Ockham's Razor, Hume's Problem, Ellsberg's Paradox, Dilation, and Optimal Truth Conduciveness

Kevin T. Kelly

December 2, 2008

Abstract

Classical Bayesianism represents ignorance, if at all, by flatness of prior probabilities. Such probabilities are an essential part of the standard Bayesian explanation of Ockham's razor. But flatness as a model of ignorance is called into question by Ellsberg's paradox, which has led to the consideration of incoherent or inexact degrees of belief, both of which undermine the usual explanation of Ockham's razor. An alternative explanation of Ockham's razor is presented, according to which always favoring the uniquely simplest theory compatible with experience keeps one on the shortest or most direct path to the truth. It turns out that minimization of total distance to the truth implies coherent degrees of belief strongly biased toward simplicity. If one focuses on retractions or drops in credence, then a more reasonably moderate bias toward simplicity results but optimal efficiency then demands either incoherence or inexact probabilities, both of which are solutions to Ellsberg's paradox. Finally, it turns out that dilation, or increasing imprecision in light of new information, is necessary if agents with inexact probabilities are to minimize total retractions. So, in place of paradox and tension, one obtains a unified perspective on Ockham's razor, Ellsberg's paradox, dilation, and the justification of inductive inference.

1 Introduction

This paper explores some new connections between several topics of interest to Henry Kyburg: inductive inference, simplicity, imprecise probabilities, and the objective truth conduciveness of scientific inference (1961, 1983). Ockham's razor is a popular name for the scientific aversion to theoretical complexity. Ockham's razor occasions the obvious question how a fixed bias toward simplicity could possibly help one find possibly complex truths. I argue that the usual, Bayesian explanation of the truth conduciveness of simplicity in inductive inference begs the question by presupposing a prior probabilistic bias toward simple possibilities. An alternative explanation of Ockham's razor is presented, according to which Ockham's razor keeps one on the shortest or straightest cognitive path to the truth. In order to minimize cognitive distance traveled to the truth, one must be coherent and one must start out with a very strong prior bias toward simplicity. In order to minimize retractions of credence en route to the truth, on the other hand, one's bias toward simplicity can be more plausibly moderate, but one must resort occasionally either to sub-additivity or to to inexact probabilities in order to retreat to a genuine state of ignorance rather than to a "flat" Bayesian prior. Sub-additivity and inexact probabilities have also been invoked as possible solutions to Ellsberg's decision theoretic paradox, which elicits intuitions inconsistent with Bayesian coherence and has been interpreted as calling "flat" probabilities into question as an adequate model of ignorance. Thus, efficient convergence to the truth via Ockham's razor yields a new, truth-directed, normative argument for Ellsberg intuitions. There is another "paradox" associated with inexact probabilities, which can become even *less* exact in light of *more* data. That phenomenon, known as *dilation*, is shown to be a natural and necessary condition for the retraction-efficiency of Ockham agents employing inexact probabilities. Thus, optimal truth conduciveness also provides a new, truth-directed explanation of dilation.

2 Synchronic Truth-Conduciveness

Coherence is a mere matter of beliefs "fitting together", whereas truth is a special relationship between one's beliefs and the world. Coherence seems to have no more to to do with truth than the soundness of a ship's framing has to do with the ship being at its intended port. But it is still the case that a ship that founders due to rot or design flaws must first be raised and repaired before it can proceed to port, wherever the intended port might be. An analogous argument can be given for the truth conduciveness of Bayesian coherence, where coherence corresponds to remaining afloat and the destination port corresponds to the unknown truth (DeFinetti 1972).[1] Suppose that you are interested in three mutually exclusive and exhaustive theories T_1, T_2, T_3. Then, as far as these theories are concerned, your degrees of belief can be reduced to a 3-dimensional vector:

$$\mathbf{b} = (b(T_1), b(T_2), b(T_3)).$$

Then extremal degrees of belief and assignments of truth values to the respective theories can both be represented by the *basis vectors*:

$$\begin{aligned}\mathbf{i}_1 &= (1, 0, 0);\\ \mathbf{i}_2 &= (0, 1, 0);\\ \mathbf{i}_3 &= (0, 0, 1).\end{aligned}$$

[1] Rosenkrantz (1981) conjectures that the result works whenever the loss function optimizes the expected loss of one's own degrees of belief. Joyce (1998) provides axiomatic sufficient conditions for the argument. Maher (2002) objects to the premises of Joyce's argument.

Consider the set of all "coherent" belief states—the vectors $\mathbf{b} = (x, y, z)$ for which $x + y + z = 1$. These triples pick out an equilateral triangle with corners $\mathbf{i}_1, \mathbf{i}_2, \mathbf{i}_3$ in three-dimensional Euclidean space whose corners are the basis vectors, corresponding to possible truth assignments (figure 1). Now, suppose that your degrees of belief $\mathbf{b}$ do not lie on the triangle. Consider alternative, "coherent" degrees of belief $\mathbf{b}'$ that result from projecting $\mathbf{b}$ orthogonally onto the surface of the triangle. Focus, for now, on truth assignment $\mathbf{i}_2$. The Euclidean distance from $\mathbf{b}$ to $\mathbf{i}_2$ is α, which is clearly longer than the distance β from $\mathbf{b}'$ to $\mathbf{i}_2$.[2] The same relationship holds for each possible truth assignment $\mathbf{i}$. In decision theory, option A_1 *dominates* option A_2 just in case A_1 achieves better results than A_2 in each possible world. Thus, your "incoherent" degrees of belief $\mathbf{b}$ are dominated in terms of distance from the truth by coherent degrees of belief $\mathbf{b}'$.[3] So, if you happen to discover that your degrees of belief $\mathbf{b}$ are incoherent, you should proceed, immediately, to adopt degrees of belief that are *admissible* (i.e., not dominated) in distance to the truth; and

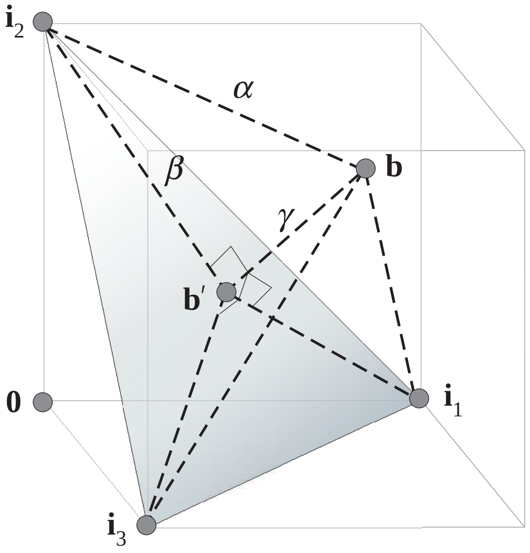

Figure 1: Coherence and static distance to the truth.

[2] By the triangle inequality and the assumption that γ is non-zero).

[3] DeFinetti actually employs squared Euclidean distance, which is minimized if and only if Euclidean distance is minimized. Rosenkrantz (1981) and Joyce (1988) discuss axiomatic generalizations of squared Euclidean distance, although neither can say how general their axiomatizations happen to be. Squared Euclidean distance has the comforting property of being a *proper scoring rule* in the sense that the expected distance from the truth of one's own belief profile **b** always turns out to look best from the viewpoint of **b**, so that the scoring rule would never leave one unsatisfied with one's current position. Joyce does not insist on that principle and neither do I. To insist on proper scoring rules is evidently to beg the question in favor of a fully subjective Bayesianism that imposes no normative restrictions on prior probability.

all such degrees of belief lie on the triangle of coherence. If, in addition, you want to revise your current degrees of belief as little as possible—a plausible, "epistemic" consideration—you should adopt the orthogonal projection $\mathbf{b}'$ of $\mathbf{b}$ onto the triangle. In particular, a Pyrrhonistic skeptic who adopts belief profile $\mathbf{0} = (0, 0, 0)$ to reflect total disbelief in everything is dominated in distance to the truth by the uniform probability measure $\mathbf{d} = (1/3,\ 1/3,\ 1/3)$, which seems to support the plausible idea that ignorance should be represented coherently as a flat probability distribution.

Inquiry is more than a mere matter of remaining afloat: it involves navigation to the intended destination—the truth. In lieu of a compass, scientists employ Ockham's razor, an instinctive bias against needless theoretical complexity. But how could a fixed bias toward simplicity help science arrive at possibly complex truths? I consider that question next.

3 Ockham's Razor

Ockham's razor is a central, characteristic, and indispensable principle of scientific inference. The rule is roughly that, when faced with a choice between alternative theories compatible with experience, one should select or favor only the simplest one, where simplicity is a matter of minimizing causes, entities, independent principles, or free parameters. Since at least the time of Kant, it has been a foundational puzzle how such a systematic bias could help one find hidden, possibly complex truths. The argument couldn't be that a bias toward complexity is *dominated* in distance to the truth by a bias toward simplicity, since the truth might be complex.

Bayesians have an easy solution to the problem: just impose higher prior degrees of belief on simpler theories, in which case simpler theories are "more probably true".[4] But that just pushes the question back by a trivial step, for why, in the interest of finding the truth, *should* one start with higher degrees of belief on simpler theories?

