Learning, Simplicity, Truth, and Misinformation

Kevin T. Kelly
Department of Philosophy
Carnegie Mellon University
kk3n@andrew.cmu.edu

March 5, 2005

Abstract

Both in learning and in natural science, one faces the problem of selecting among a range of theories, all of which are compatible with the available evidence. The traditional response to this problem has been to select the *simplest* such theory on the basis of "Ockham's Razor". But how can a fixed bias toward simplicity help us find possibly complex truths? I survey the current, textbook answers to this question and find them all to be wishful, circular, or irrelevant. Then I present a new approach based on minimizing the number of reversals of opinion prior to convergence to the truth. According to this alternative approach, Ockham's razor is a good idea when it seems to be (e.g., in selecting among parametrized models) and is not a good idea when it feels dubious (e.g., in the inference of arbitrary computable functions). Hence, the proposed vindication of Ockham's razor can be used to separate vindicated applications of Ockham's razor from spurious ones.

0.1 Introduction

In science and learning, one must eventually face up to the problem of choosing among several or even infinitely many theories compatible with all available information. How ought one to choose? The traditional answer is to choose the "simplest" and to invoke "Ockham's razor". Ockham's actual advice concerns minimizing entities in one's fundamental ontology and was intended for theological and metaphysical debate. In our day, Ockham's razor is understood in a much broader and more scientifically pertinent sense, according to which "simplicity" concerns not only entities, but free parameters, causes, independent principles, *ad hoc* hypotheses, unity, uniformity, symmetry, testability, and explanatory power. I understand the principle in this contemporary, broadly scientific sense.

Hijacking a medieval name for one's prejudices does not justify them. And Ockham's razor raises completely natural concerns upon the most casual, untutored reflection. For how could a *fixed bias* toward simplicity *possibly* be a good inductive method for finding the truth? Surely not in the sense that simplicity is guaranteed to *indicate* or point at the truth. For an indicator has to be sensitive to what it indicates. But Ockham's razor always points at "simple" no matter how complex the truth is. So if Ockham's razor *did* somehow give us an edge in locating the truth, it would be as spooky as getting information about temperature from a broken thermometer with no mercury in it.

Put this way, Ockham's razor starts to sound, well, *occult*— like Ouija boards, creation *ex nihilo*, the Philosopher's Stone, and perpetual motion machines. It seems, for all the world, to provide a free *epistemological* lunch— an unexplained, bottomless font of information that springs from no apparent source. It is small wonder that this cost-free amplifier of available information is so beloved in philosophy, science, statistics, and machine learning.

In spite of this rather severe introduction, I come not to bury Ockham, but to praise him, by explaining how, in a definite sense, Ockham's razor really is the most efficient possible means for arriving at the true theory. But first, I review some more traditional attempts at such an explanation to illustrate why an entirely new approach is necessary.

0.2 The Standard Apologies

It obviously isn't easy to explain how assuming what we don't know can help us find the true theory, so it is hardly surprising that the usual apologies are unsatsifactory. For convenience, I divide them into three categories: the *wishful*, the *circular*, and the *irrelevant*.

0.2.1 Wishful Thinking

Simple theories have attractive aesthetic and methodological properties. Aesthetically, they are more unified, uniform and symmetrical and are less *ad hoc* or messy. Method-

ologically, they are more severely testable, explain better, and predict better. Many philosophers of science have remarked on the desirability of some of these features, have derived one of the features from another, and have concluded that the latter is a reason for seeking the former (e.g., Popper 1968, Glymour 1981, Friedman 1983). It is particularly tempting, for example, to recommend simplicity because simple theories are more testable or to demand explanations and to observe that simple theories explain better. But the truth might not be simple and if the truth isn't simple, then the truth doesn't have the sort of explanatory power, testability, etc. that we wish it to have. So *inferring* that the truth is simple due to the nice properties that would follow if it were is plainly an instance of wishful thinking. It would be far better first to explain how Ockham's razor helps us find the true theory and then to use Ockham's razor to infer that the truth has the desirable properties.

0.2.2 Begging the Question

Bayesian methodology has an easy "explanation" of Ockham's razor: just put high prior probabilities on simple theories and turn the crank on Bayes' theorem (Jeffreys 1985). Then if you are forced (contrary to Bayesian ideology) to choose among theories, choosing simpler theories compatible with the data will look like a better policy to you. Of course, this argument is a bit *too* easy. The trouble with circularity is not that it is irrational or that knowledge must rest on a secure, Cartesian foundation, but that it could equally "explain" *any* prior bias and, hence, explains *none*.

There is a more subtle Bayesian argument for simplicity that does not presuppose a bias toward simple theories— at least on the surface. To keep things very concrete, suppose that we have just two theories, a simple theory S that can be true in just one way and a complex theory C that can be true in any one of k mutually incompatible ways C_1 to C_k. Now we don't want to be *unfair* to S, so we should assign even weight to S and to C *a priori*. Furthermore, we have no idea which way C might be true, so assign uniform probabilties to the cases C_i. Suppose that only S and C_i are compatible with evidence E. Then turning the crank on Bayes' theorem, $P(S|E)/P(C|E) = k$. So even though the complex theory could save the data, the simple theory that did so without any *ad hoc* fiddling ends up being "confirmed" much more sharply by the same data (cf. Rosenkrantz 1983). And who can say that the argument depends on a prior bias toward S? For at the outset, the prior probabilities of S and C are identical. Indeed, it would be a *miracle* if out of all the possible ways of being true, C turned out to be true in precisely the way that matches competitor S. Also, the unified, elegant theory S was *tested* by the data but the clunky theory C had to *use* the data to set its parameters and so didn't withstand a severe test. So S should be *rewarded* for its valor in the field of honor, etc.

Yet, there is the lingering doubt that as far as explanation is concerned, C_i and S are objectively the same and that C_i might very well be true. So what's *objectively* wrong with C_i? Nothing. Theory S beats C_i only because we imposed a lower prior probability on C_i. Two can play that game: since there are so many ways for C to be true, the prior probability of S should be much smaller— say the same as that of

C_i. Then S never gets ahead and the Ockham argument evaporates. The moral is that there is no such thing as Bayesian "fairness", since "fairness" in one partition (e.g., blue *vs* non-blue) implies bias in others (e.g., blue *vs* red *vs* yellow). Hence, the real Bayesian choice is always *which* bias to adopt, and one may as well be "fair" at the level of parameter settings as at the level of theories.

