Simplicity and Truth Conduciveness

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

Abstract

There is a long-standing puzzle concerning the connection of simplicity to truth in scientific inference. It is proposed that simplicity does not point at or indicate the truth but nonetheless keeps science on the straightest or most direct route thereto. A theorem to that effect is presented, along with a fairly general definition of empirical simplicity, a discussion of examples, and prospects for a new understanding of statistical model selection.

1. The Puzzle

When faced with a choice among alternative theories compatible with current experience, scientists tend to side with the simplest one, where simplicity has something to do with minimizing independent entities, principles, causes, or equational coefficients. Philosophers of science, statististicians, and, more recently, computer scientists, all recommend such an approach, called Ockham's razor, after the fourteenth century theologian and logician William of Ockham. But the practice raises an awkward question—a question that cuts to the very heart of scientific method and, therefore, of scientific education and outreach more generally: insofar as science is about finding true theories rather than aesthetic, useful, or comforting fictions, how *could* Ockham's razor help one find the true theory? For if it is already known that the true theory is simple, science does not require Ockham's assistance. And if it is not known that the true theory is simple, what entitles one to assume that it is?

There are many standard responses, but no satisfactory ones. It does not help to say that simple theories are better "confirmed" (Carnap 1950, Glymour 1980), more severely tested (Popper 1968), more explanatory (Harman 1965, Kitcher 1981), more unified (Friedman 1983), more symmetrical (Malament 1977), or

more compact (Rissanen 1983), since if the truth is not simple, then it does not possess these nice properties either (van Fraassen 1981).

Nor does it help to observe that a Bayesian agent whose prior probabilities are biased toward simple possibilities will judge the simple possibilities to be more probable—a Bayesian with the contrary bias would disagree and the question is why the former bias is better at finding the truth than the latter. The Bayesian argument remains circular even if complex and simple theories receive equal prior probabilities (e.g., Rosenkrantz 1983 and Schwarz 1978), for theories with more free parameters can be true in more "ways", so each way the complex theory might be true ends up carrying less prior probability than each of the ways the simple theory might be true, and that prior bias toward simple possibilities is merely passed through Bayes' theorem.

Nor does it help, so far as finding the true theory is concerned, to say that using a simple theory for purposes of predictive estimation can reduce the expected error of the resulting estimates, for that argument stands even if it is known in advance that the simple theory is false (Akaike 1973, Forster and Sober 1994). Furthermore, if one is interested in predicting the causal outcome of a policy on the basis of non-experimental data, the prediction could end up far from the mark because the counterfactual distribution after the policy is enacted may be quite different from the distribution sampled (Zhang and Spirtes 2003). Also, such arguments work only in statistical settings, but Ockham's razor seems no less compelling in deterministic ones.

Nor does it help to remark that a scientist who always favors the simplest theory will converge, eventually, to the true one (Reichanbach 1949, Sklar 1974). Scientists starting with other biases would do so as well (Salmon 1967).

Finally, appeals to Providence (Leibniz 1714), to Kant's synthetic a priori (Whewell 1840) or to unobserved evolutionary etiologies (Giere 1985, Duda et al. 2000, pp. 464-465) to argue that simplicity is correlated with truth regardless of the topic of inquiry are less plausible than the particular applications of Ockham's razor that they are invoked to justify.

In sum, the standard literature on Ockham's razor contains no plausible, non-circular, explanation of how simplicity helps science to arrive at true theories. Indeed, a recent philosophical monograph on simplicity pessimistically concludes:

> ...no single account of theory unification can be given. A philosophical consequence of that claim is that unity should not be linked to truth or increased likelihood of truth (Morrison 2000, p. 232).