There is a more subtle Bayesian explanation that focuses on prior likelihoods rather than prior probabilities (Rosenkrantz 1981). It is an immediate consequence of Bayes theorem that for arbitrary probability measure p:

$$\begin{aligned} \frac{p(T_1 \mid e)}{p(T_2 \mid e)} &= \frac{p(T_1)}{p(T_2)} \cdot \frac{p(e \mid T_1)}{p(e \mid T_2)}; \\ &= c \cdot \frac{p(e \mid T_1)}{p(e \mid T_2)}. \end{aligned}$$

The first factor of the product, the ratio of prior probabilities, is constant as evidence increases. Therefore, the crucial factor governing the impact of data e on belief is the second factor of the product, which is called the *Bayes factor*

[4] That is known as the "minimum description length" (MDL) principle (Rissannen 1983). It should be mentioned, however, that some advocates of MDL, including Rissannen, view the aim of inquiry as syntactic compression of the data rather than as finding the truth.

(Kass and Raftery 1995) for T_1, T_2, e. It is tempting to say that $p(e \mid T)$ is *objective* if T happens to be a statistical theory that entails some unique chance $c_T(e)$ for e, in which case, in the absence of other relevant information:[5]

$$p(e \mid T) = c_T(e).$$

But that is no longer the case if T is a *composite* statistical hypothesis with a free parameter θ:

$$T \equiv (\exists\theta)\ R_\theta,$$

for in that case:

$$\begin{aligned} p(e \mid T) &= \int p(e \mid R(\theta) \cdot p(R(\theta) \mid T) \quad d\theta \\ &= \int c_{R_\theta}(e) \cdot p(R(\theta) \mid T) \quad d\theta. \end{aligned}$$

The objective chances $c_{R_\theta}(e)$ are now *weighted* by the subjective prior beliefs $p(R_\theta \mid T)$ so that $p(e \mid T)$ is a *subjective* quantity. That point is crucial in connection with Ockham's razor, since many observers, including Kyburg (1961), have understood simplicity in terms of minimization of free parameters. Suppose, for an elementary example, that theory T_1 predicts phenomena e outright and that $T_2 \equiv (\exists\theta)\ R_\theta$ has discrete, free parameter θ ranging over the natural numbers from 0 to n, so that:

$$T_2 \equiv (R_0 \text{ or } R_1 \text{ or } \ldots \text{ or } R_n).$$

Moreover, suppose that R_0 entails e and every other disjunct is refuted by e. It is intuitive to say that e "severely tests" T_1, but is merely used to "set" the value of the free parameter θ in T_2 to the value $\theta = 0$. Setting $\theta = 0$ in light of e is an "ad hoc" response to e that merely "accommodates" e. One can spice the discussion with militaristic, Popperian metaphors about T_1 "passing muster", "running the gauntlet" or "proving its mettle" while T_2 "shrinks from the fray". No medals of valor for T_2.

The apparent virtue of surviving a severe test as opposed to mere accommodation is explained by the Bayes factor, assuming that there is not too much bias toward T_2:

$$p(T_1) \approx p(T_2);$$

and, equally importantly, that there is not too much subjective bias toward some value of the free parameter θ given that T_2 is true, so that:

$$p(R_i \mid T_2) \approx p(R_j \mid T_2),$$

[5] This ignores the considerable problem, of central interest to Kyburg, of what counts as other relevant information.

where i, j are distinct Boolean values. For then one has:

$$\begin{aligned}
\frac{p(T_1 \mid e)}{p(T_2 \mid e)} &= \frac{p(T_1)}{p(T_2)} \cdot \frac{p(e \mid T_1)}{p(e \mid T_2)}; \\
&\approx \frac{p(T_1)}{p(T_2)} \cdot \frac{1}{\sum_{i \leq n} p(e \mid R_i) \cdot p(R_i \mid T_2)}; \\
&= \frac{p(T_1)}{p(T_2)} \cdot \frac{1}{1/n} = n.
\end{aligned}$$

So the more equally plausible ways a theory can be true and the fewer of these ways that explain the data, the worse the theory does compared to a simple competitor that explains the same data without adjustable parameters. As the number of parameter values approaches infinity, the advantage becomes overwhelming.

4 Ignorance vs. Ignoredge

The Bayes factor computation seems to explain, from a standpoint of pure ignorance, why the simpler theory is "more likely to be true" than the complex theory in light of simple evidence. But it still hinges on a prior bias against complexity. Ignorance between T_1 and T_2 means that $p(T_1) \approx p(T_2)$ and ignorance about the true value of the parameter i given T_2 yields $p(R_i \mid T_2) \approx p(R_j \mid T_2)$, for distinct, Boolean i, j. Additivity then enforces a strong prior bias in favor of $p(T_1)$ over $p(R_0)$ (figure 2). That should give one pause. Objectively speaking, T_1 and R_0 entail the same predictions:

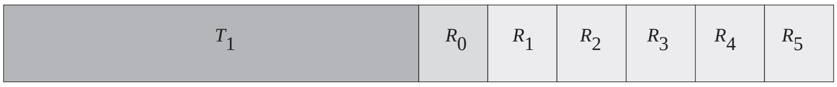

Figure 2: Bayes factor argument.

$$p(e \mid T_1) \approx 1 \approx p(e \mid R_0).$$

Therefore, the *only* reason T_2 does not end up as "likely to be true" as T_1 in light of e in the Bayes factor computation is the strong prior bias $p(T_1) \gg p(R_0)$, for:

$$\frac{p(T_1 \mid e)}{p(R_0 \mid e)} = \frac{p(T_1)}{p(R_0)} \cdot \frac{p(e \mid T_1)}{p(e \mid R_0)} \approx \frac{p(T_1)}{p(R_0)}.$$

Had one adopted, instead, an "ignorant" distribution over possibilities $T_1, R_0, \ldots, R_n$, the result would be a strong bias toward T_2 over T_1 and parity between T_1 and R_0 after updating on e (figure 3). So, after all, the Bayes factor explanation of Ockham's razor reduces to the selection of a prior bias against *complex possibilities* over a prior bias against *simple theories.* There is no neutral place

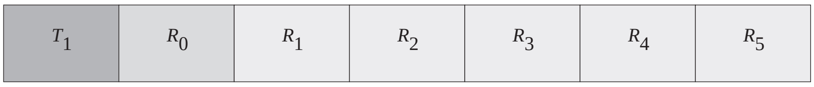

Figure 3: Alternative "ignorance".

to stand: a skeptic who confesses utter ignorance is offered a fool's bargain between two sharp biases. The point is the familiar: "indifference" depends on the question one asks: "indifference" regarding blue vs. non-blue induces a strong bias against yellow and "indifference" over a range of colors induces a strong bias against blue. The inconsistencies that arise when "indifference" is imposed simultaneously over coarser and more refined partitions of the underlying possibilities are known collectively as *paradoxes of indifference.*

The paradoxes of indifference are sometimes presented as arguments in *favor* of subjective Bayesianism: since there is no unique thing as true ignorance, anything goes as far as prior biases are concerned—indeed, *some* such bias is rationally compulsory. In applications, however, Bayesians sometimes slide into describing the resulting biases as "knowledge" or "information" and emphasize the ability of the Bayesian framework to incorporate the agent's "information" in a smooth and uniform way. In some concrete applications, some aspects of prior probability can represent a kind of knowledge or information. But when science applies Ockham's razor to theory choice in an alien application such as subatomic particles—precisely where Ockham's razor is most indispensable—it is hard to view a sharp bias against complex possibilities as arising from anything but pure ignorance. Therefore, I coin the term *ignoredge* to better describe the equivocal status of prior probabilities in Bayesian explanations of Ockham's razor.

5 Ignoredge and Ellsberg's Paradox

The essential difference between ignorance and ignoredge is laid bare in a celebrated counterexample to expected utility theory due to Daniel Ellsberg (1961). Subjects are informed that an urn contains thirty balls, ten of which are color 1, and twenty of which are either color 2 or color 3. Let T_i be the proposition that the color is i. There is a ninety dollar prize at stake and subjects are asked to choose one gamble from each of the following pairs:

$$\begin{array}{rcl} T_1 & \text{vs.} & T_2, \\ T_1 \text{ or } T_3 & \text{vs.} & T_2 \text{ or } T_3. \end{array}$$

Most humans choose T_1 from the first pair and T_2 or T_3 from the second pair, for a fairly obvious reason: the objective chance of winning is at least 1/3 for T_1 but could be zero for T_2, whereas the objective chance of winning is at least 2/3 for T_2 or T_3 but could be as low as 1/3 for T_1 or T_3. These preferences are incompatible with expected utility theory, which affords no way to represent the intuitive distinction between pure ignorance between three possibilities and knowledge that one of the possibilities occurs with chance 1/3. Note that since the colors are mutually exclusive:

$$\begin{aligned} p(T_1) > p(T_2) \quad & \text{iff} \quad p(T_1) + p(T_3) > p(T_2) + p(T_3) \\ & \text{iff} \quad p(T_1 \text{ or } T_3) > p(T_2 \text{ or } T_3). \end{aligned}$$

Moreover, subjects persist unabashedly in their preferences when their deviation from expected utility theory is explained. Perhaps, everyone is irrational—or maybe Bayesian ignoredge equivocates fatally between knowledge and ignorance.