There have been attempts to select some particular prior distribution as "special" and to show that Ockham's razor follows. For example, Rudolf Carnap (1950) imposed prior probabilities favoring uniform possibilities, with uniformity of nature as a not so surprising result.

A more recent and ambitious program for vindicating a special prior probability is developed within *algorithmic information theory* (Li and Vitanyi 1997). The algorithmic complexity of a string corresponds (roughly) to the length of the shortest computer program (in some fixed language) that generates the string. The intuitive idea is that simple strings have structure that a short program can exploit to reproduce it. This gives rise to the notion that good explanations are short theories that compress the data and that Ockham's razor is a matter of minimizing the sum of the lengths of the theory and of the compressed data. The idea that one should infer the best explanation in this sense is called the *minimum description length* principle or MDL for short (Rissannen 1983). Algorithmic information theorists have also developed the notion of a *universal prior* probability over bit strings with the property that more compressible strings tend to have higher prior probability. It can also be shown that under certain conditions the MDL approach is similar to Bayesian updating with the universal prior probability (Vitanyi and Li 2000).

First, there is a degree of internal arbitrariness in algorithmic complexity, for the length of the shortest program that produces a given string depends on how programs are interpreted, and that is pretty arbitrary. One interpreter could interpret bit string 0 as the successor function and the string with a million repeated units as some horror like the Ackermann function, while a second interpreter could just as easily reverse this convention. This makes it hard to take program length seriously as an indicator of how the world must be.

There is a partial response to this objection. The first interpreter, being a universal machine, can be arranged so that when provided with a program, it simulates the other interpreter on this program. Then the length of the shortest program producing string x in the first interpreter can't be too much bigger than the shortest program producing x in the second system, since at worst there is the shortest program in the second system packaged with the fixed overhead of a program of the second interpreter in the first interpreter's language. In other words, only a constant (depending on the length of the program of the second interpreter in the first interpreter's language) separates the complexities assigned by *these two* intepreters. But that is scant comfort when the application of the formalism is Ockham's razor, for it is still the case that an arbitrarily complex theory in the first interpreter could be the best possible theory for a second. Moreover, the constants connecting systems could be arbitrarily large, so no matter how many reversals of simplicity ranking one wishes to effect, one could fish for an alternative machine that effects them. It remains very hard to imagine how

the arbitrary conventions by which a computer chomps on symbols could explain how favoring short theories helps us find the truth about particles, galaxies, societies, and other objective truths that have nothing to do with computer science or coding.

But aside from that, there is the crucial issue, entirely beyond the purview either of logic or of computation, of how the data are processed and encoded into bit strings by the environment or the laboratory prior to their entry into the learning program that employs the special prior probability. Suppose that in the original experiment "green" is encoded by one's experimental instruments as 0 and "blue" by 1 and that these code numbers are fed to the learning agent. Then the "right" prior opinion favors constantly 0 sequences or constantly 1 sequences. Now translate into the evidential language "grue" (Goodman 1983) which means "green up to t and blue thereafter" and "bleen", which means "blue up to t and green thereafer" and code "grue" by one and "bleen" by zero. Then the rule for assigning simplicity-biased opinions *a priori* reverses. Goodman's point in the grue construction was that since logic is preserved under translation, such probability assignments can't be a feature of logic, as Rudolf Carnap would have it. My point is a bit different: no bias that depends upon mere conventions about how the data are passed along to the learner could possibly be an indicator of truths lying behind the data-reporting process that do not depend upon those mere conventions. These environmental contingencies are not mere philosophical quibbles. In machine learning practice, it is widely recognized that a successful MDL learner's coding scheme must be tuned to the particular character of the problem at hand and that the algorithmic complexity setup is more of an idealized theoretical backdrop than a practical guide (Mitchell 1997).

0.2.3 Missing the Point

(Bias *vs.* Variance) There is a familiar and important sense in which simplicity is connected with a kind of short-run truth-finding optimality. Suppose that instead of trying to infer the true theory (i.e., the theory with the true functional form, the true variables, the true free parameters, the true causal relationships, etc.) you wish only to *use* the theory to estimate the sampling distribution from which the data are generated. Success in this task is measured in terms of expected overall "error" or mis-match between the estimated distribution and the true distribution underlying the data (where error be measured in various ways). The expected error of the estimated distribution can be factored into two components, *bias* or "off-centered-ness" of the average estimate and *variance*, or "spread" of the estimate due to sampling variation (e.g., Wasserman 2003). It is a familiar and ubiquitous statistical fact that adding parameters to the model used for estimating a distribution reduces bias at the expense of variance. So to minimize expected error in one's estimate of the sampling distribution, one should choose a model that is neither too simple nor too complex. Erring to either side is called underfitting or overfitting, respectively.

Overfitting is most definitely principled, intuitive, *a priori*, and truth-directed argument for preferring simple theories for the purpose of estimating an underlying sampling distribution. But the story is *not* directed toward finding the truth of the recommended

theory. For even if God were to present you with the true, complex theory generating the data, the true theory handed to you might overfit the data and result in needlessly high variance. In that case, you would prefer an oversimplified theory to the true one you have in hand for the purposes of estimating the underlying sampling distribution. Hence, estimating the underlying distribution is quite different from finding the true theory.

Moreover, the actual sampling distribution is good only for predictive purposes, whereas the true theory (e.g., the true causal pathways or equational form) can determine counterfactual relationships (Spirtes et al. 2000) of considerable scientific and practical interest (e.g., a tiny human foible that advertisers or stock traders can strategically exploit for tremendous gain). Indeed, in typical problems, the match in theoretical truth (measured in terms of numbers of free parameters, causes) can be arbitrarily *bad* when the match in sampling distributions is arbitrarily *good*. In other words, plausible measures of *theoretical* loss are discontinuous with the notion of error of estimated distributions assumed in the overfitting story.

