Sober as a Judge[†]

Gordon Belot[‡]

Elliott Sober has a longstanding interest in delimiting the epistemic (as opposed to pragmatic or aesthetic) relevance of parsimony considerations. He tells us in his marvelous new book, *Ockham's Razors,* that his goal "is to determine when parsimony is relevant and when it is not. It is obvious that simple theories may be beautiful and easy to remember and understand. The hard problem is to explain why the fact that one theory is simpler than another tells you anything about the way the world is" (2). This is a ferociously difficult problem—and although it lies at the heart of much of contemporary philosophy of science and epistemology, it is embarrassingly often complacently shirked rather than confronted. Few philosophers have worked so hard, judiciously, and productively on the problem as Sober has—and *Ockham's Razors* provides an invaluable synthesis and overview of this work.

At the core of *Ockham's Razors* is the fascinating analysis in Chapter 2 of two "parsimony paradigms." According to the first, data often favour simpler hypotheses, in the sense of conferring higher likelihoods on them than on their more complex rivals. According to the second, data often favour simpler models over more complex ones, "because the number of adjustable parameters in a model helps you to estimate its predictive accuracy" (141).

My focus here will be on the second of these paradigms. Sober's picture is nicely encapsulated in the following passage, discussing a number of related methods of model selection: "they agree that parsimony, as measured by the number of adjustable parameters in a model, is relevant to making those estimates. It isn't just that parsimony *seems* relevant; there are mathematical arguments for why it *is* relevant. Ockham's razor is alive and well in statistics." (141) The mathematical results in question establish that certain methods of model selection enjoy good asymptotic behaviour—e.g., some are essentially guaranteed to achieve predictive accuracy in the limit of large data sets.

So we have an appealing argument: certain methods are good methods for achieving predictive accuracy because they enjoy good asymptotic behaviour; because these methods take account of the number of variables in a model, parsimony is epistemically relevant for those interested in predictive accuracy. I would like to raise some worries about this form of argument—and this to take this opportunity to become more clear about the structure of Sober's thought about his second parsimony paradigm.

[†] Forthcoming in *Metascience*, in a symposium on Sober's *Ockham's Razors: A User's Guide.*

[‡] Department of Philosophy, University of Michigan. belot@umich.edu

A Gentle Warmup

It will be helpful to begin with something simple. Suppose that a coin is to be tossed once a day. After each toss you are told the outcome (0=tails, 1=heads) and asked to put forward a conjecture as to the bias p of the coin in favour of heads (i.e., the chance p of the coin coming up heads on any given toss). A method for approaching this problem is a function from finite binary sequences to possible values for the bias of the coin.

One natural method for handling this problem is the *straight rule*: if you have seen k heads in m tosses, conjecture that the bias in favour of heads is k/m. A signal virtue of this method is that it is virtually guaranteed to lead you to the truth. More precisely, it is *statistically consistent*: whatever the true bias of the coin, there is zero chance that a data stream will be generated that when fed to the straight rule would result in a sequence of conjectures that failed to converge to the true bias.

Now consider the following question: Is the order of heads and tails in a data set—in addition to their proportions—relevant to the problem of determining the bias of a coin?

Fans of the straight rule may be tempted to argue for the irrelevance of order as follows.

> The straight rule has good asymptotic behaviour for this problem—it is statistically consistent. So it is a good method for determining the bias of a coin. But the straight rule cares only about the proportions of heads and tails in a data set, not about the order in which they occur. So order is epistemically irrelevant to the problem of determining the bias of a coin.

Now, statistical consistency is a substantive condition: many methods for approaching our problem are statistically inconsistent (e.g., making conjectures at random; or remaining certain that the coin is fair, no matter what data is seen; or throwing away every second bit of data and then applying the straight rule).

But because it cares only about behavior in the infinite long run, statistical consistency is in a sense a very weak requirement. Any method for our problem whose conjectures tend to converge to those of the straight rule will be statistically consistent. There are many such methods—some of which care about the order of heads and tails in a data set as well as about their proportions.

Consider the *unprimed straight rule*: given a data set (x_1, …, x_m) first discard the bits with prime indices (i.e., discard x_2, x_3, x_5, x_7, and so on), then apply the straight rule to the resulting data set (x_1, x_4, x_6, x_8, …). For any data stream, the outputs of the unprimed straight rule will converge to a number p if and only if the outputs of the straight rule do so. So the unprimed straight rule is statistically consistent. But unlike the straight rule, the unprimed straight rule cares about order: shown 01010, it applies the straight rule to 01 and conjectures that the bias of the coin is .5; shown 10010, it applies the straight rule to 11 and conjectures that the bias of the coin is 1.