One potential, *neo*-bayesian resolution of the paradox (e.g., Schmeidler 1989) is to drop the probabilistic axiom of finite additivity in favor of finite *sub*-additivity, defined by:

$$p(A) + p(B) \leq p(A \text{ or } B) + p(A \text{ and } B).^{6}$$

The Ellsberg preferences can be recovered by computing "expected utility" with respect to sub-additive degrees of belief **b** defined by:[7]

$$\begin{aligned} \mathbf{b}(T_1) &= 1/3; \\ \mathbf{b}(T_2) &= 0; \\ \mathbf{b}(T_3) &= 0; \\ \mathbf{b}(T_1 \text{ or } T_3) &= 1/3; \\ \mathbf{b}(T_2 \text{ or } T_3) &= 2/3. \end{aligned}$$

In figure 4, $\mathbf{b} = (1/3, 0, 0)$ is seen to fall off of the triangle of coherent measures. Now, pyrrhonistic ignorance can be represented as $\mathbf{0} = (0, 0, 0)$, as opposed to the Bayesian state of ignoredge $\mathbf{d} = (1/3, 1/3, 1/3)$. That also addresses the paradoxes of indifference, for in the coarsened partition T_1 vs. T_2 or T_3, sub-additive ignorance can still be represented by $(0, 0)$, whereas ignoredge induces the bias: $(1/3, 2/3)$.[8]

Another possible neo-bayesian response to the Ellsberg paradox (e.g., Levi 1974, Gilboa, Schmeidler 1989, and Walley 1991) is to allow for inexact probabilities modeled as interval-valued probabilities or as sets of probabilities. The

[6] This property is called "convexity" in (Schmeidler 1989). Schmeidler provides a unique representation theorem for sub-additive probabilities in terms of ratioal preference.

[7] In fact, an arbitrarily small violation of additivity suffices to recover the Ellsberg phenomenon. Note, also, that an incoherent belief profile b is not uniquely determined by the vector **b** of values on partition T_1, T_2, T_3. It is understood in such cases that **b** is defined for every disjunction of theories.

[8] Although the prospect theory of (Kahneman and Tversky 1979) allows for sub-additive weights on probabilities, Fox and Tversky (1995) do not invoke sub-additive probabilities to explain Ellsberg's paradox because of results suggesting that the values of the uncertain Ellsberg's gambles go down only in comparison with the risk version of the same gamble. The empirical results are disputed by Arló-Costa and Helzner (2008). But, in any event, Fox and Tversky do not view their treatment of the phenomenon as normative.

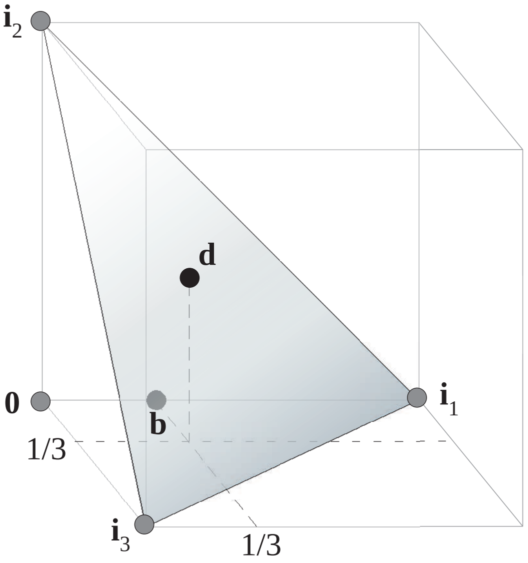

Figure 4: Ignorance as sub-additive probability

two ideas are related,[9] for each set B of probabilities determines *upper* and *lower* probabilities:[10]

$$\underline{b}_B(T) = \inf_{b \in B} \ p(T);$$
$$\bar{b}_B(T) = \sup_{b \in B} \ p(T);$$

and, thus, the interval-valued probability function:

$$b_B(T) = (\bar{b}_B(T), \underline{b}_B(T)).$$

Kyburg (1983) explicitly recommended that the intervals reflect known intervals around objective chances. In the case of the Ellsberg problem, that idea picks out the set S of all probability measures compatible with the initial information provided, namely, that $b(A) = 1/3$. Set B, representing genuine ignorance about T_2, T_3, is depicted in figure 5 along with measure $\mathbf{d}$, representing ignoredge in the same situation. The sub-additive function $\mathbf{b}$ depicted in figure 5 is seen to be the projection of the end-points of B. Then:

[9] For a detailed discussion, see the introduction to (Gilboa and Schmeidler 1989).

[10] The set picture and the interval picture are not identical, since each set B determines b_B uniquely, but b_B does not determine B uniquely. That can matter decision theoretically, For example, Isaac Levi (1980) and (Seidenfeld et al. 1999) recommend admissibility over B (i.e., elimination of dominated alternatives over B). Dominance can depend upon exactly which set B is chosen to represent p_B. But the results that follow don't depend on these differences.

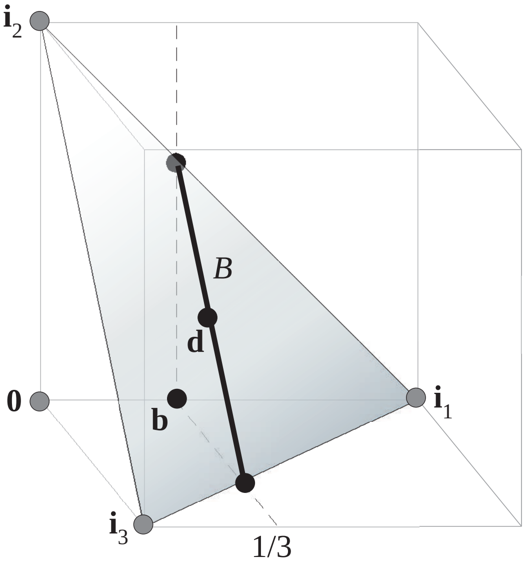

Figure 5: Ignorance as imprecise probability

$$\begin{aligned}
b_B(T_1) &= (1/3,\ 1/3); \\
b_B(T_2) &= (0,\ 2/3); \\
b_B(T_3) &= (0,\ 2/3); \\
b_B(T_1 \text{ or } T_3) &= (1/3,\ 2/3); \\
b_B(T_2 \text{ or } T_3) &= (2/3,\ 2/3).
\end{aligned}$$

Each probability measure b in B determines an expected value for each bet mentioned in the Ellsberg choices. The Ellsberg preferences are then recovered by always selecting the bet whose minimum expectation over B is maximum (i.e., by maximin-ing expectations over B).[11] Following that rule, the minimum expected value for betting on T_1 is 30 dollars (recall that the prize is ninety dollars) whereas the minimum expected value for betting on T_2 is 0 dollars. In the second choice, the minimum expected value of betting on T_1 or T_3 is 30 dollars whereas the minimum expected value of betting on T_2 or T_3 is 60 dollars.[12] Like sub-additivity, inexact probabilities allow for a clear distinction between ignorance and ignoredge, for if B contains every probability measure

[11] The maximin rule is recommended by I. Levi *after* dominated alternatives are eliminated. Again, the relative merits of that view are not directly relevant to our discussion. Other advocates of sets of probabilities (Seidenfeld et al.1999) do not recommend any uncertain choice rules beyond admissibility over B, so for them both the Ellsberg preference and the usual preference are admissible.

[12] The same preferences emerge from less draconian ignorance: as long as B contains any two points on the dark line on either side of the mid-point of the triangle the same preferences result.

than the probability intervals are vacuously $(0, 1)$ for each possible disjunction of propositions T_1, T_2, T_3.

Distinguishing ignorance from ignoredge in some manner or other is a good idea. But then one must also revisit the pivotal role played by ignoredge in the Bayes factor explanation of Ockham's razor. Sub-additivity allows one to set:

$$b(T_1) = b(T_2) = b(R_0) = \ldots = b(R_n) = 0.$$

Now a true, pyrrhonistic ignoramus is no longer forced to play favorites either at the level of theories or at the level of parameter settings, so the Bayes' factor argument for Ockham's razor based on ignoredge collapses. Similarly, a true ignoramus can adopt the set of all probability measures over the algebra of propositions generated by $T_1, R_0, \ldots, R_n$, respectively. Again, symmetry is respected in both partitions and the Bayes' factor explanation collapses.

So extensions of classical Bayesianism that deal with the Ellsberg paradox undermine the explanation of Ockham's razor. Does the Ellsberg paradox, therefore, unmask Ockham's razor as an illusion born of a misguided model of ignorance?

6 Ockham's Razor and Truth Conduciveness

The Bayesian explanation of Ockham's razor presupposes a counter-intuitive model of ignorance. But aside from that, the crux of the simplicity puzzle is to explain, without unedifying circularity, why a prior bias toward simplicity is better for *finding the truth*, whatever the truth might happen to be. That may seem hopeless, for how could favoring a complex theory do worse if that very theory happens to be true?

Nonetheless, an argument can be given. Inquiry is pursuit of truth. A strategy of pursuit is more *conducive* to its goal insofar as it keeps the pursuer on a shorter or more direct path to the goal. The best path may, perforce, reverse course any number of times if there are obstacles in the way or if some search or guesswork is required. Nonetheless, the strategy of pursuit that keeps one on the most direct cognitive path to the truth is still the most truth conducive and is, therefore, justified by the aim of finding the truth. In fact, one can argue that Ockham's razor is uniquely optimally truth-conducive in the sense of achieving the minimum achievable bound on the length of one's cognitive path to the truth. The argument may be viewed as a diachronic extension of the usual, static, distance-from-the-truth argument for probabilistic coherence described above.