This skeptical challenge to the most powerful and characteristic principle of scientific inference is not idle, as the preceding discussion of standard explanations illustrates. It would be better, therefore, to have a clear, relevant, sound, noncircular, and readily intelligible argument to the effect that Ockham's razor is the most efficient possible method for finding the true theory when the problem involves theory choice. This note presents just such an explanation.[1]

2. The Most Direct Route to the Truth

Here is the basic idea. Nature presents new empirical effects to the scientist at times of her own choosing. Assume that the correct answer to the scientist's question is uniquely determined by these effects (otherwise, even convergence to the true answer in the limit is hopeless). Even given that assumption, which characterizes many everyday applications of Ockham's razor , there is no guarantee that the true answer can be infallibly or even reliably inferred, for a crucial effect might be so subtle or small or implausible as to elude detection until later. At least one can say that the Ockham strategy of choosing only the unique answer corresponding to the set of currently observed effects is guaranteed to converge to the true answer in the limit (after the overly-simple theories are refuted by effects). But many alternative strategies also converge to the truth in this problem (every finite variant of a convergent strategy is also convergent). So the simplicity puzzle stands: in what sense does the Ockham strategy lead one to the truth better than alternative strategies?

Not in the sense that Ockham's razor is guaranteed to point at or indicate the truth in the short run, for such a guarantee would imply that one already knows that the truth is simple. Nor in the sense that Ockham's razor is guaranteed to converge to the truth in the limit, for alternative methods that differ sharply from Ockham's razor in the short run can claim the same. The new idea is that these are not exhaustive alternatives, for it may be that Ockham's razor somehow converges to the truth along the straightest or most direct path, where directness is, roughly speaking, a matter of not altering one's opinion more often or later than necessary. Since methods that approach the truth more directly have a superior *connection* to the truth or are more *conducive* to finding the truth, it

[1]The approach is based on concepts from computational learning theory. For a survey of related ideas, cf. (Jain et al., 1999) and (Kelly 1996). Earlier versions of the following argument may be found in (Schulte 1999, Kelly 2002, 2004, 2006a, 2006b, and Kelly and Glymour 2004). A version of the idea was also presented as a tutorial at the 2005 FEW conference at the University of Texas, Austin. The current version of the theory responds to an interesting objection to to S. Sarkar at the conference and is much more generally applicable.

is a relevant and non-circular response to the simplicity puzzle to prove that Ockham strategies approach the truth more directly than all competitors.

Consider the most aggressive Ockham strategy that always chooses the answer corresponding to the effects observed so far. This strategy retracts exactly when each new effect is presented. So, however the data for a given answer are presented, the obvious strategy's worst-case bound on retractions is the number of effects the answer corresponds to. Now consider an arbitrary, alternative, convergent strategy M. Strategy M might retract less than the obvious Ockham strategy on some presentations of the data for an answer (e.g., M might guess that ten more effects are coming and may then be lucky enough to observe them all in sequence, skipping all the intervening retractions the Ockham strategy would perform). But in terms of worst-case bounds on retractions over all the ways the data for a given answer could have been presented by nature, the Ockham strategy does as well, for suppose that the answer T in question corresponds to n effects. Nature can present no effects to M until M converges to the simplest answer, for if M never converges to T, nature never presents an effect, so the simplest theory is the correct answer and M fails to converge to it. Then nature can present an effect compatible with T, followed by no further effects, until M converges to the answer corresponding to one effect, and so forth, up to answer T, forcing M to retract ten times, arbitrarily late. It follows that the Ockham strategy is *efficient*, in the sense that its worst-case guarantee concerning retractions and retraction times over a given answer is as good as an arbitrary, convergent strategy's.

Furthermore, violating Ockham's razor results in a weaker guarantee. Suppose that M adopts an answer that posits unseen effects (a violation of Ockham's razor). Then nature is free to withhold the anticipated effects until, on pain of not converging to the true answer, M backtracks to the (simpler) theory corresponding to just the effects presented thus far. Then nature can proceed to present more effects, leading M through each successive theory, as in the preceding argument. So nature can force M to perform an extra reversal of opinion prior to all of the retractions the Ockham strategy would have performed. So Ockham strategies are not merely efficient; they are *uniquely* efficient, in the sense under discussion.