Finally, one must be careful about the kind of connection to the truth model selection methods provide. A familiar technique in model selection is to choose a model that maximizes a score called the *Akaike Information Criterion* (Akaike 1973, Forster and Sober 1994). Let $T[\theta]$ be a theory with n free parameters and let E be a particular sample. Let $\hat{\theta}$ denote the value of θ that maximizes $P(E \mid T[\hat{\theta}])$, in which case we say that $\hat{\theta}$ is the *maximum likelihood estimate* of θ using $T[\theta]$ from sample E. Then the Akaike information criterion is given by

$$AIC(T[\theta] \mid E) = \log(P(E \mid T[\hat{\theta}]) - n.$$

Notice that this formula is almost too good to be true as an explanation for Ockham's razor, as it congratulates a theory for its explanatory value but then taxes its complexity in the most obvious manner imaginable by subtracting off the number of free parameters it employs. Furthermore, it has something to do with *information*, so isn't simplicity informing us which theory to choose?

Here is what AIC does. As it happens (cf. Wasserman 2003), there is a quantity x depending on the estimated sampling distribution and the true distribution such that maximizing x minimizes expected error. AIC is nothing but an (almost) unbiased estimate of x. So *maximizing AIC amounts to maximizing an (almost) unbiased estimate of something that when maximized minimizes expected error of the maximum likelihood estimate of the underlying sampling distribution using the theory so selected*. Got it? Does that mean that you will converge to the true theory? No way. For example, in the easy problem of determining how many components of a bivariate normal mean are zero it doesn't come close to converging to the simplest theory (0, 0) in probability. It starts out very optimistic about (0, 0) and after sample size around 200 it gets skeptical and remains so forever after. Does it mean that you will at least converge to a true estimate of the sampling distribution? Not even that, because AIC is not guaranteed to converge in probability to the true estimate.

Therefore, the overfitting story gives little comfort to those who thought that Ockham's razor had something to do with finding the true theory. Indeed, it arguably em

undermines such hopes by explaining how they might be based on a natural confusion between a theory and the sampling distribution determined by the theory.

(Probably Approximately Confusing) Theoreticians in machine learning have coined and carefully examined a concept of *probably approximately correct* or PAC learning (Kearns and Vazirani 1994). Suppose that all we care about is not being embarrassed by future counterexamples to our inferred classification rule, so we measure "approximate correctness" of a rule by the chance that a single, sampled individual is a counterexample to the rule. Call this probability ϵ. Suppose we want a method that has at least a certain probability $1 - \delta$ of producing such a classification rule. Now we can ask how large the sample has to be for our method to achieve probability $1 - \delta$ of producing a rule that has probability at most ϵ of being refuted by a single draw from the urn of instances. If the sample size doesn't grow too fast (i.e., is polynomial in the reciprocals of ϵ, δ), then the method is said to be PAC.

The PAC program has many interesting and important results to its credit, but I am concerned here only with its surprising argument in favor of Ockham's razor. The theorem (Kearns and Vazirani 1994, Li and Vitanyi 1997) is that producing a nearly-shortest-possible classification rule suffices for PAC learning, where the sense of "nearly" can be expressed as a function of α and β. The suggestion is that data compression is deeply related to finding the truth. What is really amazing about this is that it is based on a kind of short-run success and nonetheless could not be the result of a circular, prior bias toward simplicity since there are no prior probabilities in the story whatever. So somehow, down in the mysterious bowels of objective chance and information, brevity somehow tracks or divines or *informs* us of the truth?

Not by a mile. The theorem says only that data-compression *suffices* for PAC learning. It is also true that drawing a classification rule from a sufficiently small set of classification rules suffices for PAC learning. Indeed, the Ockham result is proved as a corollary of this fact about small sets of answers: just choose the notion of "sufficiently close to the simplest rule compatible with the data" to keep the set of all such rules sufficiently small. By the same argument, it suffices for PAC learning to always choose a rule sufficiently close to a very verbose rule— hardly a resounding victory for simplicity. But how could it be otherwise? In the PAC model, classification rules are assigned inscriptions arbitrarily, so it would truly be a miracle if inscription length were a divining rod for the unseen truth.

(Convergence theorems) In many scientific and learning problems of interest, it is possible to show that various methods, Bayesian or otherwise, that are armed with a prior bias toward simplicity converge, eventually, to the true theory, either necessarily or almost surely or in probability, depending on the application (cf. Wasserman 2003). The story is that the simplicity bias "washes out" in the long run as data accumulate or that eventually the overly simple hypotheses get refuted and eventually the simplest hypothesis compatible with experience is true. That's a fine thing to know, but it falls far short of a *recommendation* for Ockham's razor, for just about any alternative, prior bias will "wash out" in the limit of inquiry in the same way (Salmon 1967). The convergence theorems simply say that a given prior bias doesn't *prevent* one from arriving at the truth eventually. But to say on that basis that a given bias *helps* you

find the truth is plainly an exaggeration— like saying that a flat tire helps you get to work because you can fix it.

Still, the general strategy of looking for senses of "finding the true theory" that don't require an occult connection between simplicity and truth is on the right track. Perhaps there is some refinement or strengthening of the notion of convergence to the truth that avoids occult connections but that still singles out simplicity as the right prior bias for success. That is the approach I will now pursue.[1]

0.3 The Obstacle

A perusal of the standard Ockham literature gives the impression of a subject that has lost the philosophical forest for the mathematical trees. Statistical and computational discussions of Ockham's razor now fill the better parts of fat volumes, replete with formidable-looking formulas and proofs. But in spite of all the minute detail and calculation that takes years to learn, surprisingly little attention is devoted to the most important element of the simplicity puzzle, namely, what it even *means* for a method to help one find the true theory. Maybe if we were clearer about what help in finding something *is*, it would be clearer that Ockham provides the best possible help.