Consider a simplistic example, in which you are watching a black box that may emit at most two marbles, at any times whatever. Let T_i be the proposition that the box emits at most $i - 1$ marbles, for i ranging from 1 to 3. This marble arrangement gives rise to nested problems of induction: any amount of marble free experience is compatible with another marble appearing later, until three marbles have been seen. Ockham's razor seems to suggest not counting your marbles until they appear, since marbles are "entities" and Ockham's razor prohibits one from positing entities without necessity. A pure implementation of Ockham's razor would, therefore, output belief profile $\mathbf{i}_1$ until the first marble is seen, belief profile $\mathbf{i}_2$ until the second marble is seen, and belief profile $\mathbf{i}_3$ thereafter, which is guaranteed to converge to the truth. This Ockham strategy corresponds to the dashed line around the edges of the triangle in figure 6. Each side of the triangle has length $\sqrt{2}$, so the Ockham strategy travels at most

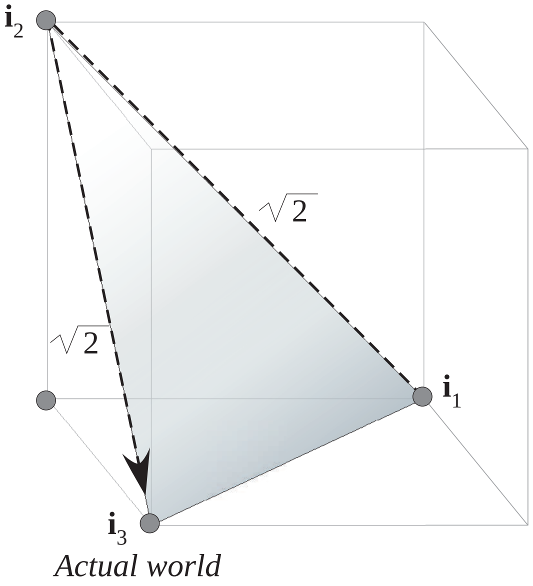

Figure 6: Ockham's razor

distance $(n-1)\cdot\sqrt{2}$ to the truth if the truth is $\mathbf{i}_n$, (for $n = 1, 2, 3$).

Furthermore, *no method that converges to the truth can do better*. For suppose that nature withholds marbles forever. Then the convergent method must move ever closer to $\mathbf{i}_1$ on pain of failure to converge to the truth. When the convergent method is close to $\mathbf{i}_1$, nature is free to extend the data presented so far with one marble followed by no more marbles. Again, on pain of failure to converge to the truth, the method must move ever closer to belief profile $\mathbf{i}_2$. When the method is arbitrarily close to $\mathbf{i}_2$, nature can present another marble, forcing the method to converge to $\mathbf{i}_3$. Thus, if the truth is $n \in \{1, 2, 3\}$, then the method travels a distance arbitrarily close to $(n-1)\cdot\sqrt{2}$.

Figure 7 depicts garden-variety Bayesian performance. Starting with some uncertainty between the three theories, a long run of marble-free experience results in movement toward T_1. Seeing the first marble refutes T_1, resulting in a leap to some point of uncertainty between T_2 and T_3. Failure to see more marbles results produces motion toward T_2. Seeing the final marble results in an immediate leap to T_3. The initial motion toward T_1 and the subsequent motion from T_3 toward T_2 are *added* to the motions performed by the simple

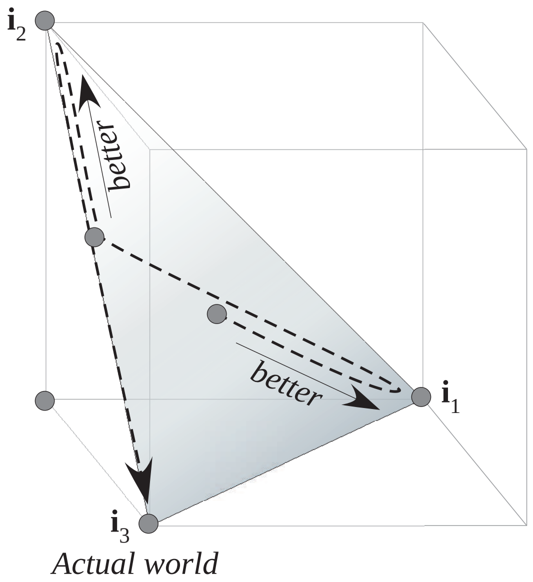

Figure 7: Bayesian approximation to Ockham

Ockham method depicted in figure 6. Indeed, by the triangle inequality, *any* deviation from the path taken by simple Ockham method follows a longer path to T_3. But the pure Ockham strategy corresponds to the *strongest possible* bias toward simplicity at every stage. So even if the truth is complex, the strongest possible bias toward the truth minimizes the length of one's path to the truth in the worst case (taking the worst case with respect to time of appearance of the two marbles).[13]

Next, consider a slightly different example. Suppose that the task is to determine how many marbles of each color you will ever see. Furthermore, you know that you will see at most one marble.[14] Ockham's razor favors T_1 a priori. Now, suppose that a marble is heard bouncing down the spout of the emitter before its color is seen. At that point, the only possible truths are T_2, T_3, so consistency with the data rules out probability mass on T_1. Ockham's razor cannot choose between $\mathbf{i}_2$ and $\mathbf{i}_3$, which are equally simple, so it adopts belief profile $\mathbf{d'} = (0, 1/2, 1/2)$. Ockham's razor then moves to $\mathbf{i}_2$ or to $\mathbf{i}_3$ depending on the color observed (cf. figure 8). If the color of the marble is seen immediately, Ockham's razor instructs you to proceed immediately to $\mathbf{i}_2$ or to $\mathbf{i}_3$, depending on the color observed. The method just described travels distance 0 if no marble

[13] Bayesian updating precludes the strongest possible bias toward simplicity, since no basis vector $\mathbf{i}$ can be changed by conditioning. So Bayesian updating can approach optimality but cannot ever achieve it exactly.

[14] The point of the restriction to one marble is merely to restrict the question to three answers so that you can continue to use the three dimensional convergence diagram. The point is general.

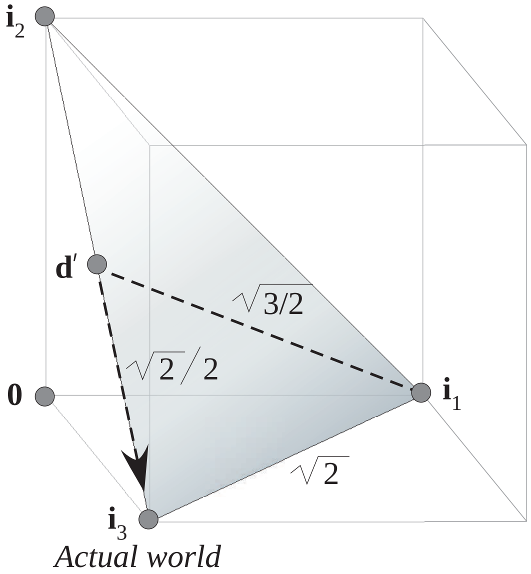

Figure 8: Ockham in suspense

appears, which is obviously optimal. It travels at worst distance $\sqrt{3/2}+\sqrt{2}/2$ if a marble is heard before it appears. Since nature is free to choose the color of the marble, waiting midway between $\mathbf{i}_2$ and $\mathbf{i}_3$ is the best an arbitrary convergent method that respects consistency with the data can do to reduce the length of its path to the truth.[15]

I refer to optimality results of the preceding sort as *Ockham efficiency theorems*. A few general comments are in order about such theorems.

(1) Even if you know only that you will see some finite number of marbles, moving directly from T_k to T_{k+1} incurs distance $\sqrt{2}$, because the Euclidean distance between $\mathbf{i}_k$ and $\mathbf{i}_{k+1}$) is still computed by taking the difference $\mathbf{i}_k - \mathbf{i}_{k'}$ and computing the Euclidean norm:

$$\|\mathbf{i}_k - \mathbf{i}_{k+1}\| = \sqrt{(1-0)^2 + (0-1)^2} = \sqrt{2}.$$

So the preceding arguments generalize: the theoretically optimal convergence distance bound over worlds in which n marbles appear is $(n-1)\sqrt{2}$, the bound achieved by the Ockham strategy of moving through the theories in order of increasing complexity.

(2) Among philosophers, I sometimes encounter the response that efficient convergence to the truth is a merely "pragmatic" consideration, as opposed to a

[15] Methods that violate consistency with the data can do better by waiting at $\mathbf{i}_1$ after T_1 is refuted and then leaping straight to T_2 or to T_3, depending on the color observed. Here, suffice it to say that consistency with the data is not the mystery to be explained—Ockham's razor is. More engagingly, it will be shown below that a different measure of cost allows one to explain both consistency with the data and Ockham's razor.

genuinely philosophical, “epistemic” consideration. The idea is that happy endings or threats are pragmatic reasons to believe whereas evidence is a properly epistemic reason to believe. But what would count as a purer or more Platonic explication of “epistemic justification” than optimal truth conduciveness? And what could be more truth conducive than a strategy that minimizes the length of one’s path to the truth?

(3) It may seem as though efficient convergence could not be the true justification of Ockham’s razor because justifications should promote credence in what they justify. The preceding argument allows for the unavoidable possibility of any number of arbitrarily bad surprises in the future after Ockham’s razor is applied, which may, upon reflection, decrease confidence in simple theories. Hence, it cannot be a justification. I respond that to specify rhetorical causes of belief is one thing and to justify such causes as the right or best causes for purposes of finding the truth is quite another. Ockham’s razor’s rhetorical force is a given. That force may engender vague hopes that simplicity is a magical divining rod that points directly and immediately toward hidden truths. A philosophical justification should be praised rather than discredited for displacing that sort of vague, unfounded, wishful thinking.

(4) The structure underlying the Ockham efficiency argument is a hybrid of maximin and dominance; namely, dominance in terms of worst-case performance bounds over each *empirical complexity class* of possible worlds, where empirical complexity corresponds simply to the number of marbles presented. That form of argument is commonplace in the computer science literature on algorithmic efficiency. One examines worst-case resource consumption of the algorithm over inputs of a given size and compares these bounds across algorithms.

