherence.

Two main types of objections are raised against Dutch book arguments. One relates to Elicitation. For this principle to be plausible it must be true that (a) prizes are paid out in some quantity that can be assigned units of *utility* that the agent values linearly and whose value does not depend on what previsions she might be asked to state, and (b) the agent assesses potential prizes on the basis of their *expected* utility. Point (b) is especially pressing because assessing a quantity on the basis of its expected value is equivalent to assigning probabilities to its possible values. This concern is ameliorated, to some extent, by the use of *representation theorems*, along the lines of Ramsey [1931] and Savage [1954], which aim to simultaneously derive both the requirements of probabilistic coherence and expected utility maximization from plausible constraints on rational preferences. Unfortunately, as a number of authors have noted, e.g., [Joyce, 1999], such representation theorems rely on strong structural assumptions about the richness of an agent's preferences that cannot be convincingly motivated as requirements of practical or epistemic rationality.

A second type of objection concerns the normative force of Dutch book arguments. Even if one agrees that having credences that sanction dominated choices is practically irrational, one might still wonder what specifically *epistemic* sin is committed. Some, notably Skyrms [1980], stress the penultimate sentence of Ramsey's famous statement of the Dutch book argument:

> These are the laws of probability, which we have proved to be necessarily true of any consistent set of degrees of belief. . . . If anyone's mental condition violated these laws, his choice would depend on the precise form in which the options were offered him, which would be absurd. He could have a book made against him by a cunning better and would then stand to lose in any event." [1931, p. 182]

The idea is that probabilistically incoherent credences are defective not so much because they leave agents open to sure losses, but because they cause agents to assess actions differently when viewed one-by-one than when viewed as a package. For example, an incoherent agent presented with choices $a_1 = [\text{Set } p_A \text{ to } 0.3 \text{ or } 0.35]$ and $a_2 = [\text{Set } p_{\neg A} \text{ to } 0.6 \text{ or } 0.65]$ might prefer 0.3 and 0.6, but then pick $\langle 0.35, 0.65\rangle$ when given choice $a_3 = [\text{Set } \langle p_A, p_{\neg A}\rangle \text{ to } \langle 0.3, 0.6\rangle \text{ or } \langle 0.35, 0.65\rangle]$. In a similar vein, Howson and Urbach [1989] and Christensen [1996] suggest that agents with probabilistically incoherent beliefs are committed to logically inconsistent value

judgments. To paraphrase Howson and Urbach (p. 57), to see p as the best prevision to report for A is to make a kind of intellectual value judgment, not to possess a disposition to accept particular bets when offered or to take particular actions when available. The problem with probabilistic incoherence is that it forces these value judgments to be inconsistent.

These arguments have had a mixed reception, and the normative significance of Dutch book arguments remain an active topic of philosophical controversy. For recent discussions see Joyce [1999; 2009], Hájek [2008], Howson [2008].

2.2.2 Cox's Theorem

Another influential argument for probabilistic coherence is found in Cox [1961]. Cox imagines a conditional credence function (a "plausibility") that maps each pair of propositions $A and C$ in an algebra Ω, to a number $b(A|C) \in [0, 1]$. He shows that any b that obeys four seemingly reasonable conditions is order-isomorphic to a probability. The first two conditions are:[15]

C1 If A and B are logically equivalent, then $b(\bullet|A) = b(\bullet|B)$ and $b(A|\bullet) = b(B|\bullet)$.

C2 $b(A\wedge B|C)$ is exclusively a function of $b(A|C)$ and $b(B|A\wedge C)$. More precisely, there exists a continuous binary operation $\otimes$ that is strictly increasing in each coordinate and such that $b(A \wedge B|C) = b(A|C) \otimes b(B|A \wedge C)$.

One example of such an operation is ordinary multiplication, which turns the equation into the usual definition of conditional probability.

Since conjunction is associative, the combination of C1 and C2 entail

(Assoc) $b(A\wedge B\wedge C|D) = (x\otimes y)\otimes z = x\otimes(y\otimes z)$ for $x = b(A|D), y = b(B|A\wedge D)$ and $z = b(C|A \wedge B \wedge D)$.

A function for which $F(F(x, y) = F(x, F(y, z))$ is called *associative*. Aczél [1966, p. 256] proves a theorem which has the consequence that any continuous, strictly increasing, associative function F defined everywhere on an interval $[a, b]^2$ is *order-isomorphic* to multiplication, i.e., there exists a strictly increasing, non-negative continuous function h defined on F's range with $h(F(x, y)) = h(x)\cdot h(y)$. So, if $\otimes$ is defined everywhere on $[0, 1]^2$, then there is an increasing, non negative continuous function m such that $m(b(A \wedge B|C)) = m(b(A|C))\cdot m(b(B|A \wedge C))$. In light of its continuity, we can ensure that $\otimes$ is defined everywhere on $[0, 1]^2$ by requiring b's range to be dense in $(0, 1)$:

C3 For any rational numbers $r_1, r_2, r_3 \in (0, 1)$ there are $A, B, C, D \in \Omega$ such that $r_1 = b(A|D), r_2 = b(B|A \wedge D)$ and $r_3 = b(C|A \wedge B \wedge D)$.

[15] I am not presenting Cox's conditions or result exactly as he expressed them. My presentation benefits from Halpern [1999], Paris [2004] and Jaynes [2003]. Also Cox's axioms are understood to hold for all elements of Ω subject to the proviso that propositions conditioned upon are never contradictory.

Under these conditions, it becomes possible to scale the function m so that $m(b(\top|C)) = 1$ and $m(b(\bot|C)) = 0$ for all C.

Completing the proof requires one further principle to govern negation:

C4 There is a continuous, non-negative non-increasing function $N : [0,1] \rightarrow [0,1]$ such that $b(\neg A|C) = N(b(A|C))$.

One example of such an operation is $N(b(A|C)) = 1 - b(A|C)$, a choice which would ensure that $b(A|C) + b(\neg A|C) = 1$, an important special case of the additivity law.

Given the existence of functions m and N with these properties, Cox establishes the following:

> *Cox's Theorem.* Given a credence function b on Ω that satisfies C1–C4 there exists a continuous, strictly increasing function $p : [0,1] \rightarrow [0,1]$ such that the composite mapping $p \circ b$ is a probability.

In other words, every belief function b that satisfies Cox's axioms is a member of a class of functions with domain [0, 1] that contains a unique probability as well as every continuous, increasing function of that probability which preserves endpoints. A rational belief function, in other words, is always *order-isomorphic* to a probability.

This is not the same as saying that b *is* a probability. For all Cox shows, it might be that b^k for some probability P and any $k > 0$. In light of this one might wonder how Cox's result serves to justify probabilistic coherence. The usual answer is to say that there is no substantive difference between representing a belief function using a probability or using any other function that is order-isomorphic to that probability. Suppose, say, that we represented beliefs not with probability functions but with their squares, so that b^2. The laws of rational belief would seem different. Instead of being additive, "credences" would be "square-additive," i.e., they would obey the law $b(A \vee B) = b(A) + b(B) + 2(b(A) \cdot b(B))^{1/2}$ for contraries A and B. But, the argument continues, this difference between this law and additivity is superficial: in terms of real constraints imposed on rational beliefs, the two are identical. Instead of saying that credences must be additive, we now say that their square roots must be additive; instead of saying that the credences of A and $\neg A$ must sum to one, we say that their square roots must sum to one; and so on. The form of expression is different, but the content is the same. To make this methodological stance explicit, let's add a principle to Cox's argument:

C5 If two functions $f, g : [0,1] \rightarrow [0,1]$ with $f(0) = g(0)$ and $f(0) = g(1)$ are order-isomorphic (i.e., if each is a continuous strictly increasing function of the other), then the two provide equivalent representations of belief states.

With C5 in place there is no harm in saying that credences must be probabilities, provided that one understands that this only means that they must be order-isomorphic to probabilities. Cox proves no more.

Cox's theorem has a "beauty is in the eyes of the beholder" quality. Some, e.g., Jaynes [2003] and Howson [2008], take it to decisively justify probabilistic coherence. Others, e.g., Halpern, [1999] think it delivers less than advertised because it ignores credence functions that fail to cover a dense subset of [0, 1]. In the end, the issue comes down to how compelling one finds the axioms. While they seem innocuous on a first reading, C1–C5 do impose substantial constraints on rational credences. Indeed, it is hard to see how to justify C2 or C4 without appealing to explicitly probabilistic considerations. Consider, for example, Jaynes's [2003, p. 24-25] justification of C2.

> For $A \wedge B$ to be a true proposition, it is necessary that A is true. Thus, $[b(A|C)]$ should be involved [as a component of $b(A \wedge B|C)$]. In addition, if A is true it is further necessary that B should be true, so $[b(B|A \wedge C)]$ is also needed. But, if A is false, then of course $A \wedge B$ is false independently of whatever one knows about $B\ldots$ so if the [agent] first reasons about A, then the plausibility of B will only be relevant if A is true. Thus, if the [agent] has $b(A|C)$ and $b(B|A \wedge C)$ [she] will not need $b(B|C)$.

Similarly Howson [2008, p. 18] writes:

> Why should [C2] be the case? Well, knowing how likely B would be if I could assume B was true will not of course tell me how likely B is; for that I would also need to know how likely B is. But once I know that then it seems that I should know, at any rate in principle, how likely both B and A are. Nothing in this piece of informal conditional reasoning depends on any scale of measurement.

This is right as far as it goes, but misleading. Given that it is being imposed in a context where C1, C3 and C5 hold, it is unclear how C2 differs from saying outright that $b(A \wedge B|C)$ is directly proportional to $b(A|C)$ and $b(B|A \wedge C)$. After all, we know that a continuous, non-negative function whose domain is dense in $[0, 1]^2$ is associative if and only if it is order-isomorphic to multiplication, and we have accepted the idea that there is no meaningful difference between a probability and its order-isomorphisms. In such a context, saying $b(A \wedge B|C)$ is a function of $b(A|C)$ and $b(B|A \wedge C)$ is no different from saying that, up to order-isomorphism, $b(A \wedge B|C)$ is $b(A|C) \cdot b(B|A \wedge C)$. And, if this is all we are saying, then it seems that we have given up on trying to *justify* the laws of probability from more basic principles: we are imposing the laws directly, up to order-isomorphism.

The situation is the same for C4. In the presence of the other requirements, if we insist that $b(\neg A|C)$ be a non-decreasing continuous function of $b(A|C)$, then we are stipulating that $b(\neg A|\ C)$ is order-isomorphic to $1 - b(A|C)$. Again, in the presence of C5, this seems no different from an overt invocation of the probabilistic requirement that $b(A|C)$ and $b(\neg A|C)$ must sum to one.

Overall then, while Cox's result offers an illuminating way of rewriting the requirement of probabilistic coherence, it is not clear how much it provides in the way of an independent justification for that requirement.

2.2.3 Quantitative Probability from Qualitative Probability

Another approach begins by assuming that believers make *comparative* probability judgments that can be represented by orderings over an algebra Ω. Intuitively, $A. \geq . B$ means that the believer is at least as confident in A as she is in B. Relations of strict confidence, $. >.$, and equi-confidence, $. =.$, are defined as one would expect. Such comparative probability rankings are coherent when they conform to the judgments that some probability function would sanction, i.e., when there exists a probability on Ω such that $A. \geq . B$ only if $P(A) \geq P(B)$. The challenge is to determine the conditions under which the ordering can be represented in this way, and to decide whether these conditions are requirements of rationality.

De Finetti and others initially thought that the following *laws of comparative probability* would be both necessary and sufficient for probabilistic representability:

CP_1 (*Normalization*): $\top . \geq . A. \geq . \bot$, and $\top . > . \bot$.

CP_2 (*Completeness*): $A. \geq . B$ or $B. \geq . A$.

CP_3 (*Transitivity*): If $A. \geq . B$ and $B. \geq . C$, then $A. \geq . C$.

CP_4 (*Quasi-additivity*): If C is incompatible with both A and B, then $A. \geq . B$ if and only if $A \vee C. \geq . B \vee C$.

CP_5 (*Archimedean*): If $A. > . \top$ for every A in some subset A of Ω, then A is countable. In addition, if $A. = . B$, for all $A, B \in A$, then A is finite.

CP_6 (*Continuity*): If $\{A_1, A_2, A_3, ...\}$ is a countable set of mutually incompatible events in Ω and if $B. \geq . (A_1 \vee ... \vee A_N). \geq . C$ for all n, then $B. \geq . \vee_n A_n. \geq . C$.

Unfortunately, Kraft, *et al.* [1959] exhibited a finite ranking that obeys CP_1–CP_6 and yet cannot be ordinally represented by any probability function.

There are two ways to go at this point. One can seek sufficient but non-necessary conditions for probabilistic representability by restricting one's attention to orderings defined over rich structures, or one can try to identify stronger necessary conditions that end up being jointly sufficient. The best result of the first type is found in Villegas [1964]. Building on earlier work of Koopman [1940] and Savage [1954], Villegas showed that CP_1–CP_6 suffice to determine a unique countably additive probability representation when $(\Omega, . \geq .)$ is *atomless*, i.e., when for each $A. > . \top$ there are disjoint A_1 and A_2 with $A = A_1 \vee A_2$ and $A_1, A_2. > . \top$. Important results of the second type are found in Scott [1964] and Suppes and Zanotti [1976]. Scott, who was inspired by Kraft *et al.* [1959], imposes the following condition (which makes a number of the others redundant):

> CP_7 (Scott): If $\langle A_1, A_2, ..., A_N \rangle$ and $\langle B_1, B_2, ..., B_N \rangle$ are sequences of propositions from Ω, possibly with repeats, that contain the same number of truths as a matter of logic, then $A_n. \geq . B_n$ for all $n \leq N$ only if $B_n. \geq . A_n$ for all $n \leq N$.

Suppes and Zanotti [1976] impose a somewhat different requirement, and both show that their axioms are necessary and sufficient for (non-unique) probabilistic representation.[16]

Of course, the significance of these results depends what justifications can be given for the various constraints imposed on comparative probability rankings. Some, e.g. Savage [1954], seek to derive the constraints from decision-theoretic considerations. On this approach, $A. \geq . B$ is taken to mean that the believer prefers a prospect in which she stands to enjoy a prize if A and to suffer a specific if $\neg A$ to a prospect in which she stands to enjoy the prize if B and to suffer the penalty if $\neg B$. Constraints on $. \geq .$ are then shown to follow from allegedly reasonable principles of rational preference. An alternative strategy is to argue that the constraints capture something central to rational belief. Joyce [1998], for example, has argued for Scott's Axiom on the grounds that (a) a person's credences commit her to making estimates of truth-values, (b) in particular, given any sequence of propositions $\langle A_1, A_2, ..., A_N \rangle$, a person with the sharp credences $b(A_1), ..., b(A_N)$ is committed to using $b(A_1) + ... + b(A_N)$ as her estimate of the number of truths among the A_n, and (c) under these conditions Scott's Axiom is a straightforward non-dominance principle that forbids believers from making different estimates for the number of truths among the A_n and among the B_n when, as a matter of logic, these numbers must be the same.

2.2.4 Nonpragmatic Vindications of Probabilism

A fourth sort of justification for probabilistic coherence is proposed in Shimony [1988] and van Fraassen [1988], and developed in Joyce [1998; 2009]. The driving idea behind these "nonpragmatic vindications" is that because credences sanction estimates of epistemically salient quantities – truth-values, objective chances, frequencies – one can evaluate the quality of a person's credences in terms of the accuracy of the estimates they sanction. Credences that sanction estimates that are necessarily less accurate than they need to be are epistemically defective.