3. Illustration: Accounting for Empirical Effects

It remains to explain how the preceding story is intended to apply to concrete scientific examples. Suppose that one is interested in the structure S of an unknown

polynomial law

$$f(x) = \sum_{i \in S} a_i x^i,$$

where S is assumed to be a finite set of indices such that for each $i \in S$, $a_i \neq 0$. It seems that structures involving fewer monomial terms are simpler, so Ockham's razor favors them. Suppose that patience and improvements in measurement technology allow one to obtain ever tighter open intervals around $f(x)$ for each specified value of x as time progresses.[2] Suppose that the true degree is zero, so that f is a constant function. Each finite collection of open intervals around values of f is compatible with degree one (linearity), since there is always a bit of wiggle room within finitely many open intervals to tilt the line. So suppose that the truth is the tilted line that fits the data received so far (i.e., suppose that the mother moves to the next shop). Eventually you can obtain data from this line that refutes degree zero. Call such data a (first-order) effect. Any further, finite amount of data collected for the linear theory is compatible (due to the remaining, minute wiggle room) with a quadratic law, etc. The truth is assumed to be polynomial, so the story must end, eventually, at some finite set S of effects. Thus, determining the true polynomial law amounts, essentially, to determining the finite set S of all monomial effects that one will ever see and Ockham's razor amounts to never assuming more effects than one has seen so far.[3]

The Copernican revolution in observational astronomy can be viewed in a similar manner. Each planet might revolve around the sun or around the earth and the sun might revolve around the earth or vice versa. Revolution of a planet around the sun implies that retrograde motion (the appearance of backtracking against the fixed stars) happens precisely at opposition for superior planets or at conjunction for inferior planets (Kuhn 1957, Glymour 1980). The earth-centered account of a planet can be adjusted just so to produce the same appearance, so strictly speaking, it is impossible to distinguish the two accounts. To make the question interesting, excuse failure to find the truth in that case because it is hopeless to do so.[4] Then if retrograde motion is out of sync with solar opposition,

[2]In statistics, the situation is analogous: increasing the sample size reduces the interval estimates of the values of the function at each argument. The analogy is sketched in greater detail in the conclusion.

[3]In typical statistical applications, something similar is true: effects probably do not appear at each sample size if they don't exist and probably appear at some sample size onward if they do exist. The data model under discussion may be viewed as a logical approximation of the statistical situation, if one thinks of samples accumulating through time.

[4]Although that may seem to assume away the traditional problem of scientific realism, it

it will eventually be observed to be so (assuming the increasingly precise, inexact measurements assumed in the preceding example). Hence, the geocentric account of each planet is associated with the eventual effect of viewing retrograde motion out of sync with opposition or conjunction, and the heliocentric account of earth is associated with the effect of stellar parallax—the apparent motion of the stars due to the earth's motion. Hence, the uniquely simplest theory on purely positional data was Tycho Brahe's, in which all planets revolve around the sun, which revolves around the earth. Indeed, Brahe was keen to emphasize this fact and the failure to detect parallax was understood to be an embarrassment for the Copernican theory, although the mechanical inscrutability of Brahe's model seems to have restricted its popularity in spite of its extra simplicity (the aim is to explain our systematic bias toward simplicity, not to prove that it must carry the field against every plausibility consideration). Copernicus' heliocentric theory implies a parallax effect, so it is one effect more complex that Brahe's. Each geocentric planet adds a further effect. Hence, the ancient Pythagorean hypothesis that Mercury and Venus are the only heliocentric planets is only one effect more complex than Copernicus' theory. Ptolemy's geocentric account of every planet was very complex, but not as complex, at least, as the very awkward theory that moves the earth around the sun but all the planets around the earth, which has the dubious distinction of being as complex as possible *and* inviting all the physical conundra of a moving earth.

The recent literature on the inference of causal structure (Spirtes et al. 2000) provides a more contemporary illustration. Stripping the idea down to its bare essentials, the true causal structure over a set of variables is conceived as a network, in which arrows correspond to immediate causal influences. Formal rules relate such networks to joint probability measures over the variables in such a way that the correspondence between probabilities and causal structures depends only on which conditional statistical independence relations hold among variables. Independence cannot be detected as such: if variables are independent one simply never detects a dependency. Dependence, on the other hand, can be detected: if variables are dependent then an independence test will reject the null hypothesis of independence at a sufficiently high sample size. Since the number of detected dependencies increases through time (as more data are collected), the causal con-

does not do so, for even if failure to do the impossible is excused, there is still the problem of determining what to say when the *current* data are compatible with either hypothesis—and that is a matter of some importance, since there are practical future consequences for being wrong (Churchland 1982).

clusions based upon the underlying correspondence between graphs and probability distributions changes through time. Indeed, direct causal connections between variables correspond to more conditional dependencies, so conditional dependencies may be viewed as empirical effects and the intuitive complexity of potential answers to the causal inference question increases as the implied number of effects increases (cf. Kelly 2006a, 2006b).