(5) The Ockham efficiency argument could not be a pure dominance argument—an Ockham violator could anticipate a marble before it is seen and nature could be so kind as to present the marble before the violator loses confidence (as she must, if she converges to the truth at all). If empirical complexity is unbounded, neither could the argument be a pure maximin argument, since then the worst case distance bound would be infinite. Nor is the argument an expected case argument, and for good reason: modeling ignorance as uniform ignoredge forces either a question-begging bias against complex possibilities or a question-begging bias against simple theories. The worst-case logic of the Ockham efficiency theorem avoids both biases.

(6) Worst-case reasoning is admittedly brittle—for example, if you might possibly be killed if you leave your house, no information short of logical certainty in your safety could ever coax you out of doors. That would be a fair objection to the Ockham efficiency theorem if the intention were to show universally and unequivocally that Ockham’s razor must preempt every material consideration to the contrary, but the aim is far less ambitious: it is to obtain some sort of non-circular but truth-directed explanation why Ockham’s razor should serve as a defeasible default bias that may be preempted by genuine material considerations. For example, Ptolemy’s astronomy was preferred to Copernicus’ simpler account due to a mistaken presupposition that heliocentrism cannot be squared with terrestrial phenomena. The history of science is

replete with similar examples.[16]

(7) Another concern about worst-case reasoning in epistemological contexts is that it tends to yield implausibly timid or skeptical conclusions compared with expected case reasoning. But the Ockham efficiency argument yields the conclusion that a sharp bias toward simplicity is the best possible policy for finding possibly complex truths. That is not blandly skeptical; it is remarkable. *Too* remarkable, perhaps. It is plausible to come to believe the simplest theory after a run of experience convinces one that no further, complicated effects are forthcoming—e.g., after the marble emitter is covered with cob-webs and forgotten in the basement or after persistent attempts to locate the ether drift have yielded null results. It is less plausible to demand *full* credence in the simplest explanation *immediately* after refutation of simpler possibilities. Such Ockham extremism is avoided below by restricting attention to reductions in credence. But first, I motivate the marble counting setup as a model of inductive inference among theories that differ in simplicity.

7 Marbles, Effects, and Empirical Simplicity

The force of the preceding discussion depends on the aptness of marble counting as a model of science, so I digress in this section to explain why the model captures the fundamental features of empirical simplicity.[17] Kyburg (1961), like Harold Jeffreys (1985) and many others, viewed simplicity as a matter of minimizing free parameters. Each free parameter corresponds to an empirical *effect*, which functions, evidentially, like the marbles in our trivial example. For example, for $i = 1, 2, 3$, let T_i denote the theory that y is a polynomial function of x of degree i but not of degree $i - 1$:

$$\begin{array}{lccl}(\exists\theta_1)\ldots(\exists\theta_{i-1})(\exists\theta_i)(\forall x)(\forall y) & y & = & \theta_1 x^1 + \ldots + \theta_{i-1}x^{i-1} + \theta_i x^i; \\ (\forall\theta_1)\ldots(\forall\theta_{i-1})(\exists x)(\exists y) & y & \neq & \theta_1 x^1 + \ldots + \theta_{i-1}x^{i-1}.\end{array}$$

Real-valued data are not exact, but become ever more accurate as sample size increases. By way of idealization, suppose that the scientist can specify an arbitrary, rational value of x any number of times and would eventually receive arbitrarily small open intervals around x if x were to be queried infinitely often. The scientist can query an infinitely repetitive enumeration of rational values of x to guarantee ever more precise open intervals around x. Suppose the truth is T_2. Then the scientist will eventually receive three open intervals through which no line fits. Call that a second-order effect. Suppose the truth is T_3. Then the scientist will eventually receive four intervals through which no parabola fits.

[16] Special creationists resisted Darwin's common ancestry explanation of structural homologies across species due, in part, to the observed resistance of species to change under artificial selection. Newton rejected the elegant wave theory of light in favor of his complicated particle theory due to a faulty hunch that such a theory could never explain geometrical shadows. In light of such examples, Thomas Kuhn (1962) concluded that simplicity is merely a value *sui generis* that can be offset by others, which opens inquiry to the objection of being an extended exercise in wishful thinking (van Fraassen 1981).

[17] Extended motivational discussions may be found in (Kelly 2007, 2008).

Call that a third-order effect. Empirical effects are like marbles—they can be arbitrarily subtle and, therefore, may appear arbitrarily late. Furthermore, each theory (in this case, polynomial degree) corresponds to some finite set of effects each of which appears eventually.

Here is another example, drawn from the literature on inferring causal relations from non-experimental data (Spirtes et al. 2000). Assume that if there is a causal connection between two variables such as "smoking" and "lung cancer" if and only if smoking is correlated with lung cancer and the correlation is not broken by conditionalizing on any other variable.[18] Then each causal connection corresponds to a set of conditional correlations. Idealizing again, one may think of these conditional correlations as being arbitrarily small or subtle and, therefore, as taking an arbitrarily long time to notice, so the set of conditional correlations corresponding to a direct causal connection may be viewed as a marble that might appear at any time. It is usually thought that more causes make for a more complex theory than fewer causes, so again, more complex theories imply more effects.

The problem of multiple simplest theories also arises naturally in causal inference. Now, consider the orientation (direction) of direct causal connections rather than just their existence. A *linear triple* of variables is a triple of variables with immediate causal connections arranged as: $X - Y - Z$. In this case, Y is the *middle* of the linear triple. A linear triple admits of four possible causal orientations:

$$\begin{gathered} X \rightarrow Y \rightarrow Z, \\ X \leftarrow Y \leftarrow Z, \\ X \leftarrow Y \rightarrow Z, \\ \\ X \rightarrow Y \leftarrow Z. \end{gathered}$$

In the first three arrangements, X is correlated with Z, but only in virtue of Y, so conditional on Y the correlation between X and X disappears. In the last arrangement (a causal *collision*), X is not correlated with Y (they are independent causes) but since X and Z conspire in producing Y, it follows that X is correlated with Z given Y. The problem is to converge to as much truth as possible about causal structure. Suppose that you previously had no sign of causal connections between X, Y, Z and that you suddenly learn that $X - Y - Z$ holds. It is now *destined* that a new effect deciding between the collision and the other cases will be seen. Ockham's razor (as well as some standard software packages for learning causal networks from correlational data) demands that one wait until nature decides the matter.[19]

[18] That is a consequence of the so-called "causal Markov" and "faithfulness" assumptions (Spirtes et al. 2000).

[19] Strictly speaking, nature does not provide either open intervals around predicted quantities or a determinate verdict on correlations. In real applications, the open intervals are replaced with confidence intervals and correlations are concluded when a statistical test of zero correlation rejects and convergence to the truth is understood to be convergence in prob-

So marbles and empirical effects are analogous, but what, exactly, are empirical effects? It is tempting, from the examples, to say that an empirical effect occurs when all theories of a lower complexity are refuted—but that is circular if empirical complexity is also defined in terms of implied effects. The circle can be broken as follows. Let K denote a set of infinite input sequences, which I will refer to as possible empirical worlds. Let Π be a countable partition of K into empirical propositions. The pair (K, Π) is an *empirical question* with *presupposition* K and *potential answers* Π. Now, say that nature can *force* a sequence $(T_1, \ldots, T_n)$ of possible answers drawn from Π if and only if nature has a strategy for presenting data to an arbitrary, convergent scientific method such that $(T_1, \ldots, T_n)$ is a sub-sequence of the sequence of theories the method outputs through time in response to nature's inputs. In the curve-fitting question, nature can present ever smaller open intervals around a line until the convergent method says "linear". Since the intervals are all open and but finitely many of them have been presented, nature can add on a quadratic term with a sufficiently small coefficient to pass through each such interval and can continue to present ever smaller open intervals around this parabola until the convergent method says "quadratic". Then nature can add on a cubic term with sufficiently small coefficient to fit through all the open intervals presented so far until the convergent scientist says "cubic". Hence, ("linear", "quadratic", "cubic") is forcible from science. The order of this sequence corresponds to intuitions of simplicity (e.g., to the number of free parameters in the respective theories). Moreover, no non-trivial permutation of this sequence is forcible by nature, for Ockham's razor converges to the truth without ever producing such a sequence.

I propose that empirical simplicity is most fundamentally a reflection of the forcibility order and that the usual marks of empirical complexity (extra entities, extra causes, extra verbiage, extra free parameters, extra independent principles, less testability) are all reflections of this underlying concept.[20] Since convergence to the truth and forcibility depend on the question the method is to converge to the truth about, the proposed concept of empirical simplicity depend essentially on the semantics of that question. It does not, however, depend on notation or on how the question is asked. In that sense, the idea is more fundamental than any syntactic approach like counting free parameters, but will agree with such accounts when parameters are materially linked to the presentation of data in the usual way. For example, when the problem is to infer conservation laws explaining reactions in particle physics (Schulte 2000), each conserved quantity makes the theory *simpler* (more testable, symmetric, etc.) but each such quantity introduces new free parameters (the amount of the conserved quantity carried by each particle type). For a simpler example, suppose that instead of a box emitting marbles, we are shown a box with a false floor that reveals increasingly more empty space at irregular intervals until it comes to rest on top of the marbles. Now Ockham's razor says, properly, that

ability. Ockham efficiency has not yet been established for statistical inference, but it has been established for random empirical methods conceived very generally (Kelly and Mayo-Wilson 2008).

[20] For more explicit and general definitions of empirical complexity, cf. (Kelly 2007, 2008).

one should infer as *many* marbles as are compatible with current experience.

8 The Thrill of Incoherence and the Agony of Retreat

The Ockham efficiency argument presented above assumes that the relevant measure of cognitive cost of convergence is cognitive distance traversed prior to convergence, which weights increases and decreases in credence equally. But there are other measures of epistemic cost. For example, C. S. Peirce (1878) saw a crucial asymmetry between the fixation of belief, which relieves one from the pain of doubt, and retracting or losing a belief, which re-introduces the pain of doubt.[21] That suggests focusing not on total distance traveled to the truth, which folds together both increases and decreases in credence, but just on loss of credence along the way.

Following classical Bayesian intuitions, one might define the retraction from $\mathbf{b}$ to $\mathbf{b}'$ as total distance traveled toward the state of ignoredge $\mathbf{d}$ at the center of the triangle:

$$r'(\mathbf{b}, \mathbf{b}') = \rho(\mathbf{b}, \mathbf{d}) \dot{-} \rho(\mathbf{b}', \mathbf{d}),$$

where *cutoff subtraction* is defined by:

$$x \dot{-} x' = \max(0, x - x').$$

Unfortunately, all retraction efficient methods must still leap immediately to the next simplest hypothesis as soon as the currently simplest hypothesis is refuted. To see why, consider figure 9. It follows immediately from the definition of r' that:

$$\begin{aligned} r'(\mathbf{d}, \mathbf{i}_1) &= r'(\mathbf{d}, \mathbf{c}) = r'(\mathbf{b}, \mathbf{i}_2) = r'(\mathbf{c}, \mathbf{i}_2) = r'(\mathbf{c}, \mathbf{i}_3) = 0; \\ r'(\mathbf{i}_1, \mathbf{d}) &= \sqrt{2/3}; \\ r'(\mathbf{i}_1, \mathbf{b}) &= r'(\mathbf{i}_2, \mathbf{c}) = \sqrt{2/3} - \sqrt{1/6}. \end{aligned}$$

Suppose that T_1, T_2, T_3 are ordered by increasing complexity, so every convergent method can be forced by nature to move from $\mathbf{i}_1$ to $\mathbf{i}_2$ to $\mathbf{i}_3$. The extreme Ockham strategy that moves straight from one vertex of the triangle to the next incurs the retractions along path:

$$(\mathbf{i}_1, \mathbf{i}_2, \mathbf{i}_3),$$

which adds up to $2(\sqrt{2/3} - \sqrt{1/6})$ retractions. A more plausibly moderate Bayesian would traverse the path:

$$(\mathbf{d}, \mathbf{i}_1, \mathbf{c}, \mathbf{i}_2, \mathbf{i}_3),$$

which incurs $\sqrt{1/6}$ *more* retractions. It is pretty clear from considerations of continuity that any deviation from the extreme Ockham path from $\mathbf{i}_1$ to $\mathbf{i}_2$ would

[21] The idea is that ignorance is only "bliss" when it is masked by false belief. That admittedly contradicts the ancient skeptics, who viewed belief as the source of risk and anxiety.

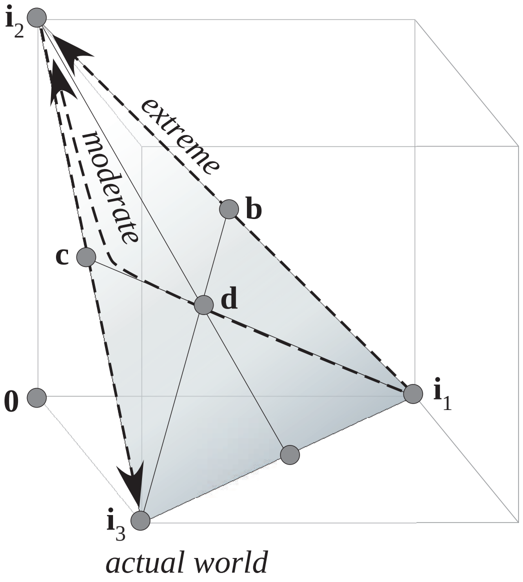

Figure 9: Retreat to ignoredge

result in some extra retractions. Furthermore, there remains the nagging objection that ignoredge, unlike true ignorance, is not invariant under coarsenings of the partition defining the empirical question to be answered. For example, suppose that the unseen marble might have three different colors. Any method that is optimal with respect to ignoredge in that problem will fail to be optimal in the coarsened problem that disjoins two of the colors (the usual paradox of indifference). On the proposed definition of retraction efficiency, however, the sub-additive Ockham method in figure 11 is optimal both in the original problem and in the coarsened problem, since **0** represents ignorance in both cases.

Those objections are avoided if retractions are measured as overall drops in credence, rather than as overall motion toward ignoredge. When computing Euclidean distance between two belief states $\mathbf{b} = (x, y, z)$ and $\mathbf{b}' = (x', y', z')$, one computes:

$$\begin{aligned} \rho(\mathbf{b}, \mathbf{b}') &= \|\mathbf{b} - \mathbf{b}'\| \\ &= \sqrt{(x - x')^2 + (y - y')^2 + (z - z')^2}. \end{aligned}$$

One can define the total *retractions*[22] $r(\mathbf{b}, \mathbf{b}')$ incurred by the move from belief

[22] Note that r is not a metric because symmetry fails:

$$r(\mathbf{0}, \mathbf{i}_1) = 0 < r(\mathbf{i}_1, \mathbf{0})$$

and the triangle inequality fails:

$$r(\mathbf{d}, \mathbf{i}_1) < r(\mathbf{d}, \mathbf{0}) + r(\mathbf{0}, \mathbf{i}_1).$$

state $\mathbf{b}$ to belief state $\mathbf{b}'$ by replacing subtraction $x - x'$ in the definition of ρ with cutoff subtraction:

$$x \dot{-} x' = \max(0, x - x'),$$

so that:

$$\begin{aligned} r(\mathbf{b}, \mathbf{b}') &= \|\mathbf{b} \dot{-} \mathbf{b}'\| \\ &= \sqrt{(x \dot{-} x')^2 + (y \dot{-} y')^2 + (z \dot{-} z')^2}. \end{aligned}$$

It follows, for each belief state $\mathbf{b}$, for each $\mathbf{b}'$ orthogonal to $\mathbf{b}$, and for each coherent belief state $\mathbf{p}$ distinct from the state of ignoredge $\mathbf{d}$ that:[23]

$$\begin{aligned} r(\mathbf{0}, \mathbf{b}) &= 0; \\ r(\mathbf{d}, \mathbf{p}) &> 0; \\ r(\mathbf{b}, \mathbf{b}') &= r(\mathbf{b}, \mathbf{0}) + r(\mathbf{0}, \mathbf{b}). \end{aligned}$$

Recall the example in which at most two marbles might be seen. When the cost of inquiry is measured in terms of retractions, one obtains the tidy result that in complexity class C_n, the best achievable cumulative retraction bound is exactly $n-1$ and that the Ockham method achieves this bound.[24] Again, synchronically coherent enquirers can come arbitrarily close to the best achievable bound.

The corresponding uniqueness result is plausibly weakened, however, for it is no longer necessary to leap to the uniquely simplest theory immediately (figure 10. To minimize retractions, an efficient method can start at the state of true ignorance $\mathbf{0}$ (by the first property). Thereafter, each time the currently simplest theory T_i is refuted, it costs no more to retreat all the way to $\mathbf{0}$ prior to advancing to the next full belief profile $\mathbf{i}_{i+1}$ than it does to advance to $\mathbf{i}_{i+1}$ directly (by the third property). It is not efficient, however, to retreat to the state of ignoredge $\mathbf{d}$ (by the second property). That is a new, truth-directed justification for distinguishing ignorance from ignoredge.

Moreover, recall the example in which one marble of unknown color will be seen. By properties already enumerated, the method that retreats to a state of ignorance when $\mathbf{i}_1$ and that waits for nature to determine which of i_2 or i_3 to move to next converges to the truth with 0 retractions if the world has simplicity 0 and with 1 retraction if the world has simplicity 1, which is optimal (figure 11). This method is also logically consistent, in the sense that it never puts

[23] The last statement is a probabilistic version of the *Levi identity* for belief revision (Levi 1980, Gärdenfors 1988). It says that each change in credence from $\mathbf{b}$ to $\mathbf{b}'$ can be represented, for purposes of calculating total retractions, as a pure retraction first from $\mathbf{b}$ to ignorance $\mathbf{0}$ followed by a pure expansion from $\mathbf{0}$ to $\mathbf{b}'$, since movement from $\mathbf{0}$ incurs no retractions according to the first statement. Carrying that idea further, define the total *expansion* or rise in credence $e(\mathbf{b}, \mathbf{b}')$ to be $\|\mathbf{b}' \dot{-} \mathbf{b}\|$. Then:

$$\begin{aligned} \rho(\mathbf{b}, \mathbf{b}')^2 &= r(\mathbf{b}, \mathbf{b}')^2 + e(\mathbf{b}, \mathbf{b}')^2 \\ &= r(\mathbf{b}, \mathbf{0})^2 + e(\mathbf{0}, \mathbf{b}')^2. \end{aligned}$$

[24] This result agrees with the qualitative retraction bounds established in (Kelly 2002, 2004, 2007a, 2007, 2008) for methods that output theories rather than degrees of belief.