This approach has affinities to de Finetti's prevision-based Dutch book argument, which explicitly rewards accurate truth-value estimates. But, whereas de Finetti's approach rests on Elicitation, this method aims to assess the accuracy of credences without the mediation of desires or choices. Following Joyce [2009], one can think of these assessments as codified in an *epistemic scoring rule* $S(b, v)$ that maps each credence function b and truth-value assignment v into a non-negative number $S(b, v)$ that measures the *epistemic disutility*[17] of having credences b when the truth-values are given by v. S's values should reflect the sorts of traits that make beliefs worth holding from an epistemic perspective. One such trait, indeed the cardinal one, is *accuracy*. S should produce lower (= better) values when b

[16] To secure uniqueness either atomlessness or the existence of a set of atoms $\langle A_1, A_2, ..., A_N \rangle$ with $A_m . \geq . A_n$ for all m, n is required

[17] I use epistemic *dis*utility rather than epistemic utility so as to more easily relate this investigation to the work on proper scoring rules in statistics and economics.

sanctions accurate estimates of epistemically salient quantities than when it sanction inaccurate estimates of those quantities.

Once an acceptable epistemic scoring rule S is identified, the objective of a nonpragmatic vindication of probabilism is to show that probabilistically incoherent credences sanction estimates of epistemically important quantities that are S-dominated whereas coherent credences never have this undesirable feature. This is supposed to establish that incoherent credences are defective from an epistemic perspective because they fare less well, in terms of epistemic utility, than some alternative set of credences *no matter what the world is like*.

Shimony and van Fraassen feel that credences are best assessed by looking at the degree to which they generate *well-calibrated* frequency estimates. The laws of probability declare that a coherent believer's estimate for the proportion of truths in a set $\mathcal{A}$ must be $\sum_A b(A)/\#(\mathcal{A})$ where A ranges over the propositions in $\mathcal{A}$ and $\#(\mathcal{A})$ is $\mathcal{A}$'s cardinality. When trying to justify probabilistic coherence, however, one cannot construe estimates as expectations without begging the question. One must begin instead from the thought that, whether coherent or not, b sanctions x as the estimate of the relative frequency of truths in any finite set of propositions all of which have credence x. In other words, setting $b(A) = x$ commits one to estimating that propositions with A's credence are likely to be true x proportion of the time. In light of this, one can partition any finite set of propositions $\mathcal{A}$ into subsets $\mathcal{A}_x = \{A \in \mathcal{A} : b(A) = x\}$, and can measure the accuracy of b's truth-frequency estimate for $\mathcal{A}_x$ relative to a truth-value assignment v using the quantity $C_x(b, v) = (F_v(\mathcal{A}_x) - x)^2$, where $F_v(\mathcal{A}_x)$ is the proportion of $\mathcal{A}_x$'s elements that are true in the assignment v. One can then obtain an overall measure of the accuracy of b's frequency estimates using the *calibration score* $Cal(b, v) = \sum_x [\#(\mathcal{A}_x)/\#(\mathcal{A})] \cdot C_x(b, v)$. This is minimized when all the propositions with credence 1 are true, three-quarters of the propositions with credence $^3/_4$ are true, two thirds of the propositions with credence $^2/_3$ are true, and so on.

By invoking some structural assumptions, Shimony and van Fraassen are able to show, in slightly different ways, that incoherent credences sanction poorly calibrated frequency estimates. In particular, for any incoherent credence function b there is a coherent credence function c such that $Cal(b, v) > Cal(c, v)$ for all logically possible truth-value assignments v. To the extent that one believes that calibration is a reasonable measure of epistemic accuracy (and accepts the structural assumptions), this shows that incoherent credences are inherently defective: they lead to frequency estimates that are, as a matter of necessity, farther from the actual frequencies than they need to be.

The problem with this argument, as pointed out in Seidenfeld [1985] and Joyce [1998] is that calibration is a poor measure of epistemic accuracy. There are a variety of problems, but the decisive one is that a believer can, in some circumstances, improve her calibration score, relative to a given truth-value assignment, by decreasing her credence in every truth and increasing her credence in every falsehood. Despite the fact that the person seems to have made herself "more wrong" about every single proposition, *Cal* says that she has improved accuracy overall. This

happens, in part, because a credence function affects *Cal* both by fixing frequency estimates for the various $\mathcal{A}_x$ but also by determining the composition of the $\mathcal{A}_x$.

To avoid this sort of thing, Joyce [1998] advocates focusing on the role of credences in estimating truth-values rather than frequencies. A clear requirement is then:

> *Truth-Directedness.* If b's credences are uniformly closer than c's are to the truth-values in v, then $S(c, v) > S(b, v)$.

One rule like this is the *Brier score*: $Brier(b, v) = [1/\#(A)]\cdot\sum_A(v(A) - b(A))^2$. This score, Brier [1950], has been used by meteorologists to evaluate the accuracy of probabilistic weather forecasts, and de Finetti relied on it to elicit previsions in his version of the Dutch Book argument. In addition to being truth-directed, the Brier score has a variety of features that make it an excellent measure of epistemic accuracy, including the following:

> *Extensionality.* $S(b, v)$ depends only on the credences in b and the truth-values in v.
>
> *Normality.* $S(b, v)$ depends only on the absolute differences between the credences in b and the truth-values in v.
>
> *Convexity.* $\lambda\cdot S(b, v) + (1 - \lambda)\cdot S(c, v) > S(\lambda\cdot b + (1 - \lambda)\cdot c, v)$ for all $\lambda \in (0, 1)$.
>
> *Symmetry.* If $S(b, v) = S(c, v)$, then $S(\lambda \cdot b + (1 - \lambda) \cdot c, v) = S((1 - \lambda)\cdot b + \lambda\cdot c, v)$ for all $\lambda \in (0, 1)$.
>
> *Propriety.* S is a *proper* scoring rule: if b is coherent, then the minimum expected Brier score is uniquely attained by b itself when expected values are computed using b.
>
> *Coherent Admissibility.* If b is coherent, then for any alternative c there is a truth-value assignment v such that $S(c, v) > S(b, v)$.

In addition, the Brier score can be decomposed, see Murphy [1973], into a sum of the calibration score and another epistemically significant quantity called the *discrimination index*, which measures to degree to which the credences in b discriminate truths from falsehoods.

Joyce [1998], inspired by both de Finetti's Dutch book argument and the Shimony/van Fraassen approach, tries to improve on these results. He imposes a series of constraints on epistemic scoring rules — including Truth-Directedness, Extensionality, Normality, Symmetry and Convexity — and argues that the property of being S-dominated relative to any rule S which meets the requirements is an epistemic defect.[18] Joyce then proves that incoherent credences are S-dominated relative to any rule that satisfies the constraints.

[18] More precisely, the claim is that b is defective when, for each scoring rule S than satisfies the constraints, there is a credence function b_S such that $S(b, v) > S(b_S, v)$ for every logically consistent truth-value assignment v.

Maher [2002] criticizes some of Joyce's constraints, especially convexity, by claiming that the non-convex, improper *absolute value score* $Abs(b, v) = [1/\#(A)] \cdot \sum_A |v(A) - b(A)|$ is the best measure of epistemic disutility. Joyce (2009) rejects this on the basis of the following sort of case:

> Example. A fair three-sided die will be tossed. Consider credence functions $b = (^1/_3, ^1/_3, ^1/_3)$ and $c = (0, 0, 0)$. Calculation shows $Abs(b, (1, 0, 0)) = Abs(b, (0, 1, 0)) = Abs(b, (0, 0, 1)) = {}^4/_9$ and $Abs(c, (1, 0, 0)) = Abs(c, (0, 1, 0)) = Abs(c, (0, 0, 1)) = {}^1/_3$. The logically inconsistent assignment c thus dominates the assignment b.

The problem here is not the assignment b, but the rule Abs, which does not do enough to penalize error and so portrays the obviously correct credence assignment (the one that agrees with the known objective chances), as less accurate than the *logically inconsistent* assignment that assigns credence zero to each of three possibilities that are known to be collectively exhaustive.

Gibbard [2008] also disputes some of Joyce's constraints, notably Symmetry and Normality, and argues that scoring rules can only be useful for purposes of guiding action if they are proper. He thus wonders whether a Joyce-styled argument for probabilism can be mounted on the basis of Propriety. It turns out that Lindley [1982] had already provided an argument of this sort, albeit not billed as such. Lindley imagines a believer who assigns credences $\langle b_n \rangle$ to members of a finite partition. These credences are scored using a rule S with these features:

- Truth-Directedness
- Additive form: $S(b, v) = \Sigma_n \lambda_n \cdot s_n(b_n, v_n)$.
- $s_n(b, 1)$ and $s_n(b, 0)$ are defined and have continuous first derivatives on $[0, 1]$.
- $\frac{d}{db} s_n(b, 0) / \frac{d}{db} s_n(b, 1)$ approaches 0 when b approaches 0 from above.
- $\frac{d}{db} s_n(b, 1) / \frac{d}{db} s_n(b, 0)$ approaches 0 when b approaches 1 from below.

Given these assumptions, Lindley proved the following:

> Lindley's Theorem: $\langle b_n \rangle$ is undominated relative to S only if the numbers
>
> $$L(b_n) = \frac{d}{db}\mathbf{s}_n(b_n, 0) / [\frac{d}{db}\mathbf{s}_n(b_n, 0) - \frac{d}{db}\mathbf{s}_n(b_n, 1)]$$
>
> obey the laws of probability. Additionally, if the map L is one-to-one, then the $L(b_n)$ obey the laws of probability only if $\langle b_n \rangle$ is undominated.

So, every undominated credence function has a "known transform" into a probability, and if this transformation is one-one, then any set of credences with a coherent transform is undominated. Lindley then observes, in passing, that if S is *proper* then $b_n = L(b_n)$ for each n and the transform is one-one. Thus, for

proper rules $\langle b_n \rangle$ is incoherent (coherent) if and only if there is a (is no) credence function c_S such that $S(b, v) \geq S(c_S, v)$ for all truth-value assignments v with the inequality holding strictly for at least one v.

This is a lovely result, and very close to what Gibbard hoped for, but it does not give us quite everything. First, it does not guarantee that incoherent credences are *strictly* dominated, since it might be that $S(b, v) = S(c_S, v)$ for all v. This issue has been largely resolved by Leib *et al.*, (manuscript) who derive the stronger conclusion from similar assumptions. Second, and more important, the requirement be a S be proper scoring rule is not appropriate in contexts where the aim is to provide a vindication of probabilistic coherence that might convince non-Bayesians. Propriety makes sense on the assumption that rationality requires maximizing *expected* epistemic utility, but, as has already been emphasized, the concept of an expectation only makes sense when expectations are generated by probabilities.

Joyce [2009] shows how to use Coherent Admissibility, a weakening of Propriety that invokes only dominance considerations, to obtain the desired result. Specifically, it is possible to prove:

> THEOREM:[19] If S satisfies Truth-Directedness and Coherent Admissibility, and is bounded and continuous for all credence functions and truth-value assignments, then
>
> - For every incoherent credence function b defined on a finite partition of propositions there is a coherent credence function c_S such that $S(b, v) \geq S(b_S, v)$ for all logically consistent truth-value assignment v, with at least one inequality being strict.
> - No coherent credence function is ever S-dominated in this way.

While this is close to what one would want, it would be desirable to remove the restriction to partitions, to relax the requirement that S be bounded, and to find a way of deriving Coherent Admissibility from more basic principles.

For further critical discussion of results like these see [Joyce, 2009; Hájek, 2008].

2.3 Is Probabilistic Coherence Enough?

Clearly, there is no shortage of proposed justifications for probabilistic coherence. Even if we are entirely convinced by one (or more) of them, however, subjective Bayesianism remains a "garbage-in-garbage-out" theory. Coherent beliefs can be absurd, but subjective Bayesians tolerate them no matter how crazy they get, (short of violating the laws of probability). Many find this objectionable. Alan Chalmers [1999, p. 188] expresses the worry nicely:

[19] Joyce [2009] actually invokes a weaker version of Coherent Admissibility in which the "$>$" is replaced by "$\geq$". This, however, was a mistake since the proof makes use of the stronger version.

> Once we take probabilities as subjective degrees of belief... a range of unfortunate consequences follow. The Bayesian calculus is portrayed as an objective mode of inference that serves to transform prior probabilities into posterior probabilities in light of given evidence. Once we see things this way, it follows that any disagreements in science must have their source in the prior probabilities held by the scientists. But these prior probabilities themselves are totally subjective and not subject to critical analysis. Consequently, those of us who raise questions about the relative merits of competing theories... will not have our questions answered by the subjective Bayesian, unless we are satisfied with an answer that refers to the beliefs that individual scientists just happen to have started out with.

Hard-core subjectivists, like de Finetti, will reject such concerns out of hand. "It's too bad," they will say, "but all we have to go on in science are 'the beliefs that individual scientists just happen to have started out with'. To think otherwise is to pretend that you have access to information you cannot possibly possess." Other Bayesians will suggest that things are not nearly so bad a Chalmers implies since (a) disagreements in science should ultimately be resolved by acquiring further evidence, and (b) as more evidence is acquired disagreements become increasingly less severe and less frequent. The next section explains how this "tempered" Bayesian response to the problem of the priors is supposed to work.

2.4 Tempered Bayesianism and "Washing Out"

Tempered Bayesians maintain that prior opinion will tend to "wash out" as believers acquire more and more information. Here is a famous statement of the view from Edwards *et al.* [1963, p. 201], one of the classic papers on the topic:

> If observations are precise... then the form and properties of the prior distribution have negligible influence on the posterior distribution. From a practical point of view, then, the untrammeled subjectivity of opinion... ceases to apply as soon as much data becomes available. More generally, two people with widely divergent prior opinions but reasonably open minds will be forced into arbitrarily close agreement about future observations by a sufficient amount of data.

This "merger of opinion" is what passes for objectivity in the tempered Bayesian's worldview, which sees objectivity as *intersubjective agreement in the long run.*

The basis of such claims is found in the "washing out" theorems, which purport to show that believers who start with widely divergent, but coherent, priors and who update on the same data streams will eventually converge on a "consensus posterior." Since the structure of all the washing out theorems is similar, we will consider only the most general version, which relies on the *Doob Martingale Convergence Theorem*, Doob [1971].

Consider two believers with priors b and c such that $0 < b(h), c(h) < 0$ for some hypothesis h. These believers will undergo a potentially infinite sequence of learning experiences involving the random variables $X_1, X_2, X_3, \ldots$, each of which may take finitely many values. Assume that:

a. Neither subject is closed-minded about the data: both b and c assign each finite data sequence $d_j = (X_1 = x_1) \wedge (X_2 = x_2) \wedge \ldots \wedge (X_j = x_j)$ an intermediate probability.

b. At time-j each subject learns the actual value of X_j.

c. Both subjects know they will condition on the evidence they receive, i.e., they know that, for each time $j, b_j(h) = b(h|d_j)$ and $c_j(h) = c(h|d_j)$.

(a)–(c) entail that each subject's credences form a *martingale sequence* in which every term is the expected value of its successor: $b_j(h) = \Sigma_x b_j(b_{j+1}(h) = x) \cdot x$ and $c_j(h) = \Sigma_x c_j(c_{j+1}(h) = x) \cdot x$. Doob's Theorem says that, with probability one, these sequences converge to a definite limit.

To show that these limits coincide, additional assumptions are required. Here are two possibilities:

- Each possible infinite data stream determines truth-value $v \in \{0, 1\}$ for h.
- Each possible infinite data stream determines an *objective chance* $p \in [0, 1]$ for h.

If first assumption holds, $\lim_j b_j(h) = \lim_j c_j(h) = v$. If the second holds, and if the subjects align their credences with their best estimates of their objective chances (of which more below), then $\lim_j b_j(h) = \lim_j c_j(h) = p$. Success?