A different sort of example concerns the determination of quantum conservation laws from a set of reactions, in which case empirical effects correspond to the discovery of reactions linearly independent of the reactions observed so far (Schulte 2000, Kelly 2006b). An interesting feature of this example is that simpler conservation theories involve more conserved quantities (symmetries), which correspond to fewer dimensions of linear independence among the reactions. Hence, the simpler hypotheses require more notation to state (the magnitude of each conserved quantity must be specified for each particle type) which reverses, in this case, the popular notion that simplicity has something to do with description length.

To reduce the preceding examples to their essential elements, let E be a denumerable set of *potential effects* and let Γ be a collection of finite sets of effects, any one of which might be the actual effects that will be observed for eternity. Nature is free to present effects from E at any time she chooses, possibly several at one time, as long as the set S of all effects presented for eternity is a member of Γ. She is never required to grant any assurance that an unobserved effect will never appear, however—that is the fundamental epistemological asymmetry upon which everything that follows depends. In the limit, nature presents an infinite *input stream* or *empirical world* w such that for each i, $w(i)$ is a finite subset of E corresponding to the effects presented at stage i. Let S_w denote the total, finite set of effects presented along w. It is assumed that each theoretical structure T_S corresponds uniquely to a finite set of effects S. Then the correct structure for world w is just T_{S_w}. Call the situation just described the *effect accounting problem* generated by Γ.

An empirical *strategy* M gets to observe, at each stage n, the initial segment $w|n$ of the actual world w presented by stage n. Strategy M responds either with some possible theoretical structure T_S or with '?', indicating a refusal to choose. Strategy M is a *convergent solution* if and only if for each possible world w, $\lim_{n\to\infty} M(w|i) = T_{S_w}$.

4. Empirical Simplicity and Empirical Effects

Empirical simplicity has been a vexed question in philosophy for many years (Goodman 1983), but often because simplicity has been conceived as an absolute property independent of a particular problem or question (e.g., Li and Vitanyi 1997). The approach adopted here less ambitiously locates simplicity in the structure of the question asked—just as computer science locates computational complexity in the problem to be solved. Let e be a finite input sequence (think of e as what has been seen so far by the scientist). Let Γ_e denote the restriction of Γ to sets of effects compatible with e, meaning that each $S \in \Gamma_e$ includes S_e. A *directed path* to S in Γ_e is a finite sequence $(S_0 \subset S_1 \subset \ldots S)$ of elements of Γ_e. Then let the *conditional complexity* $c(T_S, e)$ of theoretical structure T_S given e be the result of subtracting 1 from the length of the longest path to S in Γ_e. If Γ includes every finite subset of E, then the definition plausibly reduces to counting the effects in S that are not already in S_e: i.e., $c(T_S, e) = |S| - |S_e|$. It follows, in that special case, that the unique theoretical structure of complexity zero given e is T_{S_e}. If Γ has "gaps" because some finite sets of effects are ruled out by background information, then there may be more than one simplest structure compatible with experience and it may be that $c(T_S, e) < |S| - |S_e|$. A weaker assumption, compatible with such gaps, is that every maximal path have the same (finite or infinite) length. Say that such a problem is *uniform*.

Ockham's razor is the principle that one should not output T_S in response to e unless T_S is the unique theoretical structure compatible with e for which $c(T_S, e) = 0$. A closely related principle is *stalwartness*, which requires that M never drop an informative answer T_S unless it is no longer uniquely simplest. The intuition behind stalwartness is that there is no better explanation than the simplest one, so why drop it? One may speak of stalwartness and/or Ockham's razor as being satisfied from e onward (i.e., at each extension e' of e) or always (i.e., at each e).

5. The Straightest Path to the Truth

In accordance with the explanation sketched above, let the *loss* of convergent strategy M in world w be represented by the sequence $\lambda(M, w) = (r_1, \ldots, r_k)$ of successive times at which M retracts an informative answer in w. The only loss comparisons that matter for the following argument are the easy comparisons in which strategy M retracts as often and at least as late as strategy M' in world w, in which case it is clear that the performance of M is at least as bad as that of M' in w. For example, $(2, 3, 7) < (3, 5, 8, 12)$ but $(2, 5)$ is incomparable with

$(1, 6)$.[5]

A *potential retraction time bound* is like a retraction time sequence except that the infinite number ω may occur. Then each set X of retraction time sequences has a unique, least upper bound $\sup(X)$ among the potential retraction time bounds (cf. Kelly 2006a).[6] Now one may speak, nontrivially, of the *worst-case* timed retractions of strategy M over some collection K of worlds.

Ockham strategies do not do better than alternative strategies in every world, because a strategy that anticipates an effect before it is observed might get lucky and see the anticipated effect at the very next stage. Nor do Ockham strategies do better than alternative strategies in the worst case overall, since there is no finite bound either on a convergent strategy's retractions or on the times at which the successive retractions occur. Nor do Ockham strategies do better than alternative strategies in the expected case unless a prior probability distribution is imposed according to which complex worlds are less probable than simple worlds, which begs the question in favor of Ockham strategies (Kelly 2006a). But the unique advantage of Ockham strategies emerges if one considers worst-case costs over worlds of a given empirical complexity. Accordingly, let $C_e(n)$ denote the set of all worlds w compatible with e such that $c(w, e) = n$. Let M be an arbitrary solution to the effect accounting problem. Define the *worst-case loss* of solution M over complexity class $C_e(n)$ as: $\lambda_e(M, n) = \sup_{w \in C_e(n)} \lambda(M, w)$, where the supremum is understood in the sense of the preceding paragraph.

Suppose that input sequence e has just been received and the question concerns the efficiency of one's strategy M. Since the past cannot be altered, the only relevant alternatives are strategies that produce the same answers as M along e_-, where e_- denotes the result of deleting the last entry from e. Say that such a strategy *agrees with* M along e_- (abbreviated $M \equiv_{e_-} M'$).

Given solutions M, M', the following, natural, worst-case performance comparisons can be defined at e:

$$M \leq_e M' \quad \text{iff} \quad (\forall n)\ \lambda_e(M, n) \leq \lambda_e(M', n);$$
$$M <_e M' \quad \text{iff} \quad M \leq_e M' \text{ and } M' \not\leq_e M.$$

These comparisons give rise to two natural properties of strategies:

$$M \text{ is } \textit{beaten} \text{ at } e \quad \text{iff} \quad (\exists \text{ solution } M' \equiv_{e_-} M)\ M' <_e M;$$

[5]Formally, $\sigma \leq \tau$ if and only if there exists a 1-1 mapping g from positions in σ to positions in τ such that for each position i in σ, $\sigma(i) \leq \tau(g(i))$.

[6]A slight technicality: for potential bounds $\mathbf{b}, \mathbf{b}'$, define $\mathbf{b} \leq \mathbf{b}'$ if and only if each retraction time sequence $\leq \mathbf{b}$ is also $\leq \mathbf{b}'$.

$$M \text{ is } \textit{efficient} \text{ at } e \quad \text{iff} \quad (\forall \text{ solution } M' \equiv_{e\_} M)\ M' \geq_e M.$$

A convergent solution is beaten by another if the latter solution does as well in each complexity class and better in some. An efficient solution is as good as an arbitrary solution in each complexity class. Since efficiency can be reassessed at each time, one may speak of being efficient from e onward or always.

6. Unique Efficiency Theorem

A precise argument along the lines sketched above (Kelly 2006b) yields the following, mathematical theorem:

Theorem 1 (Ockham efficiency characterization) *Let M solve the effect accounting problem generated by Γ and suppose that the problem is uniform. Let e be a finite input sequence compatible with Γ. Then, the following statements are equivalent:*

1. *M is stalwart and Ockham from e onward;*
2. *M is efficient from e onward;*
3. *M is never beaten from e onward.*

So the set of all convergent solutions to the effect accounting problem is cleanly partitioned at e into two groups: the solutions that are stalwart, Ockham, and efficient from e onward and the solutions that are beaten at some stage $e' \geq e$ due to future violations of the stalwart, Ockham property. The argument is *a priori*, normative, truth-directed, and yet non-circular. The argument presumes no prior bias of any kind, so there is no question of a circular appeal to a simplicity bias, as in Bayesian arguments. The argument is driven only by efficient convergence to the truth, so there is no bait-and-switch from truth-finding to some other aim. There is no confusion between "confirmation" and truth-finding, since the concept of confirmation is never mentioned. There is no wishful presumption that the truth must be testable or nice in any other way. There is no appeal to the hidden hands of Providence or Evolution. The same cannot be said of any alternative explanation on the books today.

Furthermore, the argument is *diachronically stable* in the sense that it always makes sense to return to the Ockham fold no matter how many times one has violated Ockham's razor in the past. That is important, for Ockham violations are practically unavoidable in real science because the simplest theory cannot

always be formulated in time to forestall acceptance of a more easily conceived but more complex alternative (e.g., Ptolemaic astronomy *vs.* Copernican astronomy, Newtonian optics *vs.* wave optics, Newtonian kinematics *vs.* relativistic kinematics, and special creation *vs.* natural selection). So although it has been urged that scientific revolutions are extra-rational events governed only by the vagaries of scientific politics (Kuhn 1975), revision to the simpler theory when it is discovered has a clean explanation in terms of truth-finding efficiency. Stability fails for some non-uniform problems, but it is still the case that efficiency at every stage is equivalent to being a normal Ockham at every stage.

6. A General Definition of Simplicity

The preceding approach would be far deeper and more interesting if empirical complexity were defined directly in terms of the structure of an arbitrary empirical problem, rather than being presupposed and spoon-fed to the scientist. Here is a very general such an account.

Let an *empirical problem* be a pair $\mathcal{P} = (K, \Pi)$, where K is a set of infinite input sequences or *empirical worlds* and Π is a partition of K into *possible answers*. Here, there is no question of "pre-packaging" what counts as an empirical effect: the successive entries in infinite sequence $w \in K$ might be boolean bits in a highly "gruified" coding scheme with an ocean of irrelevant information added. If e is a finite input sequence, let K_e denote the set of all w in K that extend e and if w is in K.

The first step is to construct an analogue Γ'_e of the set of possible sets of effects Γ_e compatible with e entirely out of the branching structure of $\mathcal{P}$. Let p be a finite sequence of answers drawn from Π. Say that p is *forcible* by nature given finite input sequence e in $\mathcal{P}$ if and only if for each strategy M guaranteed to converge to the true answer in $\mathcal{P}$, there exists w in K that extends e such that M responds to w after the end of e with a sequence of outputs of which p is a subsequence. Let S_e denote the set of all finite sequences of answers forcible in $\mathcal{P}$ given e and refer to S_e as the *forcibility state* of $\mathcal{P}$ at e. Now let Γ'_e be the set of all forcibility states $S_{e'}$ such that e' extends e.

The next step is to recover paths through Γ'_e analogous to the inclusion paths through Γ_e. If S_0, S_1 are in Γ'_e, say that S_1 is *epistemically accessible* from S_0 given e (written $S_0 \leq_e S_1$) just in case there exists evidence e_0 extending e such that $S_0 = S_{e_0}$ and further evidence e_1 extending e_0 such that $S_1 = S_{e_1}$. Now let $\pi_e(S)$ denote the set of $\leq_e$-paths in Γ'_e that terminate with S.

It remains to associate answers in Π with elements of Γ'_e, a correspondence

merely stipulated in the effect accounting problem. Say that answer T in Π *caps* S in Γ'_e if and only if concatenating T onto an arbitrary answer sequence in S results in an answer sequence in S. Except in pathological cases (excluding all effect accounting problems), each forcibility state in Γ'_e is capped by at most one answer, so restrict attention to such *univocal* problems. Then let T_S denote the unique answer that caps S if there is an answer that caps S and let T_S be undefined otherwise.