To recapitulate, the two standard notions of finding the true theory are (1) *indication* or *pointing* at the truth in the short run and (2) *convergence* in the long run. The trouble with the pointing metaphor is that you can't explain how simplicity points at the true theory without invoking some question-begging bias toward simplicity at the outset. The trouble with mere convergence is that almost any bias will eventually be swamped by incoming information. In short, pointing is too strong and convergence in the limit is too weak to support the desired explanation. Throwing more mathematics

[1] In fact, Oliver Schulte and I have been working on this idea for some time. The basic idea of counting mind-changes is originally due to H. Putnam (1965). It has been studied extensively in the computational learning literature— for a review cf. (Jain et al. 1999). But in that literature, the focus is on categorizing the complexities of problems rather than on singling out Ockham's razor as an optimal method. I viewed the matter the same way in (Kelly 1996). Most philosophers of science have read W. Salmon's (1967) complaint that convergence results don't constrain scientific behavior in the short run. To address this complaint, Oliver Schulte and I started looking at retraction minimization as a way to severely constrain one's choice of hypothesis in the short run. Schulte's thesis work in this is summarized and extended in (Schulte 1999a, 1999b). Schulte has also applied the idea to the inference of conservation laws in particle physics (Schule 2001). In 2000, I began to extend the idea, based on a variant of the ordinal mind-change account due to (Freivalds and Smith 1993), resulting in a detailed formal derivation of Ockham's razor from efficiency (reported in Kelly 2002) and worked out in detail in an unpublished manuscript. I was not happy, however, because the ordinal retraction theory was very complicated and didn't apply to standard cases of model selection. Just when the manuscript was complete I took my wife on a trip to Niagara falls, got lost and asked for directions, and realized that what had happened yielded both a more general and much simpler analysis of Ockham's razor than the ordinal retraction theory did. The story is described below. Subsequently, I noticed that retraction minimization applies to causal inference and recommends something close to the causal inference procedures in (Spirtes et al. 2000). This led to an initial attempt to extend the idea to causal inference in (Kelly and Glymour 2004) and to an improved presentation in (Kelly 2004). The mathematics behind the following story have not yet been published and are only sketched below.

and sophistication at these failed ideas will simply result in more subtle disguises of the failure, as is apparent in the preceding review of standard arguments.

If the connection between simplicity and finding the true theory is ever to be explained, one must coin a notion of "helping to find the truth" that lies *between* indication and mere convergence in the limit, so that it neither requires circles nor vindicates alternative biases. I will now present just such a concept. Then I will use the concept to frame a simple, straightforward argument for the superiority of Ockham's razor over all competing methods. The argument *won't* imply that the simplest theory compatible with experience *is* true (that requires circles or occult channels), but it will imply that Ockham's razor gets you to the truth more efficiently than any other empirical strategy. What more could one consistently ask?

0.4 A Solution

It helps to pose the question of finding the true theory more broadly: how, in general, could *fixed* advice help you find *anything* unseen? It happens every day! Suppose you are lost in a small town on a road trip. You ask a local resident for directions. She directs you to the highway entrance ramp. You get on the highway, which cuts through a few mountain passes and traverses a few rivers on the route home.

For a more interesting story, suppose you head in a different direction, ignoring the resident's advice. The road narrows, heads through fields, then woods, and finally to rutted switchbacks in the mountains. You finally concede that it wasn't a good idea and retrace your route back to the resident, who waves as you pass. Then you follow her advice to the highway and follow the standard route home. Your reward for ignoring the advice was an unpleasant journey culminated by a humiliating U-turn prior to even beginning your proper journey home. You should have listened.

Consider the following features of the preceding story.

1. The resident's advice is *helpful*, since it minimizes troublesome false-starts on the way home. Indeed, it is helpful even if the directions to the entrance ramp lead *directly away* from the ultimate destination, for even if you happened to aim straight for your house, you would end up travelling over inferior routes.

2. The resident gives precisely the same, *fixed* advice to every city slicker who asks and it is helpful advice for all of them, even though she has no idea where they are ultimately headed. Once they are at the highway entrance they can take care of themselves, whatever their ultimate destination (of course, they may consult other residents for directions at later rest stops). So the resident required no Ouija board to give you the advice, even though she didn't know where you were going.

3. The resident's advice is the *uniquely best* advice even though the ultimate destination is unknown, for violating it leads to an extra U-turn, a dead end, or an inferior route all the way home.

4. The sense in which the resident's advice helps you get home is that it *minimizes the number of annoying course-reversals en route to your goal*. This sense of help is less than indicating or pointing to where your house actually is (that would require that the resident read each traveller's mind prior to giving advice) and more than just getting you home eventually (no matter what she says, you would get home eventually).

So getting directions to the highway entrance ramp satisfies all the apparently arcane demands that a successful explanation of Ockham's razor must satisfy and does so without Ouija boards or other occult connections between the resident and your unknown destination.

0.5 The Highway to the Truth

If there were highways to the truth and these highways had entrance ramps, then Ockham could point us toward them and the mystery would be resolved. But what does all this mean when we move from road trips to model selection in science and machine learning?

Here is a very easy example that still captures much of the spirit of the general idea. Suppose that there is an experiment that emits detectable particles of an unknown nature at irregular intervals. Something guarantees that only finitely many particles will ever be emitted but nothing indicates a time by which this will happen or how many there will be. Your assignment is to determine how many particles will be emitted. Ockham's razor recommends not counting your particles before they hatch. Indeed, Ockham, himself, recommended not assuming more entities than necessary. In a more modern idiom, the time of appearance of each particle posited by a more bloated theory amounts to a free parameter. These times of appearance are unexplained, break symmetry, and introduce non-uniformities in experience, so the intuitive marks of simplicity all speak with one voice. I will show below that this is no accident, as they are all reflections of an underlying, topological concept of empirical simplicity.

Suppose that two particles have been emitted. Time passes. On pain of not converging to the truth ever, you eventually have to cave in and conclude "two". I am *not* saying that lots of instances *confirm* and therefore *compel you to believe* that no more are coming. Rather, you are pulled along by the aim of converging to the truth eventually rather than pushed along by accumulating evidence. When you make the leap is up to you; that you make it eventually is part of what is meant by scientific success.[2]

Suppose another particle appears. Tough— nobody is safe from Hume's problem of induction. Lots of time passes without new particles. Again, on pain of failing to converge to the truth you eventually conclude "three". And so forth. So as a goody

[2]This expresses the fundamental ethos of formal learning theory as encountered, say, in (Jain et al. 1999) or in (Kelly 1996). It is an excellent tonic for what ails standard, "confirmational" approaches to the problem of induction (cf. Kelly and Glymour 2004).

two-shoes Ockhamite, you can be forced by nature to change your mind at most once per particle.

Now suppose that two particles have been emitted and you experience a bout of rebellion. You have been fooled twice already following Ockham's stupid advice. It was horrible to retract two published research reports in a row from the top journal. Given your past experience, you figure you may as well leap to "three" right away to save the ignominy of retracting "two". Time passes. And passes. You start to think about the many null experiments for detecting the ether drift that led to the downfall of Newtonian physics. And the failed attempts to locate evidence of Noah's flood in nineteenth century geology .... If you never retract back to two you won't converge to the truth if no more particles appear .... Arrgh! You cave in and change your mind to "two".