Unfortunately not. Washing out theorems that rely on either of the above assumptions will not assuage worries about subjectivity. Here prior opinion "washes out" only because the data is so incredibly informative in the limit that the subject's prior beliefs are *irrelevant* to her final view as a matter of logic. In the chance case, for example, each leaning experience might reveal successive digits of the hypothesis's objective chance. Priors will then wash out, of course, but only because they play no real role in posterior at all. If all learning situations were like this, there would be no call for priors at all, at least in the limit.

To obtain washing out theorems that might have a chance at quelling worries about subjectivism, we must start from weaker data and derive convergence from commonalities among the priors alone. One way to do this, pioneered by Savage [1954, p. 46-50] is to suppose that the subjects agree that the data statements are *statistically independent and identically distributed* (IID) conditional on both h and $\neg h$.

- If d_j is any possible data stream, and x is any possible value of X_k with $k > j$, then $b(X_k = x|h) = b(X_k = x|h \wedge d_j)$ and $b(X_k = x|\neg h) = b(X_k = x|\neg h \wedge d_j)$. The same holds with b replaced by c.

Under this assumption, it can be shown that $b_j(h)$ and $c_j(h)$ will converge to a common value with probability one according to both b and c.

Many Bayesians see this as an answer to the concerns that people like Chalmers have raised about the "anything goes" aspect of subjective Bayesianism. The case of repeated IID trials is a common one, and the result shows that subjects who believe they are watching a sequence of such trials will eventually come to agree to a greater and greater extent.

> Example. Neither Pierre not Bruno knows anything about the bias of a coin. Pierre begins with a uniform prior over [0, 1], while Bruno starts with a sharply peaked normal distribution, with mean $^1/_{10}$ and variance $^1/_{100}$. After fifteen tosses in which ten heads appear, the two will still be far apart in their estimates of the bias of the coin. But, after 3,000 tosses in which 2,000 heads have appeared both will have arrived at posterior distributions that differ only in very distant decimal places.

While this is comforting, it is important to recognize that the theorem requires a great deal of initial agreement. All parties must concur about the possible hypotheses. If, e.g., Bruno does not spread his prior over *all* of $[0, 1]$, thereby excluding some biases from consideration, convergence with Pierre is no longer assured. Likewise, all parties must agree about the possible data sequences, and they must all know that they will condition on whatever data they receive. Believers who disagree about these things will *not* tend toward consensus as evidence accumulates simply because they will not agree about what counts as evidence. If, e.g., after seeing 3,000 straight heads Bruno concludes with certainty that the coin is two headed and ignores all future data (a conclusion not sanctioned by conditioning), then the convergence result does not apply. If either party does not believe he or she is observing an IID process, then the result does not apply. While these restrictions do not render the washing out theorems irrelevant to questions about subjectivism, they underscore a significant limitation. The washing out results only secure agreement in the limit by assuming substantial amounts of agreement at the start. It's a case of *no agreement in, no agreement out*!

3 INDUCTIVE INFERENCE AS UPDATING SUBJECTIVE PROBABILITY

Whatever their views about the status of prior probabilities, all Bayesians see inductive inference as a matter of updating probabilities in light of new evidence. Abstractly, this process involves a probability P that represents some prior state of evidence, a learning experience λ that imposes constraints on a posterior probability, and an update rule, $P - \lambda \rightarrow P_\lambda$, which maps the prior to a unique posterior satisfying the constraints. As before, we think of the prior as derived from an unconditional probability distribution over a partition of hypotheses $\mathcal{H}$, and a family of normalized likelihood functions $L_x\mathcal{H} \rightarrow [0, \infty)$, where $L_x(h) = P(x|h)$ is the probability of observing datum $x \in \mathcal{X}$ if $h \in \mathcal{H}$ holds. A learning experience λ

conveys information which requires that this prior be supplanted by a posterior P_λ that satisfies certain constraints. In most cases, these constraints can be represented by a set of equations $Exp_\lambda(f_k) = \sum_{h,x} P(h \wedge x) \cdot f_k(h \wedge x) = c_k$ where each f_k is a random variable on $\mathcal{H} \wedge$ X and each c_k is a real number.

Here are some possible constraints (with $\mathcal{H} \wedge \mathcal{X}$ assumed finite):

a. $\{P_\lambda(x_2) = 1\}$

b. $\{P_\lambda(x_1) = 0.2, P_\lambda(x_2) = 0.3, P_\lambda(x_3) = 0.5, P_\lambda(x_m) = 0 \text{ for } n > 3\}$

c. $\{P_\lambda(x_1) = 2{\cdot}P(x_1), P_\lambda(x_2) = {}^1\!/_2{\cdot}P(x_2), P_\lambda(x_3) = 5{\cdot}P(x_3), P_\lambda(x_n) = 0{\cdot}P(x_3)$ for $n > 3\}$

In (a) experience delivers the verdict that x_2 is certainly true. In (b) it directly sets probabilities for the various data statements. In (c) new probabilities for the data statements are specified as a function of their priors.

Notice that (b) and (c) impose identical constraints when $P(x_1) = 0.1, P(x_2) = 0.6, P(x_3) = 0.1$ and $\sum_{n>3} P(x_n) = 0.2$. Even so, it is crucial to understand that they describe different learning experiences. On might call (b) a *hard* experience because it ignores the prior and resets each $P_\lambda(x_n)$ *de novo*, thereby requiring the posterior to satisfy $\{P_\lambda(x_1) = 0.2, P_\lambda(x_2) = 0.3, P_\lambda(x_3) = 0.5\}$ *for any prior*. In contrast, (c) fixes posterior probabilities indirectly via a specification of *Bayes updating factors* $\beta_P(\cdot, \lambda) = P_\lambda(\cdot)/P(\cdot)$ for all x_n. As a result, (c) and other *soft* experiences, impose constraints on posteriors *whose effects vary with changes in the priors*.

As these examples illustrate, experiences constrain posterior probabilities incompletely. The role of the update rule is to select that probability, from among the many that satisfy the constraints, which best preserves the information encoded in the prior. This process is governed by a kind of "minimal change" ethos which prohibits the posterior from introducing distinctions in probability among hypotheses that are not already inherent in the prior or explicitly mandated by the new evidence. There should be no "jumping to conclusions."

3.1 Simple Bayesian Conditioning

In the simplest learning experiences a person comes to know some $X \subset \mathcal{X}$ with certainty. If this is *all* she learns, her experience will not distinguish among possibilities that entail X, and the relevant minimal change principle requires that X's probability be raised to $P_\lambda(X) = 1$ in a way that does not disturb probability ratios among propositions that entail X.

> MC_1: If a learning experience's effect on a posterior is confined to fixing $P_\lambda(X) = 1$, then ratios of probabilities among propositions that entail X should remain fixed, so that $P_\lambda(A \wedge X)/P_\lambda(B \wedge X) = P(A \wedge X)/P(B \wedge X)$ for $A, B \in \mathcal{H} \wedge \mathcal{X}$.

This means that the absolute Bayes updating factor for propositions that entail X is $\beta_P(A, \lambda) = 1/P(X)$, and the relative Bayes updating factor is constant: $\beta_P(A, \lambda)/\beta_P(B, \lambda) = 1$.

It should come a no surprise that the only probability that fits the bill is $P_\lambda(\bullet) = P(\bullet|X)$. In this way, we secure the most basic principle of Bayesian inductive logic:

> *Updating by Conditioning.* Suppose a person's prior epistemic state is represented by a prior P that is not dogmatic about X, so that $1 > P(X) > 0$. If the person has a learning experience λ in which she receives the information that X is certainly true, and if this is all she learns, her posterior should be $P_\lambda(\bullet) = P(\bullet|X)$.

As we have already seen, conditional probabilities have features that make them ideal for modeling inductive learning. For example, $P_\lambda(\bullet) = P(\bullet|X)$ is a probability that is dogmatic about X. In addition, the temporal order in which data is acquired is irrelevant to its evidential import.

3.2 Jeffrey Conditioning

The weakness of Bayesian conditioning as an update rule is that it only applies to learning experiences that raise the probability of some proposition to one. This dogmatic aspect of the process can seem implausible, especially given that propositions learned via conditioning cannot be unlearned by subsequent conditioning. Richard Jeffrey, [1983], a forceful critic of dogmatist epistemologies, has advocated a model of learning that allows experiences to alter probabilities without raising them to certainty or lowering them to zero.

> Example. Consider a wine drinker who orders the house red with the expectation that it is a cabernet. Taking a sip, he has a gustatory experience that causes him to doubt his judgment. This experience might move his subjective probabilities from a prior in which $P(\text{cabernet}) = 0.95$ to a posterior in which $P_\lambda(\text{cabernet}) = 0.6$. One might try to explain this change by positing a proposition, say $x =$ "The wine has a fruity taste", that the drinker comes to learn with certainty during the experience, and to suppose his priors were such that $P(\text{cabernet}\,|x) = 0.6$. While this might make sense for sophisticated oenophiles, it is implausible for most people. If asked, the drinker might not be able to articulate any specific feature of the wine that leads him to alter his views, and even if he comes up with something vague — "It doesn't taste like other cabs" — it is even less plausible that he will have the required prior conditional probabilities. A better model, Jeffrey suggests, portrays the change as being a direct effect of the experience, without the mediation of any knowledge gained with certainty. The gustatory experience causes the drinker to move directly from $P(\text{cabernet}) = 0.95$ to $P_\lambda(\text{cabernet}) = 0.6$. If this is its only immediate effect, Jeffrey argues, then probabilities of events conditional

on {cabernet, not-cabernet} should remain fixed, so that $P_\lambda(\cdot) = 0.6{\cdot}P(\cdot|$ cabernet) $+0.4{\cdot}P(\cdot|\neg$cabernet).

More generally, represent a *Jeffrey learning* experience on $\mathcal{X}$ is a countable vector of ordered pairs $\lambda = \langle x_n, \lambda_n\rangle$ where $1 > P(x_n) > 0$ for all n and $\lambda_1, \lambda_2, \ldots$ is a sequence of non-negative real numbers summing to one. In Jeffrey's picture, the learning experience directly fixes each λ_n as the posterior for x_n. An experience with this as its *only* immediate effect will introduce no new distinctions in probability among propositions that entail the same x_n. This leads to the following minimal change principle:

> MC_2: If a Jeffrey learning experience λ's effect on a posterior is confined to fixing probabilities for elements of $\mathcal{X}$, then the ratios of probabilities of events that entail each x_n should remain fixed, so that $P_\lambda(A \wedge x_n)/P_\lambda(B \wedge x_n) = P_\lambda(A \wedge x_n)/\ P_\lambda(B \wedge x_n)$ for $A, B \in \mathcal{H} \wedge \mathcal{X}$.

The only probability that obeys MC_2 subject to the constraint that $P_\lambda(x_n) = \lambda_n$ for each n is the *Jeffrey shift*: $P_\lambda(\bullet) = P(\bullet|\langle x_n, \lambda_n\rangle) = \sum_n \lambda_n{\cdot}P(\cdot|x_n)$. Despite the notation, one should not think of $\lambda = \langle x_n, \lambda_n\rangle$ as a proposition or of $P(\bullet|\langle x_n, \lambda_n\rangle)$ as a conditional probability. Rather, λ is a direct specification of posterior probabilities. Ordinary Bayesian conditioning can be seen as a special case of Jeffrey conditioning, since $P(\bullet|X) = P(\bullet|\langle x_n, \lambda_n\rangle)$ when $\lambda_n = P(x_n|X)$ for all n, but not every instance of Jeffrey conditioning involves conditioning on a proposition.

Like ordinary conditioning, Jeffrey conditioning has a number of features that recommend it as a model for learning. In particular, it satisfies:

> *Sufficiency.* $P_\lambda(\bullet|x_n) = P(\bullet|x_n)$ for all n. Conversely, if $Q(\bullet|x_n) = P(\bullet|x_n)$ for all n, then $Q(\bullet) = P(\bullet|\langle x_n, Q(x_n)\rangle)$.
>
> *Refinement.* If $P_{\lambda,\mu}$ is the result of one Jeffrey shift $P - \langle x_n, \lambda_n\rangle \to P_\lambda$ followed by a second Jeffrey shift $P_\lambda - \langle y_k, \mu_k\rangle \to P_{\lambda,\mu}$, then $P_{\lambda,\mu}$ can always be represented as a single Jeffrey shift based on the refined partition $\langle x_n \wedge y_k\rangle$. Specifically,
>
> $$P_{\lambda,\mu}(\bullet) = \sum_{n,k} \beta_P(x_n, \lambda){\cdot}\beta_{P\lambda}(y_k, \mu){\cdot}P(\bullet \wedge x_n \wedge y_k),$$
>
> where $\beta_P(x_n, (\lambda) = \lambda_n/P(x_n)$ is the Bayes factor for x_n in the first shift, and $\beta_{P\lambda}(y_k, \mu) = \mu_k/P_\lambda(y_k)$ is the Bayes factor for y_k in the second shift.

There is an asymmetry lurking in this Refinement. Since β and β^* are defined relative to a given prior and posterior, reversing the order of updating can alter the final result.

> Example. You discover Colonel Mustard lying on the floor of the Conservatory. "Either Mrs. Peacock (x_1), Mrs. White (x_2) or Miss Scarlet

(x_3) did it," he mutters, before dying. You set about trying to determine the culprit, starting from equal priors $\langle P(x_1), P(x_2), P(x_3)\rangle = \langle 1/3, 1/3, 1/3\rangle$. The air is thick with perfume, and (though you can't say exactly why) it strikes you as smelling unlike Mrs. Peacock. On the basis of this olfactory experience you become less inclined to suspect her, and your subjective probability for x_1 falls to 1/5. Since this is the only effect of the experience (the scent fails to discriminate Mrs. White from Miss Scarlet), you Jeffrey condition to arrive at $\langle P_\lambda(x_1), P_\lambda(x_2), P_\lambda(x_3)\rangle = \langle 0.2, 0.4, 0.4\rangle$. The coroner arrives and examines the body. "Aha!," he says, "the Colonel was killed with a r$__$," but just then, as he is about to name the weapon, a lead pipe crashes through the window and kills him dead. Even though you could not make out what is said, you have the sense that the coroner was closer to saying "rō", which would indicate that a rope was used, than he was to saying "rŏ", which would indicate a wrench, or "rē" which would suggest a revolver. As it happens, you know that Mrs. Peacock kills with a revolver, Mrs. White prefers a wrench, and Miss Scarlet uses a rope. On the basis of your auditory experience, your probability for Miss Scarlett being the culprit rises to 0.7, and you Jeffrey condition to get $\langle P_{\lambda,\mu}(x_1), P_{\lambda,\mu}(x_2), P_{\lambda,\mu}(x_3)\rangle = \langle 0.1, 0.2, 0.7\rangle$. Surprisingly, if we imagine the evidence coming in the reverse order we get a different answer. The auditory experience, if it occurs first, induces a shift from P to $\langle P_\mu(x_1), P_\mu(x_2), P_\mu(x_3)\rangle = \langle 0.15, 0.15, 0.7\rangle$. If this is followed by the olfactory experience, a second Jeffrey shift yields $\langle P_{\mu,\lambda}(x_1), P_{\mu,\lambda}(x_2), P_{\mu,\lambda}(x_3)\rangle = \langle 0.2, 0.1412, 0.6588\rangle$.

Jeffrey conditioning clearly does not commutate as ordinary Bayesian conditioning does. One experience induced shift $P \to P_\lambda$ followed by another $P_\lambda \to P_{\lambda,\mu}$ need *not* be evidentially equivalent to $P \to P_\mu$ followed by $P_\mu \to P_{\mu,\lambda}$. Indeed, Refinement entails that Jeffrey shifts commute only when $\beta_P(x_n, \lambda)\cdot\beta_{P\lambda}(y_k, \mu) = \beta_P(y_k, \mu)\cdot\beta_{P\mu}(x_n, \lambda)$.