It is now possible to define empirical complexity. If p is in $\pi_e(S)$, let $c(p)$ denote the number of stages $i > 0$ along p such that $T_{p(i-1)}$ is defined and $T_{p(i)}$ is undefined or $T_{p(i)} \neq T_{p(i-1)}$. Then define the empirical complexity of forcibility state S given e as:

$$c(S, e) = \sup\{c(p) : p \in \pi_e(S)\}.$$

The final step is to associate forcibility states with worlds. In the effect accounting problem it is assumed that each world presents a finite set of effects, so the set of effects presented eventually stops growing. The parallel assumption in this more general construction is that $\lim_{i\to\infty} S_{w|i}$ converges in each $w \in K$. Given this assumption, define $S_w = \lim_{i\to\infty} S_{w|i}$ and, finally, let:

$$\begin{aligned} c(w, e) &= c(S_w, e); \\ c(T, e) &= \min\{c(w, e) : w \in T \cap K_e\}. \end{aligned}$$

The account of empirical complexity just presented agrees with the account assumed in the effect accounting problem (Kelly 2006b), so there is some set of assumptions under which it supports the unique efficiency theorem for Ockham's razor. It is an interesting question how broadly the efficiency result applies. Furthermore, if problem $\mathcal{P}$ is constructed by assuming a set of effects Γ and presenting them in an arbitrarily gerrymandered coding system, the above definition will recover the same empirical complexity assessments for answers that would be recovered from the naive version of the problem in which effects are directly presented by nature. Finally, in the proposed construction, empirical complexity is objectively grounded in the structure (K, Π) of the problem and is, therefore, invariant under notational changes, including the notorious "grue-like" translations of N. Goodman (1983).

7. Toward Statistical Model Selection

The preceding reasoning can be extended to random strategies in the following, straightforward way (Kelly 2006a). Answer T is *retracted in chance* by random

strategy M to degree r at stage $n+1$ if and only if the chance that M produces T drops by r from stage n to $n+1$. The *total retractions in chance* by M in world w are given by the sum over all answers $T \in \Pi$ and all stages $n > 0$ of the degree of retraction in chance by M of T at n. Then, assuming that M converges in probability to the true answer (i.e., that the chance that the strategy produces the true answer goes to unity as more data are seen), one can argue that the strategy can be forced by nature into total retractions in chance arbitrarily close to those of a deterministic strategy (Kelly and Glymour 2004).

It is less straightforward to apply the preceding approach to genuine, statistical model selection, in which theoretical structures are statistical models and the data are produced randomly according to the true model under some setting of its free parameters—but that must be done if the proposed approach is to enjoy real scientific application. Notoriously, in statistical model selection there is no such thing as "compatibility with the data", so Ockham's razor cannot be expressed as a matter of minimizing complexity over answers compatible with the data—instead, Ockham's razor involves some motivated compromise between simplicity and fit. There are theories of how best to strike the balance, but they are either circular or directed at prediction rather than finding the true model, as discussed in the introduction. To solve for objective constraints on the optimum balance between simplicity and fit from the aim of efficient convergence to the truth would, therefore, constitute a new, truth-directed foundation for statistical model selection and, hence, for scientific inference in general. The theoretical significance and potential for broader methodological impact of such a theory would be immense, both in terms of concrete methodological recommendations in such expanding areas as the inference of causal structure and in terms of providing science with a reasonable, transparent, truth-directed motive for its most powerful rule of inference.

A preliminary approach, which follows the literature on statistical causal modeling, is to set a significance level for statistical tests and to view the outcomes of the tests as the raw input to the theorist's model selection method. According to this approach, the crucial notion of "compatibility with the data" is explicated in terms of non-rejection at the chosen significance level and empirical effects may be understood as rejections of null hypotheses and Ockham's razor demands that one conclude the model that corresponds to exactly the null hypotheses rejected thus far. For example, the algorithms for inferring causal structure presented in (Spirtes et al. 2000) do precisely that. It is proposed to show that presuming against an effect until it is detected is a strategy that minimizes retractions in

chance for arbitrary, convergent methods that take test outcomes as raw inputs.

Ultimately, however, one would prefer not to evade the trade-off between simplicity and fit but to derive it from considerations of efficient convergence to the true statistical model. For example, consider the toy question of how many components of a bivariate normal distribution are zero. The simplest hypothesis is that both components are, the next simplest is that exactly one component is and the most complex is that neither is. A standard model selection technique is to maximize a quantity known as the Bayes Information Criterion or BIC for short (Schwarz 1978). The BIC score for a model $T[\theta]$ with free parameter vector θ of length k relative to sample E of size n is just:

$$BIC(T[\theta], E) = \log(\sup_{\theta} P(E, T[\theta])) - \frac{k \log(n)}{2}.$$

The intriguing thing about the BIC score is that the left-hand-term rewards models that can be "fit" closely to the data (i.e., that make the data very probable) while the right-hand-term penalizes the number k of free parameters in the theory. The official justification for the BIC score is that maximizing it picks out the (approximately) most probable model given the data (according to the simplicity-biased prior probability distribution discussed in the introduction of this summary). The concern here, however, is efficient arrival at the truth, rather than duplication of some simplicity-biased Bayesian's opinions—and that requires an examination of how the model selections of the BIC strategy reverse in chance as sample size increases. The Mathematica plots depicted in figure 1 illustrate the behavior of BIC as the sample size n increases. The (oddly shaped) white

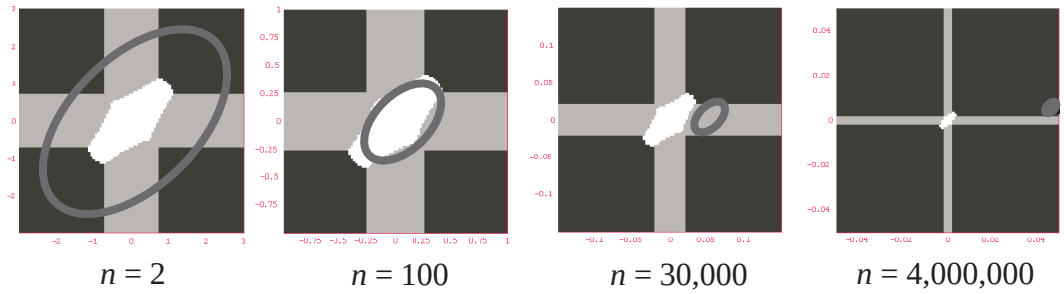

Figure 1: Retractions in chance by BIC

zone covers the points in sample mean space at which BIC selects the simplest answer (that both coordinates of the population mean are zero). The cross-shaped grey zone covers points at which the next simplest answer is selected (that one

coordinate of the population mean is zero) and the black region is where the most complex answer is selected. The unfilled oval line is the boundary of the 95% quantile or footprint of the true sampling distribution, which is (deviously) chosen to have mean vector $(.05, .005)$. The BIC strategy first "takes the bait" for the simplest answer at around $n = 100$ (note that the 95% quantile is nearly contained within the acceptance zone for the simplest answer). BIC "notices" that the first component of the mean vector is nonzero at $n = 30,000$ (the 95% quantile is now nearly contained in the acceptance zone for the next-simplest answer) and then notices that the second is nonzero at around $n = 4,000,000$, for approximately two retractions in chance—the theoretical optimum in this case. But there are also the partial retractions of complex answers between $n = 2$ and $n = 100$, due to the non-negligible acceptance regions for those answers within the 95% quantile at $n = 2$. These extra retractions could have been reduced by making the white zone larger (i.e., by making BIC more aggressively Ockham) at small sample sizes. But doing that would add extra retractions to worlds (on the axes) in which the next simplest hypothesis is true, for in such worlds the theoretically optimal performance is one retraction in chance, but favoring the simplest answer a priori adds to this retraction. Hence, there is pressure both to expand and to reduce the white acceptance zone and therein, it is proposed, lies the essential balance between simpicity and fit. Here is one proposal that appears promising. Say that a *focus point* for n retractions is a point in parameter space such that an arbitrary, convergent strategy produces retractions in chance arbitrarily close to n in arbitrarily small neighborhoods around the point. Choose a standard metric $\rho_{\theta,\theta'}$ (ranging from zero to unity) that reflects the relative distinguishability of probability measures p_θ and $p_{\theta'}$. Say that the *unexcused* retractions of strategy M at possible sampling distribution p_θ are those in excess of the *excