Simplicity is *Not* Truth-Indicative

Bruce Edmonds

Centre for Policy Modelling
Manchester Metropolitan University
`http://cfpm.org/~bruce`

In this paper I will argue that, in general, where the evidence supports two theories equally, the simpler theory is *not* more likely to be true and is *not* likely to be nearer the truth. In other words simplicity does not tell us anything about model bias. Our preference for simpler theories (apart from their obvious pragmatic advantages) can be explained by the facts that humans are known to elaborate unsuccessful theories rather than attempt a thorough revision and that a fixed set of data can only justify adjusting a certain number of parameters to a limited degree of precision. No extra tendency towards simplicity in the natural world is necessary to explain our preference for simpler theories. Thus Occam's razor eliminates itself (when interpreted in this form).

I will start by arguing that a tendency towards elaboration and the pragmatic advantages are sufficient to explain our preference for simper theories. Then I will briefly look at a couple of a priori arguments justifying a bias towards simplicity. I follow this by reviewing some evidence as to whether simpler theories are likely to be true taken from the field of Machine Learning, followed by a section discussing some special cases where we have some reason to expect there to be a bias towards simplicity. I will briefly consider some of the concepts that have been called "simplicity" in the literature before I conclude with a plea for the abandonment of the use of simplicity as justification.

Elaboration

If one has a theory whose predictions are insufficiently accurate to be acceptable, then it is necessary to change the theory. For human beings it is much easier to elaborate the theory, or otherwise tinker with it, than to undertake a more radical shift (for example, by scrapping the theory and starting again). This elaboration may take many forms, including: adding extra variables or parameters; adding special cases; putting in terms to represent random noise; complicating the model with extra equations or rules; adding meta-rules or models; or using more complicated functions. In Machine Learning terms this might be characterised as a preference for depth-first search over breadth-first search.

Classic examples of the elaboration of unsatisfactory theories include increasing the layers of epicycles to explain the observations of the orbits of planets in terms of circles and increasing the number of variables and equations in the national economic models in the UK. In the former case the elaboration *did* increase the accuracy because the system of epi-cycles can approximate the collected data as to the true orbits, but this is more informatively done with ellipses. Once the arbitrary bias towards circles is abandoned the system of epi-cycles becomes pointless. In the later case the elaboration has not resulted in the improved prediction of future trends (Moss et al. 1994), and in particular they have failed to predict *all* the turning points in the economy using these models.

Why humans prefer elaboration to more radical theory change is not entirley clear. It may be that it is easier to understand and predict the effect of minor changes to the formulation of theory in terms its content, so that, if one wants to make a change where one is more certain of improvement, minor changes are a more reliable way of obtaining this. It may be that using a certain model structure biases our view because we get used to framing our descriptions and observations in this way, using variations of the model as our 'language' of representation. It may be due to simple laziness - a wish to 'fit' the current data quickly rather than holding out for longer-term predictive success.

Regardless of the reasons for elaboration, we are well aware of this tendency in our fellows and make use of this knowledge. In particular we know to distrust a theory (or a story) that shows signs of elaboration - for such elaboration is evidence that the theory might have *needed* such elaboration because it had a poor record with respect to the evidence. Of course, elaboration is not proof of such a poor record. It may be that the theory was originally formulated in an elaborate form before being tested, but this would be an unusual way for a human to proceed.

This knowledge, along with an understandable preference for theories that are easily constructable, comprehensible, testable, and communicable provide strong reasons for choosing the simplest adequate theory presented to us.

In addition to this preference for choosing simpler theories, we also have a bias towards simpler theories in their construction, in that we tend to start our search with something fairly simple and work 'outwards' from this point. This process stops when we 'reach' an acceptable theory (for our purposes) - in the language of economics we are satisficers rather than optimisers. This means that it is almost certain that we will be satisfied with a theory that is simpler than the best theory (if one such exists, alternatively a better theory). This tendency to, on average and in the long term, work from the simpler to the less simple is a straightforward consequence of the fact that there is a lower bound on the simplicity of our constructions. This lower bound might be represented by single constants in algebra; the empty set in set theory; or a basic non-compound proposition expressed in natural language.

This constructive bias towards simplicity is also a characteristic of other processes, including many inductive computer programs and biological evolution. Evolution started from relatively simple organisms and evolved from there. Obviously when life started the introduction of variety by mutation would be unlikely to result in simplification, since the organisms were about as simple as they could get while still being able to reproduce in its environment. Thus the effective lower bound on complexity means that there is a passive drift towards greater complexity (as opposed to an active drive towards complexity, a distinction made clear by McShea, 1996). However this bias is only significant at the start of the process because the space of possible organisms is so great that once any reasonably complex organism has evolved it is almost as likely to evolve to be simpler as more complex -

the lower bound and the 'inhabited' part of the possibility space do not impinge upon the possibilities that much.