Many commentators object to this non-commutative aspect of Jeffrey's approach. For example, Kelly [2008, p. 616] rejects Jeffrey conditioning on the basis of:

> *Commutativity of Evidence Principle* (CEP). "To the extent that what it is reasonable for one to believe depends on one's total evidence, historical facts about the order in which that evidence is acquired make no difference to what it is reasonable for one to believe."

Others have made similar claims, including Doring [1999] and Lange [2000].

It turns out, however, that CEP is mistaken, and that Jeffrey conditioning fails to commute in precisely those circumstances where CEP fails. Refer back to the distinction between "hard" and "soft" learning experiences. It is tempting to think that all Jeffrey shifts reflect hard experiences, so that the posterior for each x_n is set to λ_n *no matter what prior is in play*, i.e., $P_\lambda(x_n) = Q_\lambda(x_n)$ for all priors

P and $\mathcal{Q}$. Such experiences obliterate any information about $\mathcal{X}$ that might have been contained in the prior: except in the very special case where $P(x_n) = P_\lambda(x_n)$ for all x_n, it will be impossible to infer anything about the prior distribution on $\mathcal{X}$ from the posterior distribution on $\mathcal{X}$. Consequently, for hard Jeffrey shifts λ and μ, one has $P_{\lambda,\mu}(y_k) = P_\mu(y_k) = \mu_k$ and $P_{\mu,\lambda}(x_n) = P_\lambda(x_n) = \lambda_n$. When μ occurs last it cancels out any information λ might have conveyed about $\mathcal{Y}$. When λ occurs last it cancels any information μ might have conveyed about $\mathcal{X}$. We should *not* generally expect updating rules to commute in such situations: CEP should fail when the second experience expunges information provided by the first. In general, we should only want commutation among hard Jeffrey shifts when the information in the first experience is preserved in the second, so that $\lambda_n = P_{\lambda,\mu}(x_n)$ and $\mu_k = P_{\mu,\lambda}(y_k)$.

As Diaconis and Zabell [1982] show, this happens exactly when λ and μ are *Jeffrey independent* relative to P in the sense that $P(x_n) = P_\mu(x_n)$ for $x_n \in \mathcal{X}$ and $P(y_k) = P_\lambda(y_k)$ for $y_k \in \mathcal{Y}$. Indeed, they prove that Jeffrey independence is necessary and sufficient for commutation when λ and μ are *hard* Jeffrey shifts.[20] So, in the case of hard learning, Jeffrey conditioning commutes in exactly those cases in which it should, viz., when the second experience does not nullify the information provided by the first.

Now, one might wonder how often "hard" Jeffrey shift experiences occur, especially in sequences that generate commutativity failures. How plausible is it to think, e.g., that hearing the coroner intone "r‥" effects one's beliefs the same way whether one hears it before or after smelling the perfume? It seems likely that differences in one's prior state of opinion will affect one's responses to evidence. There are two ways to cash this out. Lange [2000] has argued that seeming commutativity failures of Jeffrey conditioning arise only in cases where the subject is not really undergoing the *same* experiences in reverse order. Rather, the character of the second experience is altered as a result of the first, so that, e.g., smelling the perfume is a different experience after one has heard the coroner than when experienced *de novo*. This preserves the "hard" aspect of experience —- if λ and μ really are the same experiences, independent of the order in which they occur, then $P_{\lambda,\mu}(y_k) = \mu_k$ and $P_{\mu,\lambda}(x_n) = \lambda_n$ —- but it suggests that apparent failures of Commutativity have the form $P_{\lambda,\mu*}(y_k) \neq P_{\mu,\lambda*}(x_n)$ where λ*, the experience one has *after* μ, is not the same as λ, and μ*, the experience one has after λ, is not the same as μ.

Lange's approach has the disadvantage of requiring the qualitative character or content of experiences themselves to vary depending on the subject's prior epistemic state, which seems like an instance of the Hanson/Kuhn "theory ladenness of observation" fallacy. A more plausible, and more comfortably Bayesian, picture of the situation will portray subjects as drawing different conclusions from the same experiential data depending on their prior beliefs.

[20] While Diaconis and Zabell do not make the hard/soft distinction, there results are clearly meant to apply only to hard learning experiences.

Such a picture is found in Field [1978]. Field recognized that there can be "soft" Jeffrey shifts in which $P_\lambda(x_n)$ and $Q_\lambda(x_n)$ diverge. Such shifts have the form $\lambda = \langle x_n, \lambda_n(P)\rangle$, where the $\lambda_n(P)$ are weighting coefficients that *may* depend on P. Experience sets a posterior for each x_n indirectly by specifying a Bayes factor $\beta_P(x_n, \lambda)$ that is unique up to multiplication by a positive constant. Then, *independent of P*, experience stipulates that $P_\lambda(x_n)/P(x_n) \propto \beta_n$ for some set for non-negative real numbers $\beta_1, \beta_2, \ldots$.[21] For any such "Field shift" $\langle x_n, \beta_n\rangle$ there is always an associated Jeffrey shift λ where $\lambda_n(P) = \beta_n \cdot P(x_n)/\sum_j \beta_j \cdot P(x_j)$, with the associated posterior $P_\lambda(\bullet) = \sum_n \beta_n \cdot P(\bullet \wedge x_n)/[\sum_j \beta_j \cdot P(x_j)]$.

In addition to being soft, experiences that produce Field shifts preserve prior information about the distribution over $\mathcal{X}$. Given knowledge of the posterior over $\mathcal{X}$ and the fact that it arose from $P\langle x_n, \beta_n\rangle$ via a Field shift, the prior over $\mathcal{X}$ can be deduced. This ensures that the effects of successive Field shifts $\lambda = \langle x_n, \beta_n\rangle$ and $\mu = \langle y_k, \chi_k\rangle$ are always independent of the order in which they occur. In general, one has

$$P_{\lambda,\mu}(\bullet) = \sum_{k,n} \chi_k \cdot \beta_n \cdot P(\bullet \wedge x_n \wedge y_k) / \left[\sum_{i,j} \chi_i \cdot \beta_j \cdot P(x_j \wedge y_i)\right]$$

Since this is a Field shift on $\langle x_n \wedge y_k\rangle$ that is symmetric in n and k, it follows that $P_{\lambda,\mu}(\cdot) = P_{\mu,\lambda}(\cdot)$. Moreover, as Wagner [2002] showed the converse holds: soft Jeffrey shifts commute only if they can be represented as successive Field shifts.

The moral is that Field conditioning differs from forms of Jeffrey conditioning that do not commute because *subsequent Field conditioning on events in one partition does not expunge information already received about the other partition.* So, to emphasize the key points:

- Jeffrey conditioning commutes when it should, i.e., when subsequent updatings preserve information acquired in earlier updatings.
- Jeffrey conditioining does not commute when it should not, i.e., when subsequent updatings destroy information acquired in earlier updatings.

Turning now to a different facet of Jeffrey conditioning, let's ask whether there are good reasons for thinking that Jeffrey's rule is the correct way to generalize Bayesian conditioning for non-dogmatic learning experiences. An affirmative answer is provided by Diaconis and Zabell [1982], who show that, relative to a range

[21] Field portrayed $\langle x_n, b_n\rangle$ as an "input parameter" that captures the information conveyed by experience itself, so that "same experience" = "same sequence of b_n". As shown in Garber [1980], this interpretation founders on cases of repeated sampling. In our "Clue" example, suppose sniffing the perfume causes a small shift from $\langle 1/3,\ 1/3,\ 1/3\rangle$ to $\langle 0.34,\ 0.35,\ 0.31\rangle$, so that $b_1 = 34/300, b_2 = 35/300$ and $b_3 = 31/300$. If $\langle b_1, b_2, b_3\rangle$ were the experiential input itself, independent of the evidential context in which it occurred, then a second sniff would cause a shift to $\langle 0.3459, 0.3665, 0.2875\rangle$, a third would cause a shift to $\langle 0.351, 0.383, 0.266\rangle, \ldots$, and a fiftieth would cause a shift to $\langle 0.194, 0.804, 0.002\rangle$. But, repeatedly undergoing the same uninformative experience should not have such a dramatic effect on probabilities.

of ways of measuring divergences among probability functions, the Jeffrey shift P_λ is the "closest" probability Q to P that satisfies the constraints $Q(x_n) = \lambda_n$ for all n. These measures include all the following (with ω ranging over atomic elements of the algebra Ω over which the probabilities are defined):

Variational Distance. $V_P(Q) = \sup\{|P(A) - Q(A)| : A \in \Omega\}$
Brier Score. $B_P(Q) = \Sigma_\omega(P(\omega) - Q(\omega))^2$
Hellinger Distance. $H_P(Q) = \Sigma_\omega P(\omega) + Q(\omega) - (P(\omega)\cdot Q(\omega))^{1/2}$
Kullback-Leibler Entropy. $B_P(Q) = \Sigma_\omega Q(\omega)\cdot\log(Q(\omega)/P(\omega))$

In all cases except the first, P_λ is the unique minimum. To the extent that one is impressed with the idea that belief updating rules should make the smallest change in the prior consistent with the new data, these results make Jeffrey conditioning look like an excellent rule for updating subjective probabilities in the kinds of situation to which it applies.

3.3 Other Form of Bayesian Conditioning

There has been little systematic investigation of updating rules that apply when ordinary Bayesian conditioning and Jeffrey conditioning are inapplicable. In theory, a believer should be able to update on any informational constraint of the form $Exp(f) = S$, where f is a random variable and S is a set of real numbers. In practice, however, it is hard to know how to do this since the various measures of probabilistic divergence do not speak unequivocally for evidential constraints that cannot be expressed in terms of assignments of new probabilities to propositions.

An excellent illustration of the phenomenon is provided by the "Judy Benjamin Problem" of van Fraassen [1981]. Here experience directly provides the value of a *conditional* probability, and the question is how to update beliefs in light of such information.

> Example. Judy, a paratrooper, has just been dropped in the dead of night into a square region that is aligned along the compass points, and is divided into four equal-sized regions: *NE*, *NW*, *SE* and *SW*. Initially, she believes that she might have landed anywhere in the region, and that her chances of being in the east or west do not depend on her north/south location, and conversely. So, Judy begins with these priors:

	West [0.5]	*East* [0.5]
North [0.5]	0.25	0.25
South [0.5]	0.25	0.25

> Judy knows just two things about the region: (i) there are three times as many bears in the northwest as in the southwest, and (ii) her chances of seeing a bear in any region depends only on the number of bears in that region. If we let $\#(R)$ be the number of bears in region R, (i)

says $\#(NW) = 3{\cdot}\#(SW)$, and (ii) says that, for any regions R and R^*, $P(Bear \mid R)/P(Bear \mid R*) = \#(R)/\#(R^*)$. This means that Judy's posterior must satisfy $P_\lambda(R)/P_\lambda(R*) = [(R){\cdot}P(R)]/[(R*){\cdot}P(R*)]$ when she knows the value of $\#(R)/\#(R*)$ and encounters bear.

Judy sees a bear. Being a good Bayesian, she decides to update before fleeing. Since she assigns NW and SW equal priors her posterior should satisfy $P_\lambda(NW)\ /P_\lambda(SW) = 3$. In effect, she has acquired the conditional information $P_\lambda(N|W) = {}^3\!/_4$, which constrains her posterior so that for some p, $q \in [0, 1]$

	West $[p]$	*East* $[1-p]$
North $[q + (0.75 - q){\cdot}p]$	$0.75{\cdot}p$	$q{\cdot}(1-p)$
South $[(1-q) + (q - 0.75){\cdot}p]$	$0.25{\cdot}p$	$(1-q){\cdot}(1-p)$

The challenge is to determine what p and q should be.

People tend to have two strong intuitions here. First, given that Judy is as likely to be in any one quadrant as any other, it seems that seeing a single bear cannot indicate anything about her east/west location. So, p should remain fixed at ${}^1\!/_2$. Likewise, since Judy has no information about the distribution of bears in the eastern regions, it seem that seeing a single bear should not alter her views about the relative probabilities of *NE* and *SE*. So, q should be ${}^1\!/_2$. Together these intuitions completely determine the posterior.

	West [0.5]	*East* [0.5]
North [0.625]	0.375	0.25
South [0.375]	0.125	0.25

Surprisingly, this answer cannot be justified on the basis of standard "minimum change" principles. The Brier, Hellinger and Kullback-Liebler measures of probabilistic divergence produce the following results:

	$P_\lambda(NW)$	$P_\lambda(SW)$	$P_\lambda(NE)$	$P_\lambda(SE)$
Brier	0.3333	0.1111	0.2778	0.2778
Hellinger	0.363	0.121	0.258	0.258
K-L Entropy	0.3525	0.1175	0.265	0.265

While these are consonant with $q = {}^1\!/_2$, none supports $p = {}^1\!/_2$. Indeed, in every case Judy's confidence that she is in the east *increases* when she sees the bear.

Some, e.g., Grove and Halpern [1997], have argued that the failure to secure $p = {}^1\!/_2$ shows that it is misguided to use divergence measures as a guide to updating. However, for all its *a priori* appeal, the $p = {}^1\!/_2$ answer is incorrect. The intuitions in its favor rest on the claim that seeing a bear cannot convey any information to Judy about her east/west location. This is plainly wrong because it ignores the fact that learning a conditional probability almost always conveys information about underlying unconditional probabilities (the only exception being when the new value for the conditional probability is the same as the old). For example, as

van Fraassen noted, a learning experience that forces $P_\lambda(N|W)$ to be near one also forces $P_\lambda(SW)$ to be near zero. In the extreme case where Judy learns $P_\lambda(N|W) = 1$ (because she knows $\#(NW) > \#(NE) = 0$), the evidential impact of seeing a bear straightforward: Judy learns she is not in the southwest, and her posterior is $P_\lambda(\cdot) = P(\cdot|\neg SW)$. This effect is not confined to extreme probabilities. For example, learning $P_\lambda(N|W) = {}^4/_5$ provides Judy with information that forces $P_\lambda(SW)$ below ${}^1/_5$. It might seem that no such information can be conveyed when the prior for SW already falls below the upper bound that experience dictates, but this is not so. Even when $P(SW) = {}^1/_4$ Judy acquires data about unconditional probabilities. She learns that there is an increment $\delta \in [0, {}^1/_4]$ such that SW's probability must decrease by δ and NW's probability must increase by ${}^1/_2 - 3\delta$. While she does not learn δ's value, its existence tells Judy things about the unconditional probabilities of NW and SW, e.g., that at least one of them has to change and that $P_\lambda(NW)$ must end up three times larger than $P_\lambda(SW)$.

An updating rule that alters the unconditional probabilities of NW or SW must treat these changes in the way Bayesians usually treat changes in unconditional probabilities. In particular, the rule should respect the

> *Bayesian Updating Insight (BUI).* If a proposition has its probability diminished as the result of experience then, subject to obeying whatever other constraints experience imposes, the lost probability should be distributed among the other propositions in proportion to their probabilities.

For an example of an updating rule that respects BUI, focus on the increment δ. If the decrease in $P(SW)$'s probability was due to Jeffrey conditioning, then the posterior could be specified by the following recipe: (a) assign each of the other regions a number, $m_R = P(R)/P(SW)$, equal to the factor by which its prior exceeds or falls short of $P(SW)$, (b) normalize to attain weighting coefficients $w_R = m_R/(\sum_{R \neq SW} m_R)$, and (c) apportion δ so that each region other than SW increases according to its weight, $P_\lambda(R) = P(R) + \delta \cdot s_R$ for $R \neq SW$. This way of thinking about Jeffrey conditioning makes it feasible to process the new information Judy receives. When $\#(NW) = 3 \cdot \#(SW))$ is known, seeing the bear will affect Judy's updating process in two related ways. In addition to requiring her posterior to satisfy $P_\lambda(N|W) = {}^3/_4$, it also corrects her view of the relationship between the *unconditional* probabilities of NW and SW. Whereas Judy had thought $m_{NW} = 1$, she now knows $m_{NW} = 3$. In light of this information, Judy should choose a posterior that can be obtained from the prior using Jeffrey conditioning with values $m_{NW} = 3$ and $m_{NE} = m_{SE} = 1$, and that satisfies $P_\lambda(N|W) = {}^3/_4$. This process both respects BUI, subject to Judy's new views about the unconditional probabilities, and yields a determinate posterior. In fact, it yields $P_\lambda(NW) = {}^1/_3, P_\lambda(NW) = {}^1/_9, P_\lambda(NE) = P_\lambda(SE) = {}^5/_{18}$, the Brier score's answer.