Another particle appears— too bad! From now on, you can be forced to change your mind at least once per particle, which is as bad as Ockham. But you started out with the annoying switch from "three" to "two" that heeding Ockham's advice would have avoided. That's the initial U-turn before you get onto the highway to the truth. The highway is the route to the truth with the least reversals, which is the route an Ockham method traverses. So Ockham does as well as anyone and nobody else does as well as Ockham. And Ockham's advice is uniquely best, it helps you find the truth. And Ockham's advice is best even though Ockham has no idea where the truth is, any more than you do.

Counting particles doesn't seem that much like real science, but it is similar in relevant respects. The "particles" could be edges in causal graphs, polyomial degrees in curve fitting, extra regressors in multiple regression, etc. Each of these cases has the same intuitive structure: extra complexity, like extra particles, is eventually revealed as sample size increases, but not by any fixed time because the extra complexity may be due to arbitrarily small parameter values that require arbitrarily much experience to detect.

Here is an important detail. Suppose that you say "two" after two particles and then after a long time you fall asleep and then the third particle appears so you miss it. When you wake up a fourth particle appears and your lab assistant reminds you of the third particle, which you dozed through. Heavens! You decide you will probably fall asleep again, so you violate Ockham's razor by concluding "four" in case you sleep through the next particle again. Time passes and you change your mind to "three". Now you can be forced in the usual way to change your mind at least once for each successive particle. But you don't do worse than Ockham *overall*, because your initial nap saved you one mind-change that the alert Ockham didn't skip, which cancels the extra U-turn from "four" to "three" that you performed and Ockham didn't. Two wrongs cancel and make a right.

Not really. Efficiency should be forward-looking. No efficiency expert would countenance waste just because employees had stowed away slush funds earlier to pay for future junkets. So if we require that you always be efficient in terms of mind-changes from the present onward, Ockham beats you in the subproblem entered precisely when you conclude four. For in that subproblem you start out with a U-turn that Ockham

does not perform and he beats you.

To summarize the situation in the particle counting problem.

1. In each subproblem, everybody can be forced to change her mind as much as Ockham ever does, so we may say that Ockham is *mind-change efficient*.

2. Anybody who deviates from Ockham can be forced to change her mind more than Ockham in the subproblem entered at the time of the violation, so we may say that your strategy is *mind-change inferior*.

When I say that you can be forced to change your mind "as much" or "more" than Ockham, I mean that you can be dragged, eventually, through each of the theories adopted by the Ockham method, possibly with other guesses interspersed in between (when you think you can outsmart Ockham). That is more than saying that your total mind-change count is the same. It is as though you go down the same highway route, seeing the same trees and houses, but whereas Ockham goes straight home, you possibly get off at the wrong exit and have to turn around and get back on the highway some number of times. So Ockham's route is, as it were, *spliced into* yours.[3]

Philosophers sometimes ask why changing your mind should matter so much, because mind-changes are "merely pragmatic" and epistemology should restrict itself to epistemologically normative concerns— *as if* they had any better idea what this distinction means than they had about analytic/synthetic or essential/accidental. Well, scientific society takes a very dim view of retracting published articles. It does so for a very good reason— since it has no crystal ball to second-guess individual scientists with, it may as well punish them when there is an overt sign that something went wrong. If that isn't normative, what is?

It is also sometimes remarked that mind-changes are a healthy indication of scientific progress. I applaud that sentiment for its candid concession that science unavoidably risks mind-changes, but, of course, only *necessary* mind-changes are vindicated by the necessity of mind-changes. The needless ones should obviously be avoided.

Finally, here is an inductive proof that mind-changes are epistemologically relevant. True belief is evidently of epistemological relevance. Verifiability is also epistemically relevant, since that is what formal proofs, computers and statistical tests give us. But verifiability is just a guarantee of arriving at the truth with at most one mind-change (start with "no" and change your mind to "yes" when the verification arrives). Furthermore, since it is relevant to distinguish verifiability from non-verifiability (the whole point of Gödel's incompleteness theorems, for example). So it seems that being exactly one mind-change away from an epistemically relevant state is epistemically relevant. But being two mind-changes from the truth is being one mind-change from an epistemically relevant state, etc. So being n mind-changes from the truth is epistemically relevant. So each mind-change is epistemically relevant.

With these clarifications out of the way, it is time to take stock of the preceding explanation of Ockham's razor. First, it is driven entirely by the aim of finding the truth

[3]In the rigorous, mathematical development, this is explicated by means of homomorphisms of labelled sequences.

efficiently, where efficiency is a matter of minimizing troublesome reversals of conviction *en route* to the truth. No other preferences are presupposed, much less preferences for simple theories. So the explanation cannot be accused of wishful thinking or of missing the point of inquiry. It does not promise truth immediately, but you should not believe any such promise when it comes to theory selection.

Second, the explanation cannot be attributed to a question-begging prior bias toward simplicity, since no imposed ranking or prior probability enters into the argument at all. As a worst-case argument, it respects each world *and* each theory equally. Recall that Bayesians can't be indifferent at both levels, so they are forced into one bias or another, which makes their attempts at explanation circular one way or the other.

Third, the explanation works without Ouija boards or crystal balls and explains clearly, succinctly, and informally how Ockham's razor could help you find the hidden truth without any need for occult signals. Thus, it resolves the principal puzzle about Ockham's razor.

Fourth, the explanation is not merely that Ockham's razor *suffices* for converging to the truth eventually, as with the standard convergence that cut no ice for Ockham. It demonstrates what we really want to know: that Ockham is *efficient* and that any deviation from Ockham is *inferior*. It is worth remarking that Ockham's efficiency is an even stronger property than being uniquely non-inferior. Efficiency says that everybody else can be forced to do as badly as you do. Inferiority says, roughly, that there is nobody else such that you can be forced to do as badly as her and she cannot be forced to do as badly as you.

Comparing these points with the preceding review of the literature, it is fair to say that the preceding story is not merely the best available explanation connecting Ockham's razor with finding the true theory. There is the only available explanation.

0.6 Consistency and Symmetry

There is a deep intuition in physics, inherited from the ancient Greeks, that it is strongly inappropriate to break symmetries that are unbroken by experience and by the structure of one's theoretical question. For example, if you were told only that a triangle ABC is missing an edge, it would be very bad to say that it is the AB edge that is missing without some further reason— it seems you should suspend belief until further information breaks the symmetry. Since intuitions of symmetry are often associated with simplicity, perhaps the taboo against breaking symmetry unnecessarily is also explained by minimizing mind-changes.

Another strong intuition is not to produce answers that are already refuted by the data. If you did produce such an answer, convergence to the truth demands that you take it back, so that also sounds like it might follow from minimizing mind-changes.

Let's check. But first, it is natural not to count a move from suspension to an informative theory as a retraction, for adding beliefs and publishing theories is joyous; the pain is in losing or retracting them. So I will count as mind-changes only cases in which some content is lost.

Now, suppose the problem is that the machine emits two kinds of particles and that at most finitely many of both will be emitted, although the total number may be different for the two types. Suppose that exactly two of each kind of particle have been seen and that you currently say "(two, two)". Now suppose that you are told that another particle is coming (perhaps you can hear it rattling around in the box) but you can't yet see the type. What should you do? It seems you should suspend belief and wait for further information to break the symmetry.

Why? Suppose that you stick with answer "(two, two)", which is now refuted. The particle that is already rattling in the box pops out and is of the first type, which you must eventually acknowledge, so you bump your guess up to "(three, two)". From now on, nature can exact at least one retraction from you for each successive particle in the usual way. But an Ockham method that never breaks symmetry will retract at most once for each particle. So you are inferior. In light of this sort of argument,

> convergent methods that violate the consistency principle are mind-change inferior in the subproblem entered at the time of the violation.

Suppose, instead, that you jump the gun and guess that the rattling particle will be of the first type and say "(three, two)" when you first hear it. That breaks symmetry. Now the particle may turn out to be of the second type! If no more particles appear, convergence to the truth demands that you change your mind, eventually, to "(two, three)". That is an initial U-turn in the subproblem entered when you first say "(three, two)" that a non-breaker of symmetry wouldn't have committed in the same subproblem. Thereafter, nature can drag you through any mind-change a non-symmetry-breaking Ockham method would commit. So you are inferior. In fact,

> convergent methods that violate the symmetry principle are mind-change inferior in the subproblem entered at the time of the violation.

The preceding argument cannot be escaped by foisting your own bias onto the several simplest theories and claiming that one of them is better for that reason. Bayesian and related methods can be mind-change efficient and I expect that the the practical ones in standard usage typically are when the priors seem intuitive; but only because we have strong native intuitions against symmetry-breaking. This pre-established harmony in practice shouldn't be confused with the idealistic thesis that "anything goes" so far as priors are concerned just because they're your priors! For priors that break symmetries will result in inferior truth-finding performance.

The symmetry and consistency principles follow from a natural statement of Ockham's razor: never choose a theory unless it is *uniquely* simplest compatible with experience. So Ockham's razor entailsthe symmetry and the consistency principles as well as the familiar preference for simpler theories. Furthermore, all of these aspects of Ockham's razor follow from mind-change efficiency.

0.7 Simplicity

The nature of simplicity has long been a vexed question in science and philosophy. Thomas Kuhn (1962) thought that simplicity is a matter of taste, possibly arising from Renaissance dalliances with Neoplatonism. Nelson Goodman (1983) proposed that simplicity is a matter of entrenchment, or of formulation in terms of concepts that have tended to "work out" in the past. In statistical model selection, simplicity is usually a matter of counting free parameters, which ultimately has to do with dimensionality. In algorithmic information theory, simplicity has to do with program length and compressibility, as we have seen.

I view these familiar ideas, at best, as contingent symptoms of a deeper analysis grounded solidly in the intrinsic difficulty of inference *problems* rather than in our sociology, our linguistic conventions, our favorite universal Turing machines or the particular geometry of especially tidy problems. A *theory-choice problem* consists of a topological space in which open neighborhoods correspond to possible *information states*, together with a partition over worlds that corresponds to the *question to be answered*. Problems are understood to be given and it is the methodologist's job to investigate how best to solve them, as in the theories of computability and computational complexity.

According to the proposed analysis, *simplicity* reflects *depth of embedding of iterated problems of induction*. To get some idea of what this is supposed to mean, recall the particle counting problem. Consider all the possible streams of experience in which finitely many particles are emitted in the limit of inquiry. One can arrange these experience streams on an infinitely high bookcase so that the ith shelf consists of all worlds in which exactly i particles appear.

1. The worlds on each shelf satisfy a particular answer.

2. Each world w on a lower shelf poses the problem of induction with respect to worlds on each higher shelf (in the sense that no possible experience true in w excludes all worlds on some higher shelf).

3. It is verifiable that you are at least as high as a given shelf.

So the bookcase amounts to an infinitely iterated generalization of Hume's problem of induction. It is the iterated, Humean structure of the bookcase that drives the efficiency and inferiority arguments discussed above. For if levels of the bookcase correspond to degrees of empirical complexity, then the Ockham method changes its mind at most n times for a world on shelf n. And this performance is efficient. For let another method be given. Show it a world at level 0. On pain of failing to converge to the truth, the method must eventually say 0 on increasing information in that world. But given property (2) of the bookcase, every higher shelf is consistent with the currently available information. Hence, one may choose a world compatible with experience at level 1 and feed information consistent with what has been seen so far from that world until the method eventually concludes 1, and so forth. So the method can be forced to change its mind at *least* n times in each answer n. Indeed, the pattern of outputs

produced by the method will extend the pattern produced by Ockham, so Ockham is efficient. The U-turn argument can be implemented in bookcases in a similar manner.

The same idea works if data are open neighborhoods in an infinite-dimensional, continuous parameter space. In this example, there are parameter vectors with arbitrarily many non-zero values arbitrarily close to the all zero vector, so any open set catching the all zero vector will catch members of every other answer. So Euclidean dimension is just a particular way of generating simplicity.

Here is an important advantage of the proposed account of empirical complexity. Suppose that God tells you that the true parameter vector is rational-valued. Suddenly the infinite-dimensional Euclidean space becomes zero-dimensional! Does the concept of simplicity collapse along with dimension? Not intuitively. And not in the proposed theory, since the rational-valued vectors preserve the nested problems of induction even though they don't preserve dimensionality. Theory dimension is a *symptom* of simplicity, not the genuine article.