If the argument above for the irrelevance of order is good, then so is the following argument for the relevance of order:

> The unprimed straight rule has good asymptotic behaviour for this problem—it is statistically consistent. So it is a good method for determining the bias of a coin. But the unprimed straight rule cares about the order in which heads and tails occur in a data set as well as about their proportions. So order is epistemically relevant to the problem of determining the bias of a coin.

How can we avoid the conclusion that order both is and isn't relevant to the problem of determining the bias of a coin? There are a number of options. We could relativize relevance to commitment to a method—order is relevant for those committed to the unprimed straight rule, irrelevant for those committed to the straight rule. Relatedly, we could avail ourselves of the resources of subjective Bayesianism. Or we could deny that the fact that a good method for our problem takes a certain consideration into account suffices to show that that consideration is epistemically relevant to our problem. Each of these is an interesting option. But I will set them aside here, on the grounds that they are alien to Sober's approach.

But there is another option: we could deny that statistical consistency is a sufficient condition for the goodness of a method. This gambit has deep roots in Sobriety. Indeed, in considering a problem similar to our coin problem (namely, the problem of determining the mean height of a population from samples), Sober has argued that statistical consistency is too weak to be plausibly regarded as being a sufficient condition for a method to be reasonable, and suggested that further asymptotic considerations such as efficiency and lack of bias should be used to single out those statistically consistent methods that are genuinely reasonable (Sober 1988, 229).

However, there are ample grounds to suspect that considerations of this kind will fail to single out an elite class of statistically consistent methods that agree on such questions as whether order is relevant in determining the bias of a coin. For instance: while the straight rule is unbiased and in an interesting sense optimally efficient, there are many other unbiased methods optimally efficient in the same sense, including some that take order into account (Juhl 1994).

Sober on Parsimony and Model Selection

Sober is interested in the problem that statisticians call model selection. For concreteness, let us suppose that an infinite binary sequence will be revealed to us one bit at a time. A *statistical hypothesis* is a hypothesis about how this data stream is being generated (i.e., a probability distribution over possible data streams). A *model* of size k is a k-dimensional family of statistical hypotheses. To give some tame examples: there is a model of size zero comprising just the hypothesis that the data stream is being generated by tossing a fair coin; and there is a model of size one, that includes for each p between zero and one the hypothesis that the data stream is being generated by tossing a coin with bias p in favour of heads; and there is a model of size two, that includes for each p_1 and p_2 between zero and one, the hypothesis that the data stream is being generated by alternating between tossing a coin with bias p_1 in favour of heads and tossing a coin with bias p_2 in favour of heads.

A method of model selection is a rule that takes as input a list of models and a data set and gives as output one of the models under consideration. A paradigmatic example is the Akaike Information Criterion (AIC):

> *AIC.* For each model M under consideration, calculate for each hypothesis in the model how likely the data seen is relative to that hypothesis. Call the maximum such number $L(M)$. The *AIC score* of M is: log $L(M) - k$, where k is the size of M. Select the model with the highest AIC score.

Akaike showed that the AIC score has a remarkable feature: in the limit of large data sets, the AIC score of a model provides an unbiased estimate of the predictive accuracy of the model. (What does predictive accuracy mean here? For any model M and any (n+1)-tuple (x_1, …, x_n, y) of data points, we imagine that (x_1, …, x_n) is our data set, find the statistical hypothesis θ_0 in M that confers maximum likelihood on that data set, and find the probability p that θ_0 assigns to the next bit seen being a one. Then $(y - p)^2$ is the *loss* for M relative to (x_1, …, x_n, y). We are working in a context in which data sets are chosen randomly in accord with some underlying statistical hypothesis Θ lying in one of the models under consideration. And we know, for each way (x_1, …, x_n, y) could be chosen randomly, what the loss of M would be. So we can take the expectation value relative to Θ of the loss of M. That is the *predictive accuracy* of M for an n-point data set. Akaike's result says that, normalizing constants aside, as $n \to \infty$, this quantity is approximated by the AIC score of *M.*)

Why does the AIC score work so well? It can be thought of as weighing the size of a model against the ability of that model to fit current data. This is a reasonable strategy because the more parameters a model has the easier it is for it to fit a given data set—so a given degree of fit is not equally impressive in models of different sizes. The trick is to find the right way to weigh size against fit of current data. As Sober remarks, "AIC doesn't tell you to throw up your hands and say that it is a matter of subjective preference how much parsimony matters as compared with fitting the data. No, the criterion says that these two considerations are commensurable and tells you how to commensurate them" (148).

In explaining what it would mean for parsimony to be epistemically relevant, Sober identifies as one relevant question whether "the fact that S is simpler than C [helps to] justify the claim that S will make more accurate predictions in the future than C will" (59). And in light of the sort of result mentioned above regarding the asymptotic behaviour of AIC, it would appear that an affirmative answer is unavoidable here—"parsimony is relevant because the number of adjustable parameters in a model helps you estimate its predictive accuracy" (141).