For another process that satisfies BUI, imagine that Judy chooses a posterior for which $P_\lambda(N|W) = {}^4/_3$ via Jeffrey conditioning using her *initial* values $m_{NW} = m_{NE} = m_{SE} = 1$. This produces $P_\lambda(SW) = {}^1/_{10}, P_\lambda(NW) = P_\lambda(NE) =$

$P_\lambda(SE) = {}^3/_{10}$, so that all alternatives to SW receive an equal bump. Unlike the previous approach, which treats the data $P_\lambda(NW) = 3P_\lambda(SW)$ as an input to the conditioning rule, this approach treats it as a piece of information about SW alone, holding fixed the idea that NW, NE and SE are equiprobable. In effect, Judy is saying, "even though I now know that I am three times more likely to be in NW than in SW, and even though I have received no new evidence about the relative probabilities of SW, NE or SE, I am going to treat them all as equally probable for purposes of updating. So, I will satisfy the constraint by lowering SW's probability just the right amount to make it true that $P_\lambda(\neg SW) = 9P_\lambda(SW)$." The flaw in this reasoning is that it treats the new data Judy receives as being entirely about the distribution of probabilities over $\{SW, \neg SW\}$, but it concerns $\{NW, \neg NW\}$ as well.

There are other updating rules that satisfy BUI, but none yields $p = {}^1/_2$. BUI requires every proposition in $\neg SW$ with a positive prior probability to receive a portion of δ. The $p = {}^1/_2$ answer does not do this for either eastern region. By requiring $P(NE) = P_\lambda(NE)$ and $P(SE) = P_\lambda(SE)$, it portrays Judy as having more evidence than she actually has. If we imagine that, in addition to $\#(NW) = 3\#(SW)$, Judy knows $\#(E) = \#(W)$, then $p = {}^1/_2$ is correct because the posterior must obey $P_\lambda(E) = P_\lambda(W)$. However, in the absence of this *additional* data Judy has no basis for treating a bear encounter as irrelevant to her east/west location, which is what $p = {}^1/_2$ says. Without knowing $\#(E) = \#(W)$, all Judy has to go on is the information she receives and her prior. But, according to her prior both NE and SE have some claim to the probability that SW sheds. So, it is simply not true, in the absence of additional knowledge, that Judy should treat a learning experience which teaches her $P_\lambda(N|W) = {}^3/_4$ as irrelevant to the question of whether she is in the east or the west.

The upshot of this is that the $p = {}^1/_2$ intuitions provide no grounds rejecting solutions that the various divergence measures recommend. This does not answer the Judy Benjamin problem, of course, for we are still faced with different measures that yield different proposed solutions. While we will not adjudicate these issues here, it is perhaps worth remarking that, as noted above, the Brier score's proposal has a plausible interpretation in terms of Jeffrey conditioning. It, together with the Hellinger distance, also has the advantage of being symmetric, so that P_λ is the update for P given a set of constraints if and only if P is the update for P_λ given the inverse constraints. K-L entropy is asymmetric (and so not a true measure of "distance"), but its proponents will suggest that this is appropriate. K-L entropy is the *expected* change in information between the prior and posterior, where this expectation is calculated relative to the *posterior*. If this is the right quantity to minimize, we should not want or expect symmetry since the prior will yield different expectations from the posterior. It is interesting to note that if we minimize the expected K-L entropy as viewed from the perspective of the *prior* we get the $p = q = {}^1/_2$ solution. Presumably, proponents of K-L entropy will reject this on the grounds that the prior probabilities, being less well informed, are an unsuitable basis for computing expectations.

There is still much work to be done on generalizing Bayesian conditioning to learning situations in which the data does not fix probability assignments for propositions. The method of minimizing divergence seems fruitful, but the plethora of measures, and lack of agreement among them, means that it further efforts are required. A promising approach, taken above for the Brier score, is to use BUI to relate the posterior probabilities recommended by various divergence measures to more familiar updating rules.

4 SUBJECTIVE PROBABILITY AND OBJECTIVE CHANCE

Even if probabilities can be understood as rational degrees of confidence, and belief change can be modeled in full generality using Bayesian methods of updating, it remains to ask whether other interpretations of probability might also be legitimate, and how they might be encompassed within a Bayesian framework. As noted, there is a Bayesian tradition that allows for objective *a priori* probabilities, but the concept of probability can also be given an empirical gloss. Simplifying greatly and blurring distinctions, two main empirical interpretations of probability have been proposed: probabilities as *relative frequencies* and probabilities as *single-case propensities*.

Frequentist views have been dominant for about 150 years, beginning with Venn [1866], and buttressed by seminal contributions from Von Mises [1939], Reichenbach [1948], Neyman [1950] and many others. Roughly, frequentism identifies the probability of an event with the frequency of occurrences of its general type within a *reference class* of similar events generated by a long run of repetitions of a well-structured random experiment. While there is disagreement about what the "long run" involves, the most plausible views identify probabilities with *limits* of relative frequencies in the *infinitely* long run. To see the idea, imagine an idealized experiment involving a random variable X with values in some finite set $V = \{v_1, \ldots, v_K\}$ that can be sampled indefinitely to yield a sequence of observations $\langle x_1, x_2, \ldots, x_N\rangle$ for arbitrary large N. In such a scenario frequentism says that the probability of X having the value v, is, by definition, $P(X = v) = \lim_{n\to\infty} n_v/N$ where n_v is the number of trials among the first N on which X had value v. The probability exists if and only if this limit does.[22]

The propensity view interprets probabilities as measurements of the dispositions that experimental set-ups have to produce outcomes that exhibit stable frequencies. These are explicitly *single-case* probabilities. The propensity of a coin to

[22]Some frequentists, e.g., Von Mises [1939] and Martin-Löf [1966], augment this account with a "randomness" requirement to ensure that the limiting operation does not conceal patterns that should be taken into account. If a coin tossing experiment turns out $\langle h, t, h, t, h, t, h, \ldots\rangle$, it seems wrong to set the probability of heads at $1/2$. It is clearly in the spirit of the frequency view that, as N increases, there should *eventually* come a point at which n_v/N is almost as good as an estimate of the outcome of the next coin toss as would be provided by any other plausible method. Frequentists need not specify when this will happen, only that it eventually will. Here a "plausible method" is a recursive function that maps initial sequences of heads and tails into $\{h, t\}$.

come up heads on a given toss is physical a property of the coin, the tossing apparatus, and the particular situation. It is independent of what happens in other coin tossings. It *explains* why we observe the patterns of frequencies we do, but it is not *defined* in terms of those patterns. There are a variety of theories about how single-case chances are determined, but we shall not pursue the matter.

4.1 De Finetti's Rejection of Chance

Bayesians have complicated relationships with objective probabilities. Some maintain that there is no such thing. In his (1974, p. x), Bruno de Finetti, perhaps the most strident anti-objectivist, writes

> PROBABILITY DOES NOT EXIST. The abandonment of superstitious beliefs about the existence of the Phlogiston, the Cosmic Ether,...or Fairies and Witches was an essential step along the road to scientific thinking. Probability, too, if regarded as something endowed with some kind of objective existence, is no less a misleading misconception.

De Finetti believes that we are led, mistakenly, to believe in objective chances by a kind of psychological projection fallacy in which we misidentify symmetries found in our subjective beliefs with objective features of the world.

It takes a bit of apparatus to explain de Finetti's idea. Let $X_1, X_2, X_{3,\ldots}$ be an infinite sequence of random variables each with domain $V = \{v_1, v_2, \ldots, v_K\}$, and let $\boldsymbol{X}^n = V \cup V^2 \cup \ldots$ denote the set of all finite sequences drawn from V. Each $\boldsymbol{x} = \langle x_1, x_2, \ldots, x_N\rangle \in \boldsymbol{X}^n$ can be thought of as the result of an experiment that samples each of the first N variables. A *reordering* of $\boldsymbol{x}$ is an element $\boldsymbol{x}_\sigma = \langle x_{\sigma(1)}, x_{\sigma(2)}, \ldots, x_{\sigma(N)}\rangle$ of $\boldsymbol{X}^n$ that can be obtained by rearranging $\boldsymbol{x}$'s values using a permutation σ of $\{1,\ 2,\ldots,\ N\}$. Note that not every permutation yields a distinct reordering. In the binary case where $V = \{0, 1\}$, if $\boldsymbol{x} = \langle 1, 1, 0, 0\rangle$, then the permutation that swaps the even indicies and the odd indicies produces the same reordering as the one that swaps the first and last indicies and the middle indicies — both produce $\langle 0, 0, 1, 1\rangle$. In general, if σ and τ are permutations and $x_{\sigma(n)} = x_{\tau(n)}$ for all $n \leq N$, then $\boldsymbol{x}_\sigma$ and $\boldsymbol{x}_\tau$ are the *same* reordering of $\boldsymbol{x}$. By this criterion, the number of distinct reorderings of $\boldsymbol{x}$ is $N!/(n_1!{\cdot}n_2!{\cdot}\ldots{\cdot}n_K!)$ where n_1 is the number of times v_1 appears in $\boldsymbol{x}$, n_2 is the number of times v_2 appears in $\boldsymbol{x}$, and so on.

In de Finetti's terminology, a probability P over $\boldsymbol{X}^n$ is *exchangeable* when $P(\boldsymbol{x}) = P(\boldsymbol{x}_\sigma)$ for $\boldsymbol{x}_\sigma$ any reordering of $\boldsymbol{x}$.

> Example (IID Coin Tossing). If we assume that tosses of a coin are independent and that there is a fixed bias θ toward heads on any toss whatever the outcome of previous tosses, then, irrespective of the order in which heads and tails appear, the exchangeable probability of obtaining n heads and $N - n$ tails is $\theta^n{\cdot}(1-\theta)^{N-n}$.

Example (Sampling with Replacement). We sample randomly from an urn containing b black balls, w white balls and r red balls, and we always replace the drawn ball before the next trial. Then, independent of the order in which various colors appear, the exchangeable probability of any sequence of j black, k white and $N-(j+k)$ red is $(b^j \cdot w^k \cdot r^{N-(k+j)})/(b+w+r)^N$.

In these examples the trials $X_1, X_2, \ldots$ form an ordered sequence of *independent, identically distributed* (IID) random variables relative to P. This automatically makes P exchangeable because the probability of any given sequence of results from IID random variables is the product of the probabilities of the results taken individually. The converse does not hold: P can be exchangeable even when not IDD.

Example (Polya's Urn). Imagine an urn of unlimited capacity that initially holds b_0 black balls (color 0) and w_0 white balls (color 1). The proportion of balls in the urn is altered via an iterative process that, at each stage n, involves randomly drawing a ball and observing its color, and returning the ball to the urn with another of the same color, so that

- If the n^{th} ball is black, then $b_{n+1} = b_n + 1, w_{n+1} = w_n$.
- If the n^{th} ball is white, then $b_{n+1} = b_n, w_{n+1} = w_n + 1$.

If we begin with five blacks and three whites, and draw $\langle 0,1,1,0,1\rangle$, the contents of the urn will be

$(b_0 = 5, w_0 = 3)$, initial

$(b_1 = 6, w_1 = 3)$, after *black*

$(b_2 = 6, w_2 = 4)$, after *black*, *white*

$(b_3 = 6, w_3 = 5)$, after *black*, *white*, *white*

$(b_4 = 7, w_4 = 5)$, after *black*, *white*, *white*, *black*

$(b_5 = 7, w_5 = 6)$, after *black*, *white*, *white*, *black*, *white*

The probability of making this precise ordered sequence of draws is

$$\begin{aligned} P(0,1,1,0,1) &= [b_0/(b_0+w_0)]\cdot[w_0/(b_0+w_0+1)]\cdot[(w_0+0)/(b_0+w_0+2)] \\ &\quad \cdot[(b_0+1)/(b_0+w_0+3)]\cdot[(w_0+2)/(b_0+w_0+4)] \\ &= {}^5\!/_8 \cdot {}^3\!/_9 \cdot {}^4\!/_{10} \cdot {}^6\!/_{11} \cdot {}^5\!/_{12} = {}^5\!/_{264} \end{aligned}$$

Permuting the order makes the calculation different but the result remains the same e.g., $P(1,0,0,1,1) = {}^3\!/_8 \cdot {}^5\!/_9 \cdot {}^6\!/_{10} \cdot {}^4\!/_{11} \cdot {}^5\!/_{12} = {}^5\!/_{264}$. More generally, the probability for any sequence of B blacks and W whites, taken in any order, is

$$\frac{((b_0+B-1)!/(b_0-1)!)\cdot((w_0+W-1)!/(w_0-1)!)}{(b_0+w_0+B+W-1)!/(b_0+w_0-1)!}$$

So, the probability distribution generated by a Polya urn is exchangeable. It is not IID, however, since the probability of drawing, say, a black ball on the fourth trial depends on how many blacks were drawn on the first three trials.

For an example of a non-exchangeable process consider the following variant of the Polya urn.

Example (Polya Urn with Asymmetrical Replacement). Imagine a process like the Polya urn ($b_1 = 5, w_1 = 3$) except with a bias toward white, so that

- If the n^{th} ball is black, then $b_{n+1} = b_n + 1, w_{n+1} = w_n$.
- If the n^{th} ball is white, then $b_{n+1} = b_n, w_{n+1} = w_n + 5$.

Exchangeability fails since, e.g.,

$P(0,0,1,1) = {}^{5}/_{8} \cdot {}^{6}/_{9} \cdot {}^{3}/_{10} \cdot {}^{8}/_{15} = {}^{1}/_{15}$

$P(1,1,0,0) = {}^{3}/_{8} \cdot {}^{8}/_{13} \cdot {}^{5}/_{18} \cdot {}^{6}/_{19} = {}^{5}/_{247}$

Markov processes provide another instructive example of non-exchangeability.

Example (Markov Urns). Begin drawing, with replacement, from a black urn that contains 6 black balls and 4 white balls. Continue drawing from the black urn until a white ball is selected. Then begin drawing, with replacement, from a white urn that contains 2 black balls and 8 white balls. Continue drawing form the white urn until a black ball is selected. Then repeat the whole procedure. This is a Markov process with transition probabilities:

$P(\text{black on draw } n+1 \mid \text{black on draw } n) = {}^{3}/_{5}$

$P(\text{black on draw } n+1 \mid \text{white on draw } n) = {}^{1}/_{5}$

Exchangeability fails since, e.g., $P(0,0,1) = {}^{3}/_{5} \cdot {}^{3}/_{5} \cdot {}^{2}/_{5} = {}^{18}/_{125}$ and $P(0,1,0) = {}^{3}/_{5} \cdot {}^{2}/_{5} \cdot {}^{1}/_{5} = {}^{6}/_{125}$.

Though Markov processes are not generally exchangeable, they can possess a property much like exchangeability. To see the idea consider that

$$P(0,0,1,0,1) = {}^{3}/_{5} \cdot {}^{3}/_{5} \cdot {}^{2}/_{5} \cdot {}^{1}/_{5} \cdot {}^{2}/_{5} = {}^{36}/_{3125}$$
$$P(0,1,0,0,1) = {}^{3}/_{5} \cdot {}^{2}/_{5} \cdot {}^{1}/_{5} \cdot {}^{3}/_{5} \cdot {}^{2}/_{5} = {}^{36}/_{3125}$$

It is no accident that these probabilities agree: (a) both sequences start out with black, which contributes a factor of ${}^{3}/_{5}$ to the probability; (b) each contains two switches from black to white, which each contribute a factor of ${}^{2}/_{5}$; and (c) each contains one switch from white to black, which contributes a factor of ${}^{1}/_{5}$. Indeed, if we take a sequence of length N that starts with a black and if we count the number i of black-to-white transitions and the number j ($= i$ or $i-1$) of white-to-black transitions, then the sequence's probability is ${}^{3}/_{5} \cdot ({}^{2}/_{5})^{i} \cdot ({}^{1}/_{5})^{i-1} \cdot ({}^{4}/_{5})^{N-2 \cdot i}$

when $j = i - 1$ and is ${}^{3}/_{5} \cdot ({}^{2}/_{5})^{i} \cdot ({}^{1}/_{5})^{i} \cdot ({}^{3}/_{5})^{N-2\cdot i+1}$ when $j = i$. So, all that matters to the probability of a sequence is its first element, and the number of transitions of each type.

This suggests a fruitful chain of definitions, following Diaconis and Freedman [1980]. Given a sequence $\boldsymbol{x} \in \boldsymbol{X}^n$ and distinct $v_i, v_j \in V$, let $\tau(v_i, v_j)$ be the number of v_i-to-v_j transitions in $\boldsymbol{x}$, i.e., it is number of times that $x_n = v_i$ and $x_{n+1} = v_j$. Say that two sequences $\boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{X}^n$ are *similar* when they have the same first element, so $x_1 = y_1$, and have the same transition numbers, so that $\tau^x(v_i, v_j) = \tau^y(v_i, v_j)$. P is *partially exchangeable* when each pair of similar sequences has the same probability. Every exchangeable distribution is partially exchangeable since similar sequences can always be obtained from one another via reordering. But, not every partially exchangeable distribution is exchangeable since not every reordering preserves initial elements and transition numbers.

It is illuminating to interpret these concepts in terms of sufficient statistics. A necessary and sufficient condition for P to be exchangeable is that the frequencies of occurrence of the sample values constitute a sufficient statistic. Given $\boldsymbol{x} = \langle x_1, \ldots, x_N \rangle$ and $v_k \in V$, let $F_x(v_k) = n_k/N$ be the frequency with which v_k appears among $\boldsymbol{x}$'s elements. To say that these frequencies are sufficient statistics means that any extension $\boldsymbol{x}^+ = \langle x_1, \ldots, x_N, \ldots, x_{N+M} \rangle$ of $\boldsymbol{x}$ satisfy $P(\boldsymbol{x}^+|\boldsymbol{x} = P(\boldsymbol{x}^+|F_x(v_1), F_x(v_2), \ldots, F_x(v_K), N) = P(\boldsymbol{x}^+|n_1, n_2, \ldots, n_K)$. So, if you know $\boldsymbol{x}$'s length and know the frequencies with which the various possible v-values appear in $\boldsymbol{x}$, then further information about the sequence (e.g., information about the order in which the v_i appear) is immaterial to any conclusions you might draw about future trials. For instance, knowing the frequencies of black and white balls that have appeared in the first N draws from a Polya urn suffices to establish a definite probability for the next draw even if you are ignorant of the order in which colors appeared. If, say, you know that the urn was set up so that $(b_0 = 5, w_0 = 3)$ and are told that 60% black balls and 40% white balls came up in the first 100 draws, then you can deduce $(b_{100} = 65, w_{100} = 43)$ and conclude that the probability of the 101^{st} draw being black and the 102^{nd} draw being white is ${}^{65}/_{108} \cdot {}^{43}/_{109}$.

The connection between exchangeable sequences and frequencies is quite fruitful. Consider the following question: Given non-negative natural numbers n_1, $n_2, \ldots, n_K$ that sum to N, how likely is it that the frequencies in the first N trials will be $F(v_1) = n_1/N, F(v_2) = n_2/N, \ldots, F(v_K) = n_K/N$? When P is exchangeable there is an easy answer. Choose an $\boldsymbol{x} \in \boldsymbol{X}^n$ of length N with desired frequency profile. $\boldsymbol{x}$ will have a definite probability, $P(\boldsymbol{x})$, which it shares with all its reorderings. Since there are $N!/(n_1! \cdot n_2! \cdot \ldots \cdot n_K!)$ reorderings of $\boldsymbol{x}$, and since these reorderings are exactly the sequences of length N with $\boldsymbol{x}$'s frequency profile, it follows that

$$P(\{\boldsymbol{y} \in X^n \text{ of length } N : F_y(v_k) = n_k/N \text{ for } k = 1, 2, .., K\})$$
$$= P(\boldsymbol{x}) \cdot N!/(n_1! \cdot n_2! \cdot \ldots \cdot n_K!).$$

One can think of this as a probability over frequency profiles for sequences of length N, so that $P^N(n_1/N, \ldots, n_K/N) = P(\boldsymbol{x}) \cdot N!/(n_1! \cdot n_2! \cdot \ldots \cdot n_K!)$ is the

probability, according to P, of observing a sequence of length N with the specified profile.

> Example. In five trials from a Polya urn with $(b_0 = 5, w_0 = 3)$, what is the probability of observing $^2/_5$ black and $^3/_5$ white? Answer: the probability of any single sequence of two blacks and three a white is $^5/_{264}$. Since there are $^{5!}/(3! \cdot 2!) = 10$ reorderings of such a sequence, $P^5(^2/_5, ^3/_5) = \ ^{50}/_{264} \approx 0.19$.

Things get even more interesting when we consider P^N's *cumulative* distribution function. For each $p \in [0, 1]$, this gives the probability of obtaining a sequence of length N with a frequency of black balls that does not exceed p, $P^N(F(\text{black}) \leq p)$. Here are some values:

$$P^N(F(\text{black}) \leq p)$$

$p =$	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$N = 5$	0.026	0.026	0.121	0.121	0.310	0.310	0.576	0.576	0.841	0.841
$N = 10$	0.003	0.018	0.052	0.117	0.218	0.354	0.516	0.686	0.838	0.949
$N = 20$	$3{\cdot}10^{-4}$	0.005	0.023	0.071	0.160	0.298	0.475	0.668	0.841	0.958
$N = 50$	$5{\cdot}10^{-6}$	0.001	0.011	0.045	0.123	0.255	0.444	0.656	0.846	0.967
$N = 100$	$2{\cdot}10^{-7}$	$5{\cdot}10^{-4}$	0.007	0.037	0.110	0.242	0.432	0.652	0.849	0.970
$N = \omega$	0	$1.77{\cdot}10^{-4}$	$4.67{\cdot}10^{-3}$	0.0288	0.0963	0.2266	0.420	0.647	0.852	0.974

A clear pattern of convergence can be discerned. Indeed, one can show that, as the length of the sequence increases, the cumulative distribution for P^N converges to the bottom row. This is the cumulative distribution of the *beta density*: $\beta_{(5,3)}(\theta) = \theta^4 \cdot (1-\theta)^2 / \int_0^1 t^4 \cdot (1-t)^2 \ dt$. To appreciate what this means, think of θ as the limiting frequency of an infinitely long sequence of draws from a $(b_0 = 5,\ w_0 = 3)$ Polya urn. $P^\omega(\theta) = \beta_{(5,3)}(\theta)$ is then the probability that, in the infinitely long run, sampling from the urn will yield a sequence that has θ as its limiting frequency for black. This result generalizes to Polya urns with arbitrary initial contents $(b_0 = b, w_0 = w)$. Here taking a limit of P^N produces the beta density $\beta_{(b,w)}(\theta) = \theta^{b-1} \cdot (1-\theta)^{w-1} / \int_0^1 t^{b-1} \cdot (1-t)^{w-1} \ dt$ as the probability for obtaining a countably infinite sequence of draws that has θ as its limiting frequency for black.

De Finetti [1939] proved that something similar holds for exchangeable probabilities on binary random variables, and Hewitt and Savage [1955] extended this result to the more general setting considered here.

> *De Finetti's Representation Theorem.* Suppose that P is exchangeable over the sequence of random variables $X_1, X_2, \ldots$, each of which has values in $\langle v_1, \ldots, v_K \rangle$, and let $\Theta = \{\langle \theta_1, \theta_2, \ldots, \theta_K \rangle \in [0, 1] : \theta_1 + \ldots + \theta_K = 1\}$ be the set of all probability distributions over $\langle v_1, v_2, \ldots, v_K \rangle$. Then,
>
> (A) There is a unique probability P^ω over Θ that can be obtained as the limit of probabilities for frequencies $P^\omega(\theta_1, \ldots, \theta_k) =$

$\lim_{N\to\infty} P^N(n_1/N, \ldots, n_K/N)$ where P^N is the probability over frequency profiles for sequences of length N.

(B) If $P(\boldsymbol{x} \mid \theta)$ is the probability of obtaining $\boldsymbol{x} \in \boldsymbol{X}^n$ as the initial segment of a infinite sequence with frequency profile $\theta \in \Theta$, then

$$P(\boldsymbol{x} \mid \theta) = \theta_1^{n1} \cdot \theta_2^{n2} \cdot \ldots \cdot \theta_K^{nK}$$

where $n_1, \ldots, n_K$ are the number of times, respectively, that $v_1, \ldots, v_K$ appear in $\boldsymbol{x}$.

(C) In light of (B), the unconditional probability of any $\boldsymbol{x} \in \boldsymbol{X}^n$ can be expressed as a mixture of the IID probabilities, so that

$$P(\boldsymbol{x}) = \int_\Theta \theta_1^{n1} \cdot \theta_2^{n2} \cdot \ldots \cdot \theta_K^{nK} \cdot P^\omega(\theta) d\theta$$

This highlights the deep connection between exchangeability and limiting frequencies. It says that having an exchangeable subjective probability over $X_1, X_2, \ldots$ involves having determinate expectations about the probabilities with which various liming frequencies are likely to occur (assuming sampling to infinity), and that, conditional on one of these expectations, the subjective probability for each finite sequence is as it would be if that sequence was generated by an IID process.

Partially exchangeability relates to Markov processes in a similar way. The sufficient statistic for partial exchangeability is the pair $\langle x_1, [\tau(v_i, v_j)]\rangle$ consisting of a sequence's first entry and the $K \times K$ matrix of its transition numbers. Just as there is a limiting cumulative distribution over long-run frequencies for exchangeable random variables, there is likewise a limiting cumulative distribution over $\langle x_1, [\tau(v_i, v_j)]\rangle$ in the partially exchangeable case. This cumulative distribution yields a probability density over the set of all Markov processes that specify an initial probability distribution $\theta = \langle \theta_1, \theta_2, \ldots, \theta_K\rangle$ for X_1 and a matrix of $\theta_T = [\theta(v_i, v_j)]$ of transition probabilities. Then, if $\boldsymbol{\Theta}$* is the set of all pairs $\langle \theta, \theta_T\rangle$, we obtain the following generalization of de Finetti's theorem:

> Theorem Diaconis and Freedman [1980]: Let P be partially exchangeable over a sequence of random variables $X_1, X_2, \ldots$ that is *recurrent* in the sense that each v_k in $\langle v_1, \ldots, v_K\rangle$ appears infinitely often among the X_n. Then,
>
> (A) There is a unique probability P^ω over $\boldsymbol{\Theta}$* that can be obtained as the limit of P^N's probabilities for initial frequencies and transition numbers.
>
> (B) If $P(\boldsymbol{x} \mid \theta, \theta_T)$ is the probability that $\boldsymbol{x} \in \boldsymbol{X}^n$ is the initial segment of a infinite sequence with Markov profile $\langle \theta, \theta_T\rangle$, then
>
> $$P(\boldsymbol{x} \mid \langle \theta, \theta_T\rangle) = \theta(x_1) \cdot \theta(x_1, x_2) \cdot \theta(x_2, x_3) \ldots \theta(x_{K-1}, x_K)$$
>
> (C) In light of (B), the unconditional probability of any $\boldsymbol{x} \in \boldsymbol{X}^n$ can be expressed as a mixture of the Markov probabilities, so that
>
> $$P(\boldsymbol{x}) = \int_{\Theta*} \theta(x_1) \cdot \theta(x_1, x_2) \ldots \theta(x_{K-1}, x_K) \cdot P^\omega(\theta, \theta_T) d\langle \theta, \theta_T\rangle$$

This highlights the deep connection between partial exchangeability and Markov processes. It says that having a partially exchangeable subjective probability over $X_1, X_2, \ldots$ involves having determinate expectations about initial elements and transition frequencies (assuming sampling continues to infinity with recurrence), and that conditional on these expectations the subjective probability for a finite sequence is as it would be if that sequence was generated by a Markov process.

Friends of objective probability might take comfort in these results. Those favorably inclined toward limit-of-frequency interpretations will construe $P^{\omega}(\theta)$ as a subject's estimate of probability that, in the infinitely long run, sampling from $X_1, X_2, \ldots$ will yield a sequence that has θ_1 as it its limiting frequency for v_1, θ_2 as its limiting frequency for v_2, and so on. Similarly, under conditions of partial exchangeability and recurrence, $P^{\omega}(\boldsymbol{x} \mid \theta, \theta_T)$ will be her subjective probability that, in the infinitely long run, sampling from $X_1, X_2, \ldots$ will yield a sequence with the given transition frequencies. Proponents of propensity views will view $P^{\omega}(\boldsymbol{x} \mid \theta)$ and $P^{\omega}(\boldsymbol{x} \mid \theta, \theta_T)$ as single-case probabilities produced by an underlying IID or Markov causal process. Either way, it might seem, De Finetti's theorem can be read as saying that, under conditions of exchangeability or partial exchangeability, a person's subjective probabilities are her subjective expectations of objective chances. This makes it appear as if objective chances are required to explain why subjective probabilities have the exchangeability properties they do.

De Finetti draws the opposite moral. Instead of seeing exchangeability judgments as requiring explanation in terms of objective chances, he regards them as the bedrock phenomena. It can make sense, he maintains, for a person to view a process as exchangeable or partially exchangeable *even when she denies that it is governed by objective chances*. The probabilities $P^{\omega}(\boldsymbol{x} \mid \theta)$ and $P^{\omega}(\boldsymbol{x} \mid \theta, \theta_T)$ will still exist of course, but instead of reflecting objective chances they display the subject's personal opinion that facts about the past ordering of outcomes, or about the positions at which transitions among outcomes occur, are irrelevant to questions about future outcomes. For example, if you treat a coin tossing process as exchangeable – i.e., if you judge that the past order of heads and tails provides no relevant information about future heads or tails – then de Finetti's theorem shows that your inferences from the data will be identical to those of a person who thinks the coin has a determinate, perhaps unknown, objective chance of landing heads. Even so, there is nothing in your inductive practices that *requires* you to invoke objective chances: your exchangeability judgments do all the work. While you might act *as if* you believe that the coin has some objective chance of landing heads, the moral of de Finetti's theorem is that this attitude can be based on judgments of exchangeability alone. Indeed, de Finetti would stress, the value of $P(\boldsymbol{x})$ on the left of the (C) equations is *not* derived from the chance estimate on the right. Rather, $P(\boldsymbol{x})$ comes directly from the exchangeable probability on $\boldsymbol{X}^n$, and $P^{\omega}(\boldsymbol{x} \mid \theta)$ and $P^{\omega}(\boldsymbol{x} \mid \theta, \theta_T)$ are constructed, *post hoc*, to make the equations work out. The invocation of objective chance is a third wheel: all inductive inferences drawn on the basis of an exchangeable subjective probability can, in principle, be explained without chances.

De Finetti draws a radical conclusion. He maintains that the idea of objective chances is a kind of illusion that arises when we project our subjective beliefs onto the world. It seems like there are objective chances, he says, only because we so often reason as we would if chances existed, but this reasoning is based solely on exchangeability judgments. When we are faced with a situation in which various aspects of the order of past outcomes seems immaterial to the probability of future events we illegitimately reify some chance property to explain our inductive tendencies in much the same way that ancient peoples invoked "fairies and witches" to explain coincidences in nature. But, the objective facts actually concern not chances but our personal tendencies to draw certain sorts of inductive conclusions from specified data. Instead of saying the weaker, true thing that the frequency of outcomes in past trials is a sufficient statistic for inferences about future trials, we say the stronger, false thing that the data is being generated by an IID process governed by objective chances. Objective chance, according to de Finetti, is a man-made chimera.

4.2 The Principal Principle

Some Bayesians are more tolerant of objective chances. David Lewis, for example, writes [1980, p. 83]:

> We subjectivists conceive of probability as the measure of reasonable partial belief. But we need not make war against other conceptions of probability, declaring that were subjective credence leaves of, there nonsense begins. Along with subjective credence we should believe in objective chance. The practice and analysis of science requires both concepts. Neither can replace the other.

So, says Lewis, subjective probability and objective chance can peacefully coexist. The challenge for Bayesianism is not to show how chance can be eliminated, but to formulate *chance/credence principles* (CCPs) that explain how information about chances should affect credences. Lewis proposes a simple CCP, the *Principal Principle*, which he hoped would characterize the conditions under which credences and chances should align, but things turned out to be far more complicated than he first thought.

To appreciate the complexities, imagine an agent who has a subjective probability P_t at each time t, and who assigns credences to propositions, $Ch_t(A) = p$, which say that the chance at t of event A occurring is p.[23] On a first pass, we might try to formulate a CCP as follows:

> CCP (Incorrect). If $P_t(Ch_t(A) = p) > 0$, then $P_t(A|\ Ch_t(A) = p) = p$. More generally, a subject's unconditional probability for A should be her expectation of its objective chance, $P_t(A) = \int_0^1 P_t(Ch_t(A) = p) \cdot p\ dp$.

[23] Here "$Ch_t(A)$" is a non-rigid designator — "the chance of A at t, whatever it is" — and p is a determinate real number in [0, 1].

The trouble with this, as Lewis recognized, is that evidence about chances can be rendered moot by facts about later chances or by direct evidence of A's truth-value.

> Example. A coin was tossed ten times yesterday. You know that it was either biased 2:1 for heads or 3:1 for tails, and you think these are equally likely. You need to estimate the probability of the proposition, A, that the sixth toss was a head. If this is all you know, then it is reasonable to set your credence in A at your expectation of yesterday's chances, in which case $P_t(A) = {}^{11}/_{24}$. But, if you then learn that nine of the ten tosses were heads, it is crazy to set $P_{t+1}(A) = {}^{11}/_{24}$. The right answer is near ${}^{9}/_{10}$, an answer you cannot attain as a mixture of the chances.

Lewis [1980, p. 92] calls a proposition B *admissible* with respect to $Ch_t(A) = p$ when it is "the sort of information whose impact on credences... comes entirely by way of chances." B is inadmissible when B conveys information about A that cannot be accounted for by changes in A's chance distribution conditional on B, in which case $P_t(A|\ Ch_t(A) = p \wedge B) \neq p$ for some p. To repair our CCP we need to explain the substantive conditions under which information is and is not admissible.

There are some clear first steps:

- Current chances screen-off the past. If B is entirely about the state of the world prior to t, then B is admissible with respect to $Ch_t(A) = p$.
- Later chances screen-off earlier chances. If B entails that A's chance is anything other than p at time t or later, then B is inadmissible with respect to $Ch_t(A) = p$. So, any proposition $Ch_s(A) = q$ with $s > t$ and $q \neq p$ is inadmissible with respect to $Ch_t(A) = p$, and $P(A|\ Ch_t(A) = p \wedge Ch_s(A) = q) = q$.
- Truth screens-off everything. If B entails A, then B is inadmissible with respect to $Ch_t(A) = p < 1$, and $P(A|\ Ch_t(A) = p \wedge B) = 1$.

In his initial paper Lewis made these basic observations and proposed the following:

> *Principal Principle.* Let be P_t be any "reasonable" initial credence function with $P_t(Ch_t(A) = p) > 0$. If B is admissible for $Ch_t(A) = p$ then $P_t(A|\ Ch_t(A) = p \wedge B) = p$ and, more generally, $P_t(A|B) = \int_0^1 P_t(Ch_t(A) = p|B) \cdot p\ dp$.

Of course, this merely gives a name to the problem since, absent any substantive theory of admissibility, the Principle says little. Lewis initially thought that the incomplete theory of inadmissibility just sketched would suffice for most purposes, but it turned out that, on Lewis's favored theory of chance, many statements about *current* chances are inadmissible relative to the current chances. The reasons for

this are unimportant, but subsequent struggles with the notion of admissibility made it clear that much remained to be done before an acceptable CCP would be forthcoming.

One sticking point is Lewis's focus on admissibility for propositions rather than for credence functions. He sometimes speaks as if the Principal Principle can be applied when P_t is any reasonable (= probabilistically coherent) credence function, but this cannot be. If B contains inadmissible information then any P_t that assigns it a high probability, even short of one, would need to be ruled out. In light of this, it can be tempting to think that P_t should be restricted to "pure" priors that either contain no information about A or can be derived from such priors via learning experiences that do not bring in inadmissible information.[24] Unfortunately, there is no such thing as a pure prior in this context. Any prior that assigns unconditional probabilities to statements about A's chances (a prerequisite for this whole discussion) is thereby taking some stand on A's truth-value, and it is not clear what it could mean to say that one of these stands is evidentially neutral. So, we need a distinction between those credence functions that encode information, either as certainties or not, that make it reasonable to align credences with chances and those that do not. With this in mind, we can reformulate Lewis's principle thus:

> *Principal Principle.* If $P_t(Ch_t(A) = p) > 0$ and if P_t does not encode any information that is inadmissible with respect to $Ch_t(A) = p$, then $P_t(A|\ Ch_t(A) = p) = p$ and $P_t(A) = \int_0^1 P_t(Ch_t(A) = p) \cdot p\ dp$.

We still need a substantive admissible/inadmissible distinction if this is to be useful, but at least we now have the right distinguishata.

One might wonder, however, whether a substantive distinction is really necessary. Perhaps we can do without it if we adjust our views about the relationship between credence and chance. Such an approach has been suggested by Lewis [1994], Hall [1994] and Thau [1994]. Strevens [1995] and Meacham [2005] make similar proposals. The basic idea is that we can eliminate the need for any substantive theory of admissibility by recognizing that a believer should align her credences with known chances *only* when these chances incorporate all the information *the believer* possesses. When the Principal Principle fails it is always because the credence function encodes information not found in the chance distribution. A vivid, if unrealistic, example is provided by "crystal ball" cases, Hall [1994].

> Example. Assume the same set-up as in the previous example except that coin will be tossed tomorrow. Before making your estimate you consult a soothsayer you believe to a reliable. She tells you that nine of the ten tosses will land heads, an unlikely event whichever way the coin is biased. Taking her word, you end up assigning P_t(90% heads)

[24] Meacham [2005], for example, speaks of a "hypothetical prior" that gives the right credences to hold before the receipt of any evidence about A.

> $= 1$, but Ch_t(90% heads) is either $10{\cdot}(^2/_3)^9 \cdot (^1/_3)$ or $10{\cdot}(^1/_4)^9 \cdot (^3/_4)$, both very small numbers. So, you possess information that the chance distribution lacks, and you would be unwise to align your credences with expected chances since doing so prevents you from believing A to degree 0.9 or to any degree greater than $10{\cdot}(^2/_3)^9 \cdot (^1/_3) \approx 0.087$.

Lewis, Hall and Thau suggest that one ought to align one's credences with the chances *conditional on the extra information one possesses*. So, the chance/credence principle should be expresses thus:

> *New Principle*. If E expresses every relevant item of data that a believer has concerning A's truth or falsity, and if $P_t(Ch_t(A \mid E) = p) > 0$, then $P_t(A \mid Ch_t(A \mid E) = p \wedge E) = p$ and, more generally, $P_t(A \mid E) = \int_0^1 P_t(Ch_t(A \mid E) = p \mid E){\cdot}p\ dp$.

We no longer need a substantive theory of admissibility since any "inadmissible" information B will be incorporated into E and so into the chance distribution. B is always admissible for $Ch_t(A \mid E) = p$ when E entails B. Alternatively, as Strevens (1995) observes one can simply define B as inadmissible when it alters the chances, $Ch_t(A \mid B) \neq Ch_t(A)$, and then New Principle and the Principle Principal end up being equivalent for admissible evidence, and only the New Principle applies when a believer has inadmissible evidence.

It is instructive to think about the Principal Principle and New Principle using the theory of "expert" probabilities developed in Gaifmann [1986]. An epistemic expert is a probabilistic information source to which a believer defers by aligning her credences with the source's probabilities, to the extent that she can discern what those probabilities are. More exactly, say that an believer with a subjective probability P treats another probability Q as an *epistemic expert* with respect to the propositions in $\mathcal{A}$ exactly when $P(A|Q(A) = a) = a$ for all $a \in [0, 1]$ and $A \in \mathcal{A}$.[25]

Truth, the probability that assigns one to all truths and zero to all falsehoods, must be accorded expert status by any coherent credence function. Other commonly alleged experts include Chance, Epistemic Probability and Physical Probability (derived, e.g., from quantum mechanics or statistical mechanics). There is a pecking order among experts. While a coherent believer will always *expect* all her experts to agree with one another, in the technical sense that $Exp(Q(A){-}Q^*(A)) = 0$, this allows for the possibility that experts sometimes disagree in fact. When this happens, the pronouncements of some experts nullify those of others. Truth sits atop the heap. As a consequence of the laws of probability, it must receive complete deference in all circumstances from all other experts. As we have seen, later chances merit more deference than earlier ones, so later chances are higher in the pecking order. Likewise, current chances trump any expert whose probabilities are based solely on information about the past. In this guise, the problem

[25] Typically, a person's deferential attitudes toward information sources depend on both the content of the proposition in question and the probability assigned.

of admissibility reemerges as the problem of determining which experts trump Chance.

In considering this, it is useful to follow Hall (2004) in distinguishing two sorts of epistemic experts. A *database-expert* deserves deference owing to its superior knowledge. P defers to Q (solely) as a data-base expert when there is a random variable X such that (a) P is uncertain about X's value, (b) P knows that Q is certain about X's value, and (c) knowing X's value would make Q's views irrelevant to P. Here, P's deference to Q does not require P to admire Q's reasoning or insight: she defers to Q simply because she believes Q knows more than she does. In other scenarios, P's *lack* of deference toward Q might be based entirely on her perception that Q is missing some key datum. Here, (a*) P is certain about X's value, (b*) P knows that Q is uncertain about X's value, but (c*) P would defer to Q if Q knew the truth about X. Hall calls such a Q an *analyst expert*. Here P respects Q's general reasoning ability rather than her information. When P's total evidence can be expressed by a proposition E_{tot}, Q is an analyst-expert *in re* A for P when $P(A|Q(A|\ E_{tot}) = a) = a$ for all a.

An epistemic expert might be a data-base expert or an analyst expert or some mixture of the two. Truth is the archetypal data-base expert. No great reasoner, Truth merely consults the premises at its disposal (i.e., every truth) and asserts the conclusions it finds. Our deference is due entirely to Truth's wonderful premise set. In contrast some authors, e.g., Williamson [2000, ch. 10], believe in a kind of "epistemic or evidential" probability that serves as a pure analyst expert. Some versions of objective Bayesianism can be interpreted as attempts to cash out this notion. Jaynes, for example, is committed to the following: if Q_{ME} is the prior that results from applying MaxEnt to the evidential constraints of an inductive problem, then the right credences to adopt in that problem are always the MaxEnt prior conditioned on subsequent evidence, so $P(A|\ Q_{ME}(A|E_{tot}) = a) = a$. This is just to see the MaxEnt prior as a flawless inductive reasoner.

It should be clear that the New Principle captures the way in which we would defer to Chance as an analyst expert, whereas the Principal Principle is appropriate if our deference is grounded in "data-base" considerations. The puzzle is where to place Chance on the analyst/data-base spectrum. Should we seek to align credences with known chances because Chance encodes information we lack, or because it attains a level of inductive reasoning we could not hope to match, or is it a little bit of both? One's answer to these questions will reflect one's views about how to approach the problem of admissibility, and how to formulate the chance/credence principle. At one end, we have Hall, who writes [2004, p. 101] that "*chance is an analyst-expert*... this claim holds for chances at *any* time, and *without qualification*" (i.e., for any proposition A for which $Ch(A|E)$ is defined). Hall's proposal, then, is this:

> *Chance as Analyst Expert*: $P_t(A|\ Ch_t(A|E) = p \wedge E) = p$ for any body of evidence E with $P_t(Ch_t(A|E) = p \wedge E) > 0$.

This portrays chance as an ideal inductive reasoner: she might not know much, but give her your data and she will produce a probability to which you should defer. This, as Hall sees things, provides a rational for the New Principle.

There is no doubt that "analyst expert" considerations play a crucial role in explaining our deference to Chance. However, as Joyce [2007] emphasizes, this cannot be the whole story since there are "data-base" considerations in play as well. The really remarkable thing about Chance is that current chances screen-off past facts. When one knows that an argon-41 atom has a half-life of 109.6 minutes, one should assign credence of one-half to it decaying in the next 109.6 seconds *whatever one knows about the past history of the world*. One can best explain this by supposing that the probabilities that actually realize Chance encode far more information than any believer could have about the past and present. Chances have this property because they are realized by physical probabilities, like those found in quantum-mechanics or statistical mechanics, which serve as "summary statistics" that measure the causal tendency of the current state of the world to produce future effects. According to Joyce, our deference to such physical probabilities rests ultimately on our views about the causal structure of the world, and about the restrictions that this structure places on our ability to acquire evidence. In particular, given the (contingent) facts that (a) the past only causally influences the future by influencing the present, (b) our ability to predict (nearly) all contingent future events is based entirely on evidence about the present causes of those events, (c) physical probabilities encode all the evidence about the present causes of future events that any human being could possibly have at present (and lots more), it follows that believing at any time-t that the time-t physical probability of some event is p involves thinking that no additional information *that could be acquired by a believer at* t can undermine the chance assignment.[26] The physical probabilities that realize Chance are thus data-base experts for us with respect to all questions about future events with current causes.

This alters our view of the Principal Principle. While there are still inadmissible propositions and credence functions, and while it is still true that believers who knew these propositions or had these credences would not defer to Chance as an expert, believers who understand their epistemic situation will recognize that they cannot be warranted in believing inadmissible propositions or having inadmissible credence functions. Predictions about the future that are not expectations of current objective chances cannot be justified on the basis of the sorts of evidence that, as a matter of physical possibility, human believers can possess. Thus, if

[26]There is a subtlety here. When Chance is realized by genuinely indeterministic physical probabilities, like those in quantum mechanics, a believer cannot acquire information about the past or present that undermines Chance simply because there is no such information. Chance already knows everything relevant to the causes of future events – it's all in the world's quantum state. In contrast, when Chance is realized by statistical probabilities that average over "hidden variables", like those of statistical mechanics, there is causally relevant information about the past and present that Chance lacks. However, when we defer to such probabilities as experts part of what we are doing is admitting to ourselves that we lack the ability, perhaps for contingent reasons, to acquire evidence about the hidden variables that is not already reflected in the chance distribution.