The particle example is special, because all the worlds in the problem fit in one bookcase. For a more complicated example, recall the problem in which there are two particle types. Consider all the bookshelves we can set up for subsets of this problem that satisfy the above three rules. Each such bookcase is *demonic* in the sense that nature has a strategy to force an arbitrary learner to change its mind from one shelf to the next. Some of these bookcases will map into others in a way that preserves shelf-order and the answer corresponding to the shelf.

In this more general setting, I understand Ockham's razor as follows. Consider potential answer A to the question at hand. Now consider whether putting A on the bottom shelf of each bookcase and moving all other answers upward one step results in modified bookcases that map (in the manner just described in the preceding paragraph) into original bookcases? If so, you may select A. If there is no such A, you must suspend judgment. This general formulation of Ockham's razor entails preference for simplicity, respect for symmetry, and consistency with the data.

Furhermore, it can be shown, for a broad range of problems, that

1. Every solution to the problem that violates Ockham's razor is inferior and

2. There exists an Ockham method that is efficient.

To see why, recall the example with two particle types. Suppose you hear the fateful rattling of the first particle in the emitter but haven't seen it yet. At this point, there are demonic bookcases with worlds for "(one, zero)" and for "(zero, one)" on the bottom shelves, respectively.

For either answer, if you were to concatenate it on the bottom of every possible demonic bookcase, you would end up with a modified demonic bookcase that maps into no original demonic bookcase in the sense described above. For if you try the answer "(one, zero)", then tacking it on the bottom of the original bookcase starting with "(zero, one)" violates rule (2) for demonic bookcases, since there is information true in "(one, zero)" that excludes all worlds in "(zero, one)". The case of answer "(zero, one)" is, of course, symmetrical.

So Ockham says to wait. And Ockham is right, for if you persist in producing one of the forbidden answers, say "(zero, one)" before the matter is settled, then nature can reveal the heard particle to be of the second type, forcing you into "(one, zero)". Thereafter, you can be dragged the rest of the way up any demon book-case starting with "(zero, one)", etc. An Ockham method, on the other hand, would have started with "(zero, one)" and would do no worse than to traverse the shelves that any method could be dragged through.[4] So you are mind-change inferior.

Conditions (1) and (2) in the demon tower definition essentially involve both the underlying information topology of the problem *and* the question asked. That is as it should be. The whole point is to argue that Ockham is more efficient at solving inductive problems than are other methods. As everyone who takes a course in introductory algorithm analsysis learns, efficiency is always relative to the problem to be solved. So if Ockham is to have a truth-finding efficiency justification, simplicity *must* depend on the structure of the problem to be solved.

There is also an intuitive advantage of making simplicity relative to the question asked. Any particular law L we decide to pick out and test against alternative not-L ends up at the bottom of the demon bookcase for that (binary) problem, so L is the unique Ockham answer until L is refuted. So the typical practice of treating a point (refutable) hypothesis as the null hypothesis in a statistical test is an instance of Ockham's razor! Absolutist theories of simplicity that assign simplicity degrees to worlds independently of the question asked can only say that testing practice and Ockham's razor are two entirely different things.

Another important advantage of making simplicity relative to the question asked is that it makes simplicity "grue-proof". Uniformity and compressibility of sequences depends on the character of the sequence and can disappear when the sequence is relabelled. But such relabellings (e.g., grue) are actually topological symmetries (auatomorphisms) of the underlying sequence space. But the material content of the answers to a given question *are* preserved under translation, grue-like or otherwise, and so is the topological structure of the underlying space. Hence, *everything in the proposed statement and vindication of Ockham is immune to the sort of Goodmanian critique that has plagued other approaches.*

0.8 A Word on Statistical Applications

The preceding ideas are defined only for deterministic inference problems. In statistical problems, the idea is to apply the same definition of Ockham's razor at the level of statistical parameters. Convergence in the limit becomes convergence in probability and mind-changes are viewed as occuring *in probability* when the method probably chooses one theory at a given sample size and a different theory at a later sample size (Kelly and Glymour 2004, Kelly 2004). Then any model selection technique that is guaranteed to converge in probability to the true *model* can be dragged through at

[4]The general argument that Ockham does no worse than get dragged up a demon bookcase is too involved to present here.

least one retraction per dimension, just as in the deterministic case.

For example, the Bayes' information criterion (BIC) scoring rule (similar to AIC but capable of converging to the true model) does quite nicely. In the easy model selection problem of inferring how many components of a bivariate mean are nonzero, BIC changes its mind once for each zero. If the true mean value is (.05, .005), then BIC starts out somewhat skeptical, mixing its outputs among the various answers so that no answer is produced with high probability. Then it starts to become convinced that both components of the mean are zero (at around $n = 200$). Then it switches to thinking that just one component is zero (at around $n = 30,000$). Finally, when the sample size is in the millions, BIC realizes that both components of the mean are nonzero.

So BIC avoids an initial U-turn in probability by probably producing the answers in ascending order of complexity, just as the proposed vindication of Ockham's razor requires. Almost any natural method that converges in probability to the truth and that is set up with a simplicity bias should also satisfy Ockham's razor and be efficient in the proposed sense (e.g., Bayesian updating or MDL using codes that reflect the number of nonzero components of the mean). Since AIC does not even converge to the truth in this problem (it never fully takes the bait that the world is simple), it cannot be forced (in probability) into two mind-changes.

The statistical mind-change idea shows promise as a unifying theme for the *ad hoc* departments of classical statistics. Tests are just binary problems requiring one mind-change. Estimation is a matter of verifying degrees of approximation with no mind-changes. And model selection is a (comparatively nasty) problem that requires arbitrarily many. The piece-meal approach characteristic of the classical literature on model selection suggests that ideas suited to zero or just one mind-change are being applied to problems of much higher intrinsic complexity, with expectably confusing results.

0.9 Ockham, *Angst*, and Misinformation

It is no accident that the proposed explanation of Ockham's razor has been missed for so long. As David Hume wryly remarked, it is almost impossible for humans to take the problem of induction seriously. Thomas Kuhn (1962) even suggested that when science does change its mind, the fact is so upsetting that it must be expunged from the historical in favor of a myth of monotone progress. In statistics, the revisionist tendency emerges as the myth that every scientific inference is either a verification (of the denial of the null hypothesis) or a verified degree of approximation. But none of those verificationist tricks work for learning *theories* because choosing a theory, with all its counterfactual implications, really amounts to *accepting* a null hypothesis (i.e., that some potential complexity precluded by the theory will *never* appear)— and that null hypothesis is universal, unverifiable, and therefore subject to Hume's problem of induction.