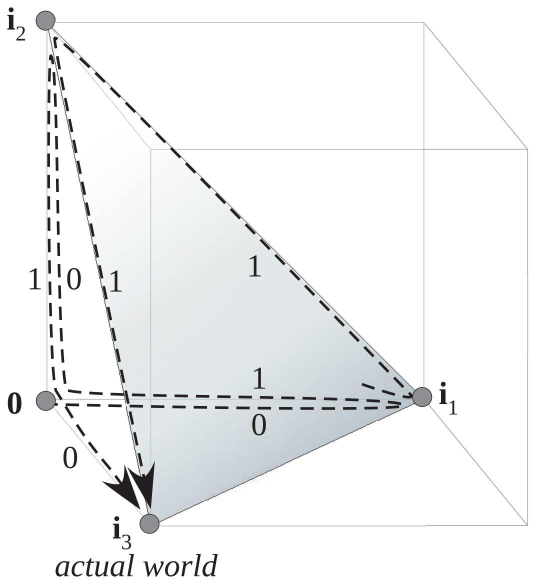

Figure 10: Ignorance as safe haven

credence in a refuted theory. But no consistent, synchronically coherent agent is efficient because retractions are incurred by any motion within the plane of the triangle of coherence. In fact, the best retraction bound a coherent agent can achieve is $3/2 > 1$. So minimization of retraction distance *requires* retreat to a genuine (sub-additive) state of ignorance rather than to a Bayesian state of ignoredge. That is a new, normative, purely truth-directed motive for the intuitions underlying the Ellsberg paradox.

So far, I have merely stipulated consistency with the data, but it is better to explain consistency in terms of retraction efficiency as well. Suppose that, in addition to minimizing retractions, one wishes to get one's retractions over with as soon as possible. That minimizes the number of subsidiary conclusions and technological applications that must be retracted along with the theory. More fundamentally, even a true belief that is destined to be retracted in light of true information fails to constitute knowledge (Gettier 1963), so the search for *knowledge* as opposed to merely justified, true belief argues for getting the retractions over as soon as possible. Now, suppose that you were to remain at $\mathbf{i}_1$ after T_1 is refuted. Had you moved immediately to $\mathbf{0}$, you would have incurred 1 retraction right away. But by remaining at $\mathbf{i}_1$, you are destined to retract later and any path either to $\mathbf{i}_2$ or to $\mathbf{i}_3$ incurs at least one retraction. So in terms of the joint (Pareto) ranking of retractions and retraction times, you lose on the time dimension and the method that departs right away does just as well on the overall retraction dimension.[25] It is left as an easy exercise to verify that the

[25] For the fully general formal details of this argument, cf. (Kelly 2007, 2008).

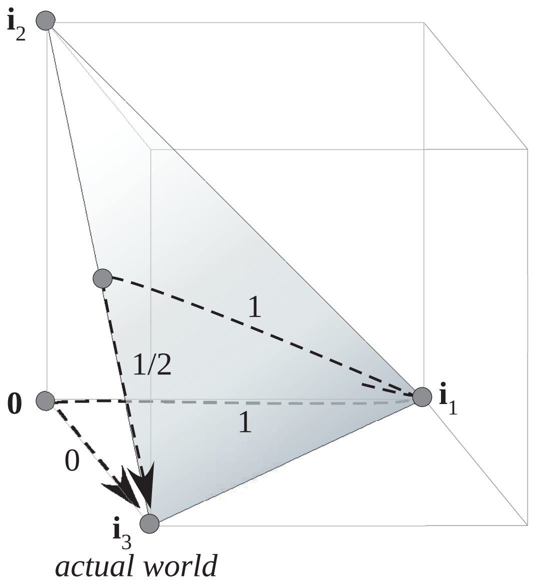

Figure 11: Victory for ignorance

same Pareto-dominance argument for consistency fails for Euclidean distance (cf. figure 8).

9 Inexact Probability

It has just been shown that if ignorance is understood sub-additively as $\mathbf{0}$, then retraction-efficiency implies incoherence. But incoherent belief profiles are dominated in terms of distance to the truth, as discussed above. In this section, it is argued that representing ignorance as inexact probability allows one both to be retraction efficient and to avoid short-run dominance in terms of distance to the truth.

Inexact probabilities have been recommended for a number of reasons, by a number of authors, including Walley (1991), Kyburg (1977, 1983), Levi (1977, 1980), Suppes (1974), and Seidenfeld et al. (1999). Kyburg's *epistemic probabilities* are intervals corresponding to known bounds on chance. Kyburg proposed a rule of direct inference for associating epistemic probabilities with propositional knowledge about chance, so the entire setup is based on objective frequencies. The *imprecise probabilities* of Walley and Levi, are convex sets of personal probabilities and are updated by Bayesian conditioning. Seidenfeld et al. drop the convexity requirement.

A method M with set-valued degrees of belief *converges uniformly* to the true answer to question (K, Π) just in case:

for each world w in K and for each $\epsilon > 0$ there exists stage of inquiry n

such that for all stages $m \geq n$, if B is the output of M at m in w, then:

$$\sup_{b \in B} \rho(b, \mathbf{i}_w) < \epsilon;$$

where $\mathbf{i}_w$ is the probability measure that assigns unit probability to the theory T in Π that contains w.

Next, it is necessary to define the retraction $r(B, B')$ that occurs when B is replaced with B'. Taking the problem in parts, it is natural to define:

$$r(B, \{\mathbf{b}'\}) = \inf_{\mathbf{b} \in B} r(\mathbf{b}, \mathbf{b}').$$

This corresponds to the case with distance: one's distance from Missouri is one's distance to the nearest point in Missouri. Then define:

$$r(B, B') = \sup_{\mathbf{b}' \in B'} r(B, \{\mathbf{b}'\}).$$

The idea here is that if B' sends a protrusion out from B, the total retractions of B by B' are the total retractions of the most distant point in the protrusion (figure 12). This definition has an important and plausible consequence that is crucial to our argument: tightening one's imprecise probabilities does not count as a retraction of prior belief. Let $\mathrm{cl}(B)$ denote the topological closure of B with respect to the usual Euclidean metric topology induced by open balls. Then:

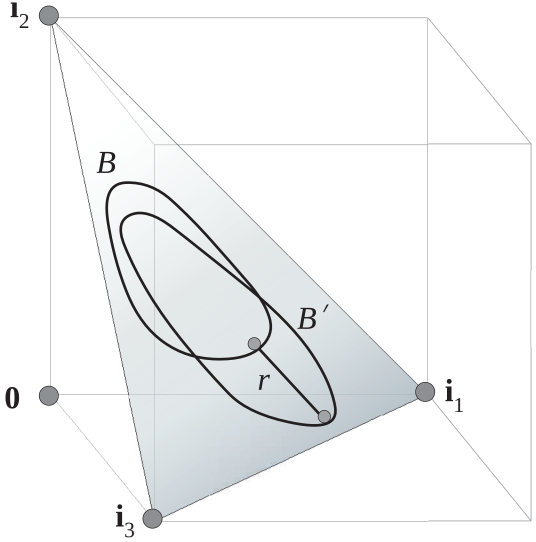

Figure 12: Retraction measure for imprecise probability

$$B' \subseteq \mathrm{cl}(B) \text{ implies } r(B, B') = 0.$$

Consider how sets of probability measures could be updated so as to implement Ockham's razor in the case of the one marble problem in which one must determine whether there will be a marble and, if so, what its color is. Initialize the method with a small, open triangle B of probability measures whose boundary coincides with the boundary of the triangle of coherent measures and includes the vertex $\mathbf{i}_1$, as depicted in figure 13.[26] Assuming that the measures in B all determine the same likelihood for data given T_0, iterated conditionalization of these measures will shrink B toward $\mathbf{i}_1$ as marble-free experience accumulates. Since each successive triangle is a subset of its predecessor, no retractions are incurred during this phase of inquiry (by the property mentioned above). Next, suppose that nature presents a marble. Now, the method outputs the set of all points on the line connecting $\mathbf{i}_2$ with $\mathbf{i}_3$, except for the endpoints. That amounts to a total retraction of 1. When the color of the marble is revealed, the method leaps to the singleton containing the corresponding endpoint. Since the endpoint is contained in the closure of the line, the above property again entails that no retractions are incurred. So the method achieves the best possible bound of one retraction for arbitrary, convergent methods. Since no method that produces a

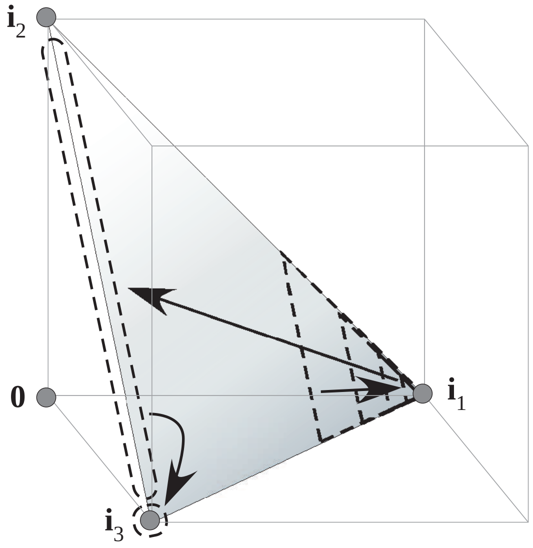

Figure 13: Iterated update of imprecise probabilities

[26] The point $\mathbf{i}_1$ is a member of the boundary of S, not S itself. In light of the dilation results of (Seidenfeld and Wasserman 1993) one might worry whether it is possible to implement uniform convergence to a corner of the triangle. However, those results concern the ubiquitous ability to find *some* event that dilates, not the ubiquity of dilation of a fixed event. In the marble case, suppose that all measures in the triangle have identical versions of the likelihoods $b(e|T_i)$ for $i \in 1, 2, 3$ and differ only in the prior probabilities assigned to the respective theories (that is implicit in the diagram, since differences in likelihood are not even represented). Every (countably additive) convex subset B of the interior of the triangle will converge uniformly to T_1 on increasing, marble-free data.

precise probability distribution at each stage can match that bound, we have a new, normative, objective, truth-directed argument for imprecise probability.

What of Ockham's razor? It is still enforced, but in a plausibly weakened form. Full efficiency demands that measures arbitrarily close to the simplest theory be included in the initial set of measures, but each time the simplest theory compatible with experience is refuted, it is admissible to skeptically adopt the set of all remaining coherent distributions. Further simple experience contracts the set around full belief in the simplest theory compatible with experience.

10 Dilation

Imprecise probabilities can become even less precise in light of new information, which "goes against our seeming intuition that when we condition on new evidence, upper and lower probabilities should shrink toward each other" (Seidenfeld and Wasserman 1993).[27] Kyburg's epistemic probability theory (1977) embodied a stipulation against dilation.[28] The preceding discussion explains, however, why dilation is both natural *and* necessary for optimal truth conduciveness. For after lengthy exposure to marble-free data, the interval between upper and lower probabilities for both T_2 and T_3 collapses on 0, but when the marble is heard but not yet seen, the intervals both for T_2 and T_3 leap to $(0, 1)$. That is as it must be, for any constraint on the interval for either answer would open the agent to extra retractions.

11 Conclusion

This paper sketches some new connections between some themes urged by Kyburg, including objectivity of scientific inference, simplicity, and imprecise probability. It has been argued that a standard, Bayesian explanation of Ockham's razor in terms of Bayes factors runs afoul of the usual paradoxes of indifference and Ellsberg. An alternative vindication of Ockham's razor in terms of minimal path length to the truth was given, but it has the implausible consequence that efficient methods must leap to the simplest theory immediately. Focusing on retractions of credence, rather than expansions of credence, allows for a more plausible Ockham efficiency theorem, according to which suspension of belief is possible prior to advancing to the next simplest hypothesis. However, that result requires, as does Ellsberg's paradox, that ignorance be represented more deeply than as Bayesian indifference. Both incoherence (sub-additivity) and imprecise probabilities have been entertained as models of ignorance in response to Ellsberg's paradox and both allow one to prove a natural Ockham efficiency

[27] Elsewhere, Seidenfeld and Wasserman adopt the much more circumscribed view that dilation is only odd when it occurs "no matter what". The following explanation does not motivate that extreme case.

[28] The rule for associating intervals of probability with information states selects the narrowest interval associated with information entailed by the information available (cf. Seidenfeld 2007).

theorem, but imprecise probabilities have the further advantage of avoiding dominance in terms of distance to the truth in the short run. Furthermore, in order to achieve retraction efficiency, imprecise probabilities must exhibit a maximum degree of "dilation" or increase in imprecision in light of new information. Thus, one arrives at a new, unified, normative explanation in terms truth conduciveness of the intuitions underlying Ockham's razor, Ellsberg's paradox, and dilation of imprecise probabilities.

More generally, I hope this study focuses more attention on truth conduciveness as a source of normative methodological explanations. It is regrettable that some Bayesian epistemologists now *define* truth conduciveness as whatever raises Bayesian credence (e.g., Shogenji 1999, Olsson 2005, Bovens and Hartmann 2004), an idea dispelled by a quick glance at the Bayesian grand tour depicted in figure 6. Real truth conduciveness is just what it sounds like: finding the truth more directly than alternative strategies. Taking truth conduciveness seriously is the key, at least, to obtaining a unified, non-circular, truth-directed, normative explanation of Ockham's razor, the Ellsberg Pardox, and the dilation phenomenon.

12 Acknowledgements

This work was supported generously by NSF grant 0750681. I am indebted to the editors of the volume, Horacio Arló-Costa and Greg Wheeler, for detailed, useful comments. The manuscript also benefited from comments by Conor Mayo-Wilson and from Claus Beisbart, Leah Henderson, Hanti Lin, and David Infortunio, who attended my 2008 seminar on simplicity.

13 Bibliography

Arló-Costa, H. and Helzner, J. (2880) "Ambiguity Aversion: The Explanatory Power of Indeterminate Probabilities", unpublished manuscript.

Bovens, L. and Hartmann, S. (2004) *Bayesian Epistemology*, New York: Oxford University Press.

DeFinetti (1972) *Probability, Induction, and Statistics*, Chichester: Wiley.

Ellsberg, D. (1961) "Risk, Ambiguity, and the Savage Axioms", *Quarterly Journal of Economics* 75: 643-669.

Fox, C. and Tversky, A. (1995) "Ambiguity Aversion and Comparative Ignorance", *The Quarterly Journal of Economics*, 110: 585-603.

Gärdenfors, P. (1988) *Knowledge in Flux*, Cambridge: M.I.T. Press.

Gettier, E. (1963) "Is Justified True Belief Knowledge?", *Analysis* 23: 121-123.

Gilboa, I. and Schmeidler, D. (1989) "Maxmin Expected Utility with Non-unique Prior", *Journal of Mathematical Economics* 18: 141-153.

Jeffreys, H. (1985) *Theory of Probability*, Third edition, Oxford: Clarendon Press.

Joyce, J. (1998) "A Nonpragmatic Vindication of Probabilism", *Philosophy of Science*, 65: 575-603.

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

Kass, R. and Raftery, A. (1995) "Bayes Factors", *Journal of the American Statistical Association* 90: 773-795.

Kelly, K. (2002)"Efficient Convergence Implies Ockham's Razor," *Proceedings of the 2002 International Workshop on Computational Models of Scientific Reasoning and Applications*, Las Vegas, USA, June 24-27.

Kelly, K. (2004) "Justification as Truth-finding Efficiency: How Ockham's Razor Works," *Minds and Machines* 14: 485-505.

Kelly, K. (2007a) "Ockham's Razor, Empirical Complexity, and Truth-finding Efficiency," *Theoretical Computer Science* 317: 227-249.

Kelly, K. (2007) "How Simplicity Helps You Find the Truth Without Pointing at it", V. Harazinov, M. Friend, and N. Goethe, eds. *Philosophy of Mathematics and Induction*, Dordrecht: Springer.

Kelly, K. (2008) "Ockham's Razor, Truth, and Information", in *Handbook on the Philosophy of Information*, Van Benthem, J. Adriaans, P. eds. Dordrecht: Elsevier, pp. ***-***.

Kelly, K. and Glymour, C. (2004) "Why Probability Does Not Capture the Logic of Scientific Justification", forthcoming, C. Hitchcock, ed., *Contemporary Debates in the Philosophy of Science*, Oxford: Blackwell, 2004 pp. 94-114.

Kelly, K. and Mayo-Wilson, C. (2008) "Ockham Efficiency Theorem for Empirical Methods Conceived as Empirically-Driven, Countable-State Stochastic Processes", unpublished manuscript.

Kuhn, T. (1962) *The Structure of Scientific Revolutions*, Chicago: University of Chicago Press.

Kyburg, H. (1961) "A Modest Proposal Concerning Simplicity", *The Philosophical Review* 70: 390-395.

Kyburg, H. "Randomness and the Right Reference Class", *The Journal of Philosophy* 74: 501-521.

Kyburg, H. (1983) *Epistemology and Inference*, Minneapolis: University of Minnesota Press.

Levi, I. (1974) “On Indeterminate Probabilities”, *Journal of Philosophy* 71: 397-418.

Levi, I. (1977) “Direct Inference”, *The Journal of Philosophy* 74: 5-29.

Levi, I. (1980) *The Enterprise of Knowledge*, Cambridge: M.I.T. Press.

Maher, P. “Joyce’s Argument for Probabilism”, *Philosophy of Science* 69: 73-81.

Olsson, E. (2005) *Against Coherence: Truth, Probability, and Justification.* New York: Oxford University Press.

Peirce, C. (1878) “How to Make Our Ideas Clear”, *Popular Science Monthly* 12: 286-302.

Rissanen, J. (1983) “A universal prior for integers and estimation by inimum description length,” *The Annals of Statistics*, 11: 416-431.

Rosenkrantz, R. (1981) *Foundations and Applications of Inductive Probability*, Atascadero, CA: Ridgeview Press.

Schmeidler, D. (1989) “Probability and Expected Utility without Additivity”, *Econometrica* 57: 571-587.

Schulte, O. (1999) “Means-Ends Epistemology”, *The British Journal for the Philosophy of Science*, 50: 1-31.

Schulte, O. (2000) “Inferring Conservation Principles in Particle Physics: A Case Study in the Problem of Induction”, *The British Journal for the Philosophy of Science* , 51: 771-806.

Spirtes, P., Glymour, C.N., and R. Scheines (2000) *Causation, Prediction, and Search.* Cambridge: M.I.T. Press.

Shafer, G. (1976) *A Theory of Evidence*, Princeton: Princeton University Press.

Seidenfeld, T. (2007) “Forbidden Fruit: When Epistemological Probability may *not* take a bite of the Bayesian apple”, in *Probability and Inference: Essays in honour of Henry E. Kyburg jr.*, Harper, W. and Wheeler, G., Kings college publications, London.

Seidenfeld, T., Schervish, M., and Kadane, J. (1999) “Decisions Without Ordering”, in *Rethinking the Foundations of Statistics*, J. Kadane, M. Schervish, and T. Seidenfeld, eds., Cambridge: Cambridge University Press.

Seidenfeld, T. and Wasserman, L. (1993) “Dilation for Sets of Probabilities”, *The Annals of Statistics*, 21: 1139-1154.

Shogenji, T. (1999) “Is Coherence Truth Conducive?”, *Synthese* 157: 361-372.

Suppes, P. (1974) "The Measurement of Belief", *Journal of the Royal Statistical Society*, 36: 160-175.

van Fraassen, B. (1981) *The Scientific Image*, Oxford: Clarendon Press.

Walley, P. (1991) *Statistical Reasoning with Imprecise Probabilities*, London: Chapman and Hall.