Chance is realized by physical probabilities that encode all present information that is relevant to future events, then we have a *practically sufficient* theory of admissibility for use with the Principal Principle. In practice, we defer to current chances more-or-less unconditionally because we know that no evidence we can actually obtain will undermine them. Go back to the "crystal ball" example. When you are certain the coin is biased 2:1 for heads, part of what you believe is that acquiring additional information about the past, or more detailed information about the present, will provide no further insight into the causal processes that lead the coin to come up heads or tails. And, insofar as you believe that the basic evidence available to human beings at a time is restricted to facts that can be learned at that or previous times, you will never defer to a soothsayer who pretends to know more than the current chances. It is not that reliable crystal balls are metaphysically impossible. The point, rather, is that it is *physically* impossible for us, or other physical creatures, to have evidence that would warrant taking any information source whose pronouncements we can know to be a more reliable guide to the future than the current chances.

In the end, both the Principal Principle and the New Principle capture central aspects of the relationship between credence and chance. The New Principle expresses our deference to chance's prowess at drawing inductive conclusions, while the Principal Principle captures the idea that chance is a data-base expert whose access to causally relevant information about the world, though not perfect, far outstrips anything to which humans can aspire.

5 CONCLUSION

Though not monolithic, Bayesianism offers a powerful and compelling set of methods for drawing inductive inferences. Its unifying ideas are (a) Pascal's recognition that uncertainty is best expressed probabilistically and that values of unknown quantities are best estimated using the principle of mathematical expectation, and (b) Bayes's insight that learning and inductive inference can be fruitfully modeled using conditional probabilities and Bayes's theorem. The two central challenges for Bayesianism are the problem of the priors, and the development of general methods for Bayesian conditioning. Bayesians have responded to the problem of the priors by proposing the use of ignorance priors that are justified *a priori*, embracing a radical subjectivism in which probabilities are mere degrees of coherent credence, or have sought refuge in the idea that subjective prejudices will wash out as evidence increases. On the conditioning front, Jeffrey has extended Bayes's basic approach to account for non-dogmatic learning experiences, and further developments based on measures of divergence among probabilities seem promising. Bayesians have a vexed relationship with objective chance. Some reject the notion outright and portray chances as projections of personal inductive tastes onto the world. Others hope to make room for chances by developing chance/credence principles that clarify and explain the evidential relationships between the two kinds of probability. At bottom, however, all Bayesians agree that inductive reasoning

involves drawing conclusions from new data on the basis of prior information using update rules that require conditioning on the evidence.

BIBLIOGRAPHY

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

[Bayes, 1763] T. Bayes. "An Essay Toward Solving a Problem in the Doctrine of Chances," *Philosophical Transactions of the Royal Society of London* **53**: 370-418, 1763.

[Brier, 1950] G. W. Brier. "Verification of Forecasts Expressed in Terms of Probability," *Monthly Weather Review*, 75: 1-3, 1950.

[Chalmers, 1999] A. F. Chalmers. *What is This Thing Called Science*, 3^{rd} ed. Indianapolis: Hackett, 1999.

[Christensen, 1996] D. Christensen. "Dutch-Book Arguments De-Pragmatized," *Journal of Philosophy* **93**: 450-479, 1996.

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

[Doring, 1999] F. Doring. "Why Bayesian Psychology is Incomplete," *Philosophy of Science* **66** (Proceedings): S379-389, 1999.

[de Finetti, 1937] B. de Finetti. La prévision: ses lois logiques, ses sources subjectives, *Ann. Inst. Henri Poincaré* 7: 1–68. Translation reprinted in H.E. Kyburg and H.E. Smokler (eds.) (1980), *Studies in Subjective Probability*, 2nd ed.: 53–118. New York: Robert Krieger, 1937.

[de Finetti, 1974] B. de Finetti. *Theory of Probability*, Vol. 1. New York: John Wiley and Sons, 1974.

[Diaconis and Freedman, 1980] P. Diaconis and D. Freedman. "De Finetti's Theorem for Markov Chains," *Annals of Probability* **8**: 115-130, 1980.

[Diaconis and Zabel, 1982] P. Diaconis and S. Zabell. "Updating Subjective Probability," *Journal of the American Statistical Association* **77**: 822–830, 1982.

[Doob, 1971] J. Doob. "What Is a Martingale?," *American Mathematical Monthly* **78**: 451-462, 1971.

[Edwards *et al.*, 1963] W. Edwards, H. Lindeman, and L. Savage. "Bayesian Statistical Inference for Psychological Research," *Psychological Review* **70**: 193-242, 1963.

[Fermat and Pascal, 1679] P. Fermat and B. Pascal. "Correspondence," *Varia Opera Mathematica D. Petri de Fermat*: 179-188. Toulouse, 1679. Available in English translation on the web as "Fermat and Pascal on Probability," `http://www.york.ac.uk/depts/maths/histstat/pascal.pdf`.

[Field, 1978] H. Field. "A Note on Jeffrey Conditionalization," *Philosophy of Science* **45**: 361–367, 1978.

[Fisher, 1922] R. A. Fisher. "On the Mathematical Foundations of Theoretical Statistics," *Philosophical Transactions of the Royal Society,* **Series A 222**: 309-368 1922.

[Fisher, 1959] R. A. Fisher. *Statistical Methods and Scientific Inference*. Edinburgh: Oliver and Boyd, 1959.

[Gaifman, 1986] H. Gaifman. "A Theory of Higher Order Probabilities," *Proceedings of the 1986 Conference on Theoretical Aspects of Reasoning about Knowledge*: 275-292. San Francisco: Morgan Kaufmann Publishers, 1986.

[Garber, 1980] D. Garber. "Field and Jeffrey Conditionalization," *Philosophy of Science* **47**: 142-145, 1980.

[Gibbard, 2008] A. Gibbard. "Rational Credence and the Value of Truth," in T. Szabó Gendler and J. Hawthorne, eds., *Oxford Studies in Epistemology*, vol. 2: 143-164, 2008.

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

[Grove and Halpern, 1997] A. Grove and J. Y. Halpern. "Probability Update: Conditioning vs. Cross Entropy," *Proceedings of the Thirteenth Annual Conference on Uncertainty in Artificial Intelligence*: 208-214, 1997.

[Hall, 1994] N. Hall. "Correcting the Guide to Objective Chance," *Mind*, **103**: 504-17, 1994.

[Hall, 2004] N. Hall. Two Mistakes About Credence and Chance," *Australasian Journal of Philosophy* **82**: 93-111, 2004.

[Halpern, 1999] J. Y. Halpern. "Cox's Theorem Revisited," *Journal of Artificial Intelligence Research* **11**: 429–435, 1999.
[Hájek, 2008] A. Hájek. "Arguments For – Or Against – Probabilism?," *The British Journal for the Philosophy of Science* **59**: 793-819, 2008.
[Hewitt and Savage, 1955] E. Hewitt and L. J. Savage. "Symmetric Measures on Cartesian Products," *Transactions of the American Mathematical Society*, **80**: 470–501, 1955.
[Howso, 2008] C. Howson. "De Finetti, Countable Additivity, Consistency and Coherence," *British Journal for the Philosophy of Science* **59**: 1-23, 2008.
[Howson and Urbach, 1989] C. Howson and P. Urbach. *Scientific Reasoning: The Bayesian Approach*. La Salle: Open Court, 1989.
[Jaynes, 1968] E. Jaynes. "Prior Probabilities," *IEEE Transactions on Systems Science and Cybernetics,* **SSC-4**: 227-241, 1968.
[Jaynes, 1973] E. Jaynes. "The Well-Posed Problem," *Foundations of Physics* **3**: 477-493, 1973.
[Jaynes, 2003] E. Jaynes. *Probability Theory: The Logic of Science*. Cambridge, U.K.: Cambridge University Press, 2003.
[Jeffrey, 1983] R. Jeffrey. *The Logic of Decision*, revised 2nd edition. Chicago: University of Chicago Press, 1983.
[Jeffrey, 1983a] R. Jeffrey. "Bayesianism with a Human Face," In *Testing Scientific Theories*, edited by J. Earman, *Minnesota Studies in the Philosophy of Science* **10**. Minneapolis: University of Minnesota Press 1983.
[Jeffrey, 1987] R. Jeffrey. "Indefinite Probability Judgment: A Reply to Levi," *Philosophy of Science* **54**: 586-591, 1987.
[Jeffreys, 1939] H. Jeffreys. *Theory of Probability*. Oxford: Clarendon Press, 1939.
[Joyce, 1998] J. M. Joyce. "A Nonpragmatic Vindication of Probabilism," *Philosophy of Science* **65**: 575-603, 1998.
[Joyce, 1999] J. M. Joyce. *The Foundations of Causal Decision Theory*. New York: Cambridge University Press, 1999.
[Joyce, 2005] J. M. Joyce. "How Degrees of Belief Reflect Evidence," *Philosophical Perspectives* **19**: 153-179, 2005.
[Joyce, 207] J. M. Joyce. "Epistemic Deference: The Case of Chance," *Proceedings of the Aristotelian Society*, **107**: 1-20, 2007.
[Joyce, 2009] J. M. Joyce. "Accuracy and Coherence: Prospects for an Alethic Epistemology of Partial Belief," in F. Huber and C. Schmidt-Petri, eds., *Degrees of Belief*: 263-300. Berlin: Springer, 2009.
[Kaplan, 1996] M. Kaplan. *Decision Theory as Philosophy*. Cambridge: Cambridge University Press, 1996.
[Kelly, 2008] T. Kelly. "Disagreement, Dogmatism, and Belief Polarization," *Journal of Philosophy* **105**: 611-633, 2008.
[Keynes, 1921] J. M. Keynes. *A Treatise of Probability*. London: Macmillian, 1921.
[Koopman, 1940] B. O. Koopman. "The Bases of Probability," *Bulletin of the American Mathematical Society* **46**: 763-774, 1940.
[Kraft *et al.*, 1959] C. Kraft, J. Pratt, and A. Seidenberg. "Intuitive Probability on Finite Sets," *Annals of Mathematical Statistics* **30**: 408-19, 1959.
[von Kries, 1871] J. von Kries. *Die Principien der Wahrscheinlichkeitsrechnung*, 2nd ed., Tübingen, 1871.
[Lange, 2000] M. Lange. "Is Jeffrey Conditionalization Defective By Virtue of Being Non-Commutative? Remarks on the Sameness of Sensory Experience," *Synthese* **123**: 393-403, 2000.
[Laplace, 1774] P. Laplace. "Mémoire sur la probabilité des causes par les évenements," *Mémoires de l'Académie royale des sciences presentés par divers sarans* **6**: 621-56, 1774.
[Lee, 1997] P. M. Lee. *Bayesian Statistics: An Introduction*. New York: Wiley 1997.
[Levi, 1980] I. Levi. *The Enterprise of Knowledge*. Cambridge, Mass.: MIT Press, 1980.
[Lewis, 1980] D. Lewis. "A Subjectivist's Guide to Objective Chance," 1980. reprinted in *Philosophical Papers: Volume II*. New York: Oxford University Press, 1986. All page references are to the 1986 publication.
[Lewis, 1994] D. Lewis. "Humean Supervenience Debugged," *Mind*, **103**: 473-90, 1994.
[Lieb *et al.*, unpublished] E. H. Lieb, D. Osherson, J. Predd, V. Poor, S. Kulkarni, and R. Seiringer. "Probabilistic Coherence and Proper Scoring Rules", unpublished.

[Lindley, 1982] D. Lindley. "Scoring Rules and the Inevitability of Probability," *International Statistical Review* **50**: 1-26, 1982.
[Maher, 2002] P. Maher. "Joyce's Argument for Probabilism," *Philosophy of Science* **96**: 73-81, 2002.
[Martin-Löf, 1966] P. Martin-Löf. "On the Concept of a Random Sequence," *Information and Control* **9**: 602-619, 1966.
[Meacham, 2005] C. J. G. Meacham. "Three Proposals Regarding a Theory of Chance," *Philosophical Perspectives*, **19** (*Epistemology*): 281-307, 2005.
[Murphy, 1973] A. H. Murphy. "A New Vector Partition of the Probability Score," *Journal of Applied Meteorology* **12**: 595-600, 1973.
[Neyman, 1950] J. Neyman. *First Course in Probability and Statistics*. New York: Henry Holt, 1950.
[Paris, 1994] J. B. Paris. *The Uncertain Reasoner's Companion*. Cambridge, U.K.: Cambridge University Press, 1994.
[Popper, 1959] K. Popper. *The Logic of Scientific Discovery*. London: Hutchinson, 1959.
[Ramsey, 1931] F. Ramsey. "Truth and probability," in *Foundations of Mathematics and other Logical Essays*. London: Kegan Paul, 1931.
[Reichenbach, 1948] H. Reichenbach. *The Theory of Probability, an Inquiry into the Logical and Mathematical Foundations of the Calculus of Probability*. Berkeley: University of California Press, 1948.
[Rényi, 1955] A. Rényi. "On a New Axiomatic Theory of Probability," *Acta Mathematica Academiae Scientiarium Hungaricae* **6**: 285-335, 1955.
[Savage, 1954] L. J. Savage. *Foundations of Statistics*. New York: Wiley, 1954.
[Savage, 1971] L. J. Savage. "Elicitation of Personal Probabilities," *Journal of the American Statistical Association* **66**: 783-801, 1971.
[Scott, 1964] D. Scott. "Measurement Structures and Linear Inequalities," *Journal of Mathematical Psychology* **1**: 233-247, 1964.
[Seidenfeld, 1985] T. Seidenfeld. "Calibration, Coherence, and Scoring Rules," *Philosophy of Science* **52**: 274-294, 1985.
[Shimony, 1988] A. Shimony. "An Adamite Derivation of the Calculus of Probability," in J. H. Fetzer, ed., *Probability and Causality*: 151-161. Dordrecht: D. Reidel, 1988.
[Skyrms, 1980] B. Skyrms. "Higher Order Degrees of Belief," in D. Mellor, ed., *Prospects for Pragmatism*. Cambridge: Cambridge University Press, 1980.
[Skyrms, 1984] B. Skyrms. *Pragmatics and Empiricism*. New Haven: Yale University Press, 1984.
[Strevens, 1995] M. Strevens. "A Closer Look at the 'New' Principle," *British Journal for the Philosophy of Science*, **46**: 545-56, 1995.
[Suppes and Zanotti, 1976] P. Suppes and M. Zanotti. "Necessary and Sufficient Conditions for the Existence of a Unique Measure Strictly Agreeing with a Qualitative Probability Ordering," *Journal of Philosophical Logic* **5**: 431-38, 1976.
[Thau, 1994] M. Thau. "Undermining and Admissibility," *Mind*, **103**: 491-503, 1994.
[van Fraassen, 1981] B. van Fraassen. "A Problem for Relative Information Minimizers in Probability Kinematics," *British Journal for the Philosophy of Science* **32**: 375–37, 1981.
[van Fraassen, 1983] B. van Fraassen. "Calibration: A Frequency Justification for Personal Probability," in R. Cohen and L. Laudan, eds., *Physics Philosophy and Psychoanalysis*: 295-319. Dordrecht: D. Reidel, 1983.
[Villegas, 1964] C. Villegas. "On Qualitative Probability σ-Algebras," *Annals of Mathematical Statistics* **35**: 1787-96, 1964.
[Venn, 1866] J. Venn. *The Logic of Chance*. London, Macmillan, 1866.
[von Mises, 1957] R. von Mises. *Probability, Statistics and Truth*. New York: Macmillan, 1957.
[Wagner, 2002] C. Wagner. "Probability Kinematics and Commutativity," *Philosophy of Science* **69**: 266-278, 2002.
[Walley, 1991] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. New York: Chapman and Hall, 1991.
[Williamson, 2000] T. Williamson. *Knowledge and its Limits*. Oxford: Oxford University Press, 2000.