So we appear to have an argument for the relevance of parsimony to the problem of model selection:

> AIC has a certain desirable asymptotic behaviour in regard to predictive accuracy. So AIC is a good method of model selection when we value predictive accuracy. In implementing AIC, one takes model size into account. So parsimony is epistemically relevant in the context of model selection for the purpose of predictive accuracy.

A Dilemma

This argument for the relevance of parsimony has an evil twin, standing to it as our argument for the relevance of order stood to our argument for the irrelevance of order in the problem of the biased coin.

Consider a widely-used alternative to AIC, cross-validation (CV).

> CV. Fix a model M and n-point data set $S=(x_1, \ldots, x_n)$. Let $S_{(1)}$ be the (n-1)-point data set that results from dropping x_1 from S. Choose the hypothesis in M that confers maximum likelihood on $S_{(1)}$ and let p_1 be the probability that that hypothesis confers on the missing data point x_1 and let d_1 be the square of the difference between p_1 and the actual value of the missing data point. Likewise define $S_{(i)}$, p_i, and d_i for each i=2, …, n. The *CV score* of the model is sum $d_1+\ldots+d_n$. Select the model with the smallest CV score.

As Sober notes (133 fn. 52), although CV "makes no overt use of parsimony," it is known that for a wide class of model selection problems, AIC and CV are asymptotically equivalent to one another—for sufficiently large data sets, they always select the same model.

So in the limit of large data sets, following CV is as good a strategy as following AIC (since the two strategies select exactly the same models). So if the fact that the AIC score is a good asymptotic estimator of predictive accurcy implies that AIC is a good method of model selection in the limit of large data sets, then following CV is also a good method in that regime.

So if the argument given above for the epistemic relevance of parsimony to predictive accuracy is sound, so is the following argument for the epistemic irrelevance of parsimony.

> CV has a certain desirable asymptotic behaviour in regard to predictive accuracy. So CV is a good method of model selection when we value predictive accuracy. In implementing CV, one does not take model size into

> account. So parsimony is epistemically irrelevant in the context of model selection for the purpose of predictive accuracy.

We have landed in the same sort of situation as in the case of the biased coin. We are in danger of being forced to accept that parsimony is both relevant and irrelevant to predictive accuracy in model selection.

The most attractive route to avoiding this conclusion proceeds via denying that good asymptotic behaviour is sufficient for a method of model selection to be good. Careful readers (and authors!) of Sober's work will be justifiably impatient to observe that there is a passage in *Ockham's Razors* that makes just this move: on pp. 132 f., Sober notes that Akaike's result concerning the asymptotic relation between expected predictive accuracy and the AIC score leaves open questions about other aspects of the behaviour of AIC (e.g., it could be pathologically erratic in the sense of having high variance), and concludes that "Akaike's result, by itself, does not suffice to justify your using AIC. This, of course, leaves it open that there are other mathematical results that close the gap" (see also Sober 2008, 87).

From this perspective, it is at present an open question either AIC or CV is a good method of model selection. In one sense, that is promising for Sober's project—if it should turn out that all good methods of model selection take model size into account, then we could say that parsimony is epistemically relevant without also having to say that opposite.

But in the meantime, it leaves us able only to assert very weak conditional conclusions: if future discoveries determine that all good methods of model selection are like AIC rather than CV, then parsimony is epistemically relevant to model selection for predictive accuracy—but if all good methods are like CV rather than AIC, then parsimony is epistemically irrelevant. These are much weaker than the verdicts handed down by Sober in *Ockham's Razors.*

Worse: as in the biased coin case, there is ample room doubt that results will emerge that serve to single out a determinate elite class of good methods for model selection that agree as to questions like the relevance of parsimony. On the question of efficiency, for instance, the situation is reminiscent of the biased coin case: there is a sense in which AIC is an especially efficient method; and across a wide range of situations, AIC and CV are known to be equivalently efficient in that sense (Shao 1997).

I claim, then, that Sober faces a choice: either accept that parsimony is both relevant and irrelevant to model selection; or allow that as things stand we have no grounds for claiming either that parsimony is relevant or that it is irrelevant. That is: either render a verdict of *both guilty and not guilty* or a verdict of *not proven*.

References

Juhl, Cory. 1994. The speed-optimality of Reichenbach's straight rule of induction. *British Journal for the Philosophy of Science* 45: 857–863.

Shao, Jun. 1997. Asymptotic theory of linear model selection. *Statistica Sinica* 7: 221–242.

Sober, Elliott. 1988. Likelihood and convergence. *Philosophy of Science* 55: 228–237.

Sober, Elliott. 2008. *Evidence and evolution: The logic behind the science*. Cambridge: Cambridge University Press.