We do infer theories, they are subject to the problem of induction, and we may have

to take them back. Indeed, there is no bound on the number of times you might have to change your theory as new layers of complexity are successively revealed. Ockham's razor puts you on the highway to the truth but doesn't straighten the highway for you— you have to traverse all of the upcoming curves and passes on your own. If it were otherwise, then Ockham would have to see your unknown route to the truth in advance, which only an epistemic crystal ball could do. So *embracing* the existential *angst* of unsupported inductive leaps with their attendant risk of disastrous retraction is the key to understanding Ockham's razor. Ironically, the very longing for certainty and verification that attracts us to Ockham's razor prevents us from ever understanding how it works.

I suspect that much of the appeal of information theory in the study of learning stems from our perennial unwillingness to concede that science is ultimately guesswork. "Information" is really just a rubric for some formal analogies between string length and logarithms of probabilities. In the technological problem of optimizing signals, the formal relationships really do have something to do with sending genuine information to somebody else, for if you want to minimize expected message length, it is a good idea to set up your coding scheme so that more frequent signals have shorter lengths.

But sending signals down a wire isn't remotely the same thing as choosing a theory on the basis of empirical data. In learning applications, one uses the formulas that apply to communication channels, but only in a purely formal, metaphorical way (Mitchell 1997). With ingenuity, one can make shortest program length approximately satisfy the probability calculus. With some more ingenuity, one can construct agents whose prior opinions are biased toward strings with shorter programs. And since the same formal techniques are involved, one is licensed by academic courtesy to continue calling it "information theory". But this otherwise harmless professional courtesy is extremely unfortunate in the application to learning because the word "information" already means something there: it refers to input data and background knowledge in learning problems. For this reason, any talk of "information" can readily be interpreted as a kind of knowledge or data beyond what is really known or observed. Compounding the misfortune is the perennial temptation to view simplicity as a kind of occult information about the unknown. How easy it is for the gullible to clutch at the hope that *somehow*, deep in the cabalistic mysteries of information, computation, Gödel, Escher, and Bach, there *really is* an occult connection between simplicity and reality that will direct us unswervingly to the Truth; that prior probabilities constructed to favor computationally compressible strings really are *informative* and that learning can be *defined* as data-compression. After all, aren't these things constructed out of "information"? I know whereof I speak— I have met these glassy-eyed wretches (full professors, even) and they are beyond salvation. It would be better, all things considered, to revoke conventional academic courtesy in this particular instance and to simply *banish* the word "information" outright from the the study of learning and induction. There is already a symbol for the logarithm of a probability function. There is already a word for "data". Nobody is even tempted to confuse them.

0.10 Acknowledgements

I dedicate this paper to Clark Glymour, who stamped the conundrum of Ockham's razor irrevocably upon me in 1981. My admirably open-minded statistics colleague, Larry Wasserman kindly encouraged this research but is definitely not to be blamed for the wicked views expressed. I am also grateful for the indefatigable curiosity of my current students Seth Casana and John Taylor. Finally, I reserve particular thanks for my stalwart colleagues Oliver Schulte, Richard Scheines, and especially Joseph Ramsey, who exhibited patience above and beyond the call of collegiality in discussing the endless and ongoing evolution of these ideas.

0.11 Bibliography

Akaike, H. (1973) "Information theory and an extension of the maximum likelihood principle", *Second International Symposium on Information Theory.* pp. 267-281.

Carnap, R. (1950) *Logical Foundations of Probability*, Chicago: University of Chicago Press.

Forster, M. R. and Sober, E. (1994): How to Tell When Simpler, More Unified, or Less Ad Hoc Theories will Provide More Accurate Predictions. *The British Journal for the Philosophy of Science* 45: 1-35.

Freivalds, R. and C. Smith (1993) "On the Role of Procrastination in Machine Learning", *Information and Computation* 107: pp. 237-271.

Friedman, M. (1983) *Foundations of Space-Time Theories*, Princeton: Princeton University Press.).

Glymour, C. (1980) *Theory and Evidence*, Princetion: Princeton University Press.

Goodman, N. (1983) *Fact, Fiction, and Forecast*, fourth edition, Cambridge: Harvard University Press.

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

Jain, S., Osherson, D., Royer, J. and Sharma, A (1999) *Systems That Learn: An Introduction to Learning Theory.* Camabridga: M.I.T. Press.

Kearns, M. and Vazirani (1994) *An Introduction to Computational Learning Theory*, Cambridge: M.I.T. Press.

Kelly, K. (1996) *The Logic of Reliable Inquiry.* New York: Oxford.

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. 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.

Kechris, A. (1991) *Classical Descriptive Set Theory*, New York: Springer.

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

Li, M. and Vitanyi, P. (1997) *An Introduction to Kolmogorov Complexity and Its Applications*, New York: Springer.

Mitchell, T. (1997) *Machine Learning.* New York: McGraw-Hill.

Putnam, H. (1965) "Trial and Error Predicates and a Solution to a Problem of Mostowski", *Journal of Symbolic Logic* 30: 49-57.

Popper, K. (1968), *The Logic of Scientific Discovery*, New York: Harper.

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

Rosenkrantz, R. (1983) "Why Glymour is a Bayesian", in *Testing Scientific Theories*, J. Earman ed., Minneapolis: Universitiy of Minnesota Press.

Salmon, W. (1967) *The Logic of Scientific Inference*, Pittsburgh: University of Pittsburgh Press.

Schulte, O. (1999a) "The Logic of Reliable and Efficient Inquiry", *The Journal of Philosophical Logic*, 28:399-438.

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

Schulte, O. (2001) "Inferring Conservation Laws 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.

Vitanyi, P. and Li, M. (2000) "Minimum Description Length Induction, Bayeisanism, and Kolmogorov Complexity". *IEEE Transactions on Information Theory* 46: 446-464.

Wasserman, L. (2003)*All of Statistics: A Concise Course in Statistical Inference.* New York: Springer.