A Priori Arguments

There have been a number of *a priori* arguments aimed at justifying a bias towards simplicity - (Kemeny 1953) and (Li, M. and Vitányi, 1992) are two such. The former makes an argument on the presumption that there is an expanding sequence of hypotheses sets of increasing complexity and a completely correct hypotheses - so that once one has reached the set of hypotheses that contains the correct one it is not necessary to search for more complex hypotheses. However this does not show that this is likely to be a better or more efficient search method than starting with complex hypotheses and working from there. The later shows that it is possible to code hypotheses so that the shorter codes correspond to the more probable ones, but in this case there is no necessary relation between the complexity of the hypotheses and the length of the codes that is evident *before* the probabilities are established.

To show that such prior arguments are unlikely to be successful, consider the following thought experiment. In this experiment there are two 1kg masses, A and B, of the same weakly radioactive material, in which atoms currently decay at an average rate of 1 atom per minute. By each mass there is a Geiger counter which detects when an atom in the mass decays and sends a particle towards the counter. The task is to predict which counter will register an particle first after each hour on the clock begins. Now any model which predicts A and B half the time will, in the long run, do equally well. In this case it is abundantly clear that simpler theories are not more likely to be correct - correctness is determined by the proportion of A and B that the theory predicts and nothing else.

Now, quite reasonably, one might object that a sensible model concerning radioactive decay is not a directly predictive one but one which specifies the unpredictability of the phenomena and concentrates on 'second-order' properties such as the probability distribution. However, this is beside the point - it is a truism to say that those phenomena where our simple theorising succeeds do have some simple behaviour and those where such theories do not hold require more complex ones. If the thesis that simplicity is truth-indicative is restricted to only those aspects of the natural world where it works, it has force but then can not be invoked to justify the selection of theory about phenomena in general. We rightly do not attempt to predict the *exact* position of *each* grain of sand with our mathematical models of sand piles but instead concentrate on those aspects of that *are* amenable to our modelling techniques,such as relation between the frequency and size of avalanches (Bak 1997). In general we are highly selective about what we attempt to model - we usually concentrate upon that tip of the natural world iceberg which is not overly complex.

Theoretical results in Machine learning (Schaffer 1994, Wolpert 1996) show that, in general, no learning or search algorithm is better than another. In particular that if a bias towards simplicity is sometimes effective, there must be other domains in which it is counter-productive. To gain any improvement in inductive ability one must apply knowledge about the particular domain one is concerned with. However, these results are extremely abstract and dominated by search spaces that are seemingly random and discontinuous almost everywhere. It may be that nature is biased towards producing data that is more amenable and, in particular, simple than these extreme cases. Thus we look to some evidence as to this.

Some Evidence from Machine Learning

We have two explanations for our preference for simpler theories once the pragmatic advantages are factored out (all evidence being equal): *firstly*, our knowledge that theories tend to be elaborated when unsuccessful and, *secondly*, an inherent bias towards simplicity in the natural world. If we *were* to hold to Occam's razor (in the form that simplicity is truth-indicative) then we would choose the first because this is sufficient to explain the phenomena - the postulated bias in the natural world is an 'unnecessary entity'.

Since I don't hold with this form of Occam's razor I need to look for some evidence to distinguish between the two explanations. Since the tendency towards elaboration is a characteristic of human theory construction, we look to situations where theory construction is not biased towards elaboration to see if simplicity is truth-indicative there. Recently there have been such studies in the field of Machine Learning - where a computer program (rather than a human) attempts the induction. This gives one a test bed, for one can design the induction algorithm to use a simplicity bias or otherwise and compare the results. In one of these studies (Murphy and Pazzani 1994) a comprehensive evaluation of *all* possible theories in a given formal language (to a given depth) were analysed against some real-world data series as follows: *firstly* as to their effectiveness at fitting some initial portion of the data (the in-sample part of the series), *secondly* as to their success predicting the continuation of this data (the out-of-sample part), and *finally*, as to the theory's complexity (measured in this case by the size or depth of the formal expression representing the theory). The theories with best success at fitting the in-sample data were selected. Within this set of 'best' theories it was examined whether the simpler theories predicted the out-of-sample data better than the more complex theories. In some cases the simpler hypotheses were not the best predictors of the out-of-sample data. This is evidence that on real world data series and formal models simplicity is not necessarily truth-indicative.

In a following study on artificial data generated by an ideal fixed 'answer', (Murphy 1995), it was found that a simplicity bias was useful, but only when the 'answer' was also simple. If the answer was complex a bias towards complexity aided the search. Webb (1996) exhibited an algorithm which systematically extended decision trees so that they gave the same error rate on the in-sample data, and, on average, gave smaller error rates on the out-of-sample data for several real-life time series. This method was based upon a principle of similarity, which was used to restrict the set of considered hypotheses. A useful survey of results in Machine Learning, that can be seen as a parallel paper to this one is (Domingos 2000).

Thus, the evidence, is that when considering non-human induction, that a simplicity bias is not necessarily helpful or truth-indicative. Rather that it is often used as an ill-defined satand-in form some domain knowledge. A bias towards simplicity does seem to be a particular feature of human cognition (Charter 1999).

Special Cases

Although, simplicity is not *in general* truth-indicative, there are special circumstances where it might be. These are circumstances where we have some good reason to expect a bias towards simplicity. I briefly consider these below.

The first is when the phenomena are the result of deliberate human construction. Deliberate human constructions are typically amenable to an almost complete analysis assuming a design stance, they are frequently modular, and the result of simple principles iterated many times. If someone asks you to guess the next number is the sequence: 2 ,4, 8, 16 you will correctly guess 32, because the n^{th} *power of two* is the simplest pattern that describes these five numbers, and you an rely on the fact that the human will have chosen a simple (albeit possibly obscure) rule for their construction. It would not be sensible to guess the number 31, despite the fact that there is *a* rule that would make this the correct answer (the number of areas that *n* straight lines, each crossing the perimeter of a circle twice and such that no three lines intersect in a single point, cut that circle into).

The simplicity of these kinds of phenomena is only a hallmark of deliberate, conscious human construction. Products of our unconscious brain or social constructs such as language may be extremely complex for these were not the product of an intentional design process. Thus artists may construct extremely complex artefacts because they do not design every detail of their work but work intuitively a lot of the time with parts and media that are already rich in complexity and meaning.

Apart from human construction there are some circumstances where one has good reason to expect simplicity, namely the initial stages of processes that start with the simplest building blocks and work from there. That is the process is known to be one of elaboration. Examples of these might include the construction of higher elements in the early universe, the reactions of bacteria to external stimuli, or, possibly, the first stages in the evolution of life.

Another situation is where one already *knows* that there is some correct model of some minimum complexity. In this case one heuristic for finding a correct model is to work outwards, searching for increasingly complex models until one comes upon it. There are, of course, other heuristics - the primary reason for starting small are pragmatic; it is far easier and quicker to search through simpler models. In more common situations it might be the case that increasingly complex models may approximate the correct model increasingly, but never completely, well or that no model (however complex) does better than a certain extent. In the first case one is forced into some trade-off between accuracy and convenience. In the second case maybe no model is acceptable, and it is the whole family of models that needs to be changed.

In such circumstances as those above there is some reason to err towards simplicity. However in these circumstance the principle is reducible to a straight forward application of our knowledge about the phenomena that leads us in that direction - principles of simplicity do not give us any 'extra' guidance. In these circumstances instead of invoking simplicity as a justification the reason for the expectation can be made explicit. Simplicity as a justification is redundant here.

Versions of "Simplicity"

In order to justify the selection of theories on the basis of simplicity, philosophers have produced many accounts of what simplicity *is*. These have included almost every possible non-evidential advantage a theory might have, including: number of parameters (Draper 1981), extensional plurality (Goodman 1966, Kemeny 1953), falsifiability (Popper 1968), likelihood (Rosenkranz, 1976 Quine 1968), stability (Turney, P 1990), logical expressive power (Osherton and Weinstein 1990) and content (Good 1969).

In some cases this has almost come full circle. Sober (1975) characterises simplicity *as* informativeness - so that instead of asking whether simplicity is informative he seeks to show that simplicity (as informativeness w.r.t. a specified question) is, in fact, simple.

If, as I have argued, simplicity is *not* truth-indicative, this whole enterprise can be abandoned and the misleading label of 'simplicity' removed from these other properties. This mislabelling, far from producing insight has produced a fog of differing 'simplicities' and 'complexities' which do much to hinder our understanding of the modelling process. Theories can posses a lot of *different* advantages that are not directly linked to its success at explaining or predicting the evidence, restoring the correct labels for these advantages will help (rather than hinder) their elucidation.

An Example - Curve Fitting by parameterisation

A particular case of hypothesis selection that has been discussed in the literature is curve fitting. This is simply a case of deciding which of a variety of hypotheses (in different functional forms) one will select given a set of data (in the form of points). Typically these forms include parameters that are adjusted to fit the data, so that each form corresponds to a family of curves. Curve fitting can be a misleading example as it can be difficult to rid oneself of one’s intuitions about what sort of curves are useful to posit in the case one has personally come across. One can have strong visual intuitions about the suitability of certain choices which strongly relate to a set of heuristics that are effective in the domains one happens to have experienced.

In particular, one might happen to know that there is likely to be some noise in the data, so that choosing a curve that goes through every data point is not likely to result in a line that reflects the case when more data is added. In this case one might choose a smoother curve, and a traditional method of smoothing is choosing a polynomial of a lower order or with fewer parameters. This is not, of course, the only choice for smoothing one might instead use, for example, local regression (Cleveland et al. 1988) where the fitted curve is a smoothed combination of lines to fit segments of the data. Thus the choice of a curve with a simpler functional form depends on: *firstly*, that one has knowledge about the nature of the noise in the data and, *secondly*, that one chooses the simplicity of the functional form as one’s method of smoothing. If, on the other hand, one knew that there was likely to be a sinusoid addition to the underlying data one might seek for such regularities and separate this out. Here a preference for simplicity is merely an expression of a search bias which encodes one's domain knowledge of the situation.

A recent series of papers (Forster and Sober 1994, Forster 1999) argues that simplicity is justified on the grounds that its use can result in greater predictive accuracy on unseen data. This is based on results obtained in (Akaike 1973). Simplicity in this case is defined as (effectively) the Vapnik-Chervonenkis (VC) dimension (Vapnik and Chervonenkis 1981) of the set of curves which in some circumstances is equivalent to the number of adjustable parameters in the equation form. The advantages of 'simplicity' in this account amount to the prescription not to try and fit more parameters that you have data for, since the larger the set of hypotheses one is selecting from the more likely one is to select a bad hypothesis that 'fits' the known data purely by chance. The extent of this overfitting can sometimes be estimated. If you have two models whose predictive accuracy, *once adjusted for its expected overfitting*, is equal then there would be no reason to choose the family which might be considered simpler to have a simpler form. In circumstances with a fixed amount of data the estimation of the extent of overfitting might or might not tip the scales to lead one to select the simpler model.

This account gives no support for a thesis that the simplicity of a model gives any indication as to its underlying model bias. In circumstances where one can always collect more data, so that effectively there is an indefinite amount of data, these arguments provide no reason to select a simpler model. In this case, the decision of when to stop seeking for a model which gives increased predictive accuracy is a pragmatic one: one has to balance the cost of collecting the additional data and using it to search for the most appropriate model against the utility of the parameterised model.

Also the connection between the VC dimension and any recognisable characteristic of simplicity in the family of curves is contingent and tenuous. In the special case where the only way of restricting the VC dimension (or in finite cases, number of hypotheses) is through the number of adjustable parameters, then it is the case that an equational form with more adjustable parameters will require more data for accurate parameterisation. However there are other ways of restricting the set of hypotheses; as discussed above (Webb 1996) successfully uses a similarity criterion. Thus one can avoid overfitting by restricting the VC dimension of the set of hypotheses without using any criteria of simplicity or parsimony of adjustable parameters. Of course, one can decide to define simplicity *as* the VC dimension, but then one would need to justify this transferred epithet.

To summarise this section, there *is* a limit to the accuracy with which one can adjust a certain number of parameters given a certain amount data - one is only justified in specifying in a curve to the extent that one has information to do so. Information in terms of a tightly parameterised curve has to come from somewhere. However, in the broader picture where different families of curves are being investigated (by competing teams of scientists continually searching out more data) as to which explains or predicts the data better, these considerations give no support to the contention that the simpler family has an advantage.

Concluding plea

It should be clear from the above that, if I am right, model selection 'for the sake of simplicity' is either: simply laziness; is really due to pragmatic reasons such as cost or the limitations of the modeller; or is really a relabelling of more sound reasons due to special circumstances or limited data. Thus appeals to it should be recognised as either spurious, dishonest or unclear and hence be abandoned.

However, there is a form of Occam's Razor which represents sound advice as well as perhaps being closer to its Occam's original formulation (usually rendered as "*entities should not be multiplied beyond necessity*"), namely: that the elaboration of theory in order to fit a known set of data should be resisted, i.e. that the lack of success of a theory should lead to a more thorough and deeper analysis than we are usually inclined to perform. It is notable that this is a hallmark of genius and perhaps the reason for the success of genius - be strict about theory selection and don't stop looking until it *really* works.

References

Akaike, H. Information Theory and an Extension of the Maximum Likelihood Principle, in Petrov, B. N. and Csaki, F. (eds.) *2nd International Symposium on Information theory*, 267-281. Budapest: Akademai Kiado, 1973.

Bak, P. *How Nature Works: The Science of Self Organized Criticality*. Oxford, Oxford University Press, 1997.

Charter, N. The Search for Simplicity: A Fundamental Cognitive Principle? *The Quarterly Journal of Experimental Psychology*, 1999, **52A**: 273-302.

Cleveland W. S., Devlin S. J., Grosse E. Regression By Local Fitting - Methods, Properties, And Computational Algorithms. *Journal Of Econometrics*, 1988, **37**: 87-114.

Domingos, P. Beyond Occam's Razor: Process-Oriented Evaluation. Machine Learning: ECML 2000, 11th European Conference on Machine Learning, Barcelona, Catalonia, Spain, May 31 - June 2, 2000, Proceedings, *Lecture Notes in Artificial Intelligence*, **1810**, 2000.

Draper, N. R.; Smith, H. *Applied Regression Analysis*. New York: John Wiley, 1981.

Forster, M. Model Selection in Science: The Problem of Language Invariance. *British Journal for the Philosophy of Science*, 1999, **50**, 83-102.

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

Good, I. J. Corroboration, Explanation, Evolving Probability and a Sharpened Razor. *British Journal for the Philosophy of Science*, 1969, **19**, 123-43.

Goodman, N. *The Structure of Appearance*. Indiapolis: Bobbs-Merrill, 1966.

Kemeny, J. G. Two Measures of Complexity. T*he Journal of Philosophy*, 1953, **52**, 722-733.

Li, M. and Vitányi, P. M. B. Philosophical Issues in Kolmogorov Complexity, in Automata, Languages and Programming, 19th International Colloquium, *Lecture Notes in Computer Science*, **623**, 1-15, Springer-Verlag, 13-17 July 1992.

Mcshea, D. Meatzoan Complexity and Evolution: is there a trend? *Evolution*, 1996, **50**, 477-492

Moss, Scott, Artis, M. and Ormerod, P., A Smart Macroeconomic Forecasting System, *The Journal of Forecasting* **13**, 299-312, 1994.

Murphy, P. M. An empirical analysis of the benefit of decision tree size biases as a function of concept distribution. Technical report 95-29, Department of Information and Computer Science, Irvine, 1995.

Murphy, P.M.; Pazzani, M.J. Exploring the Decision Forest: an empirical investigation of Occam's razor in decision tree induction, *Journal of Artificial Intelligence Research*, 1994, **1**, 257-275.

Osherson, D.N. and Weinstein, S. On Advancing Simple Hypothesis. *Philosophy of Science*. 1990, **57**, 266-277.

Pearl, J. On the Connection Between the Complexity and Credibility of Inferred Models, *International Journal of General Systems*, 1978, **4**, 255-264

Popper, K. R. *Logic of Scientific Discovery*. London: Hutchinson, 1968.

Quine, W. V. O. Simple Theories of a Complex World. In *The Ways of Paradox*. New York: Random House, 1960, 242-246.

Rosenkrantz, R. D. *Inference, Method and Decision*. Boston: Reidel, 1976.

Schaffer, C. 1994. A conservation law for generalization performance. In *Proceedings of the 11th International conference on Machine Learning*, 259-265. New Brunswick, NJ: Morgan Kaufmann.

Sober, E. *Simplicity*. Oxford: Clarendon Press, 1975.

Turney, P. The Curve Fitting Problem: A Solution. *British Journal for the Philosophy of Science*. 1990, **41**, 509-530.

Vapnik V. N. and Chervonenkis A. Y. , Necessary And Sufficient Conditions For The Uniform-Convergence Of Means To Their Expectations, *Theory Of Probability Applications*, **26**, 532-553, 1981.

Webb, G. I. Further Evidence against the Utility of Occam's Razor. *Journal of Artificial Intelligence Research*, 1996, **4**, 397-417.

Wolpert, D. The lack of a priori distinctions between learning algorithms. *Neural Computation*, 1996, **8**, 1341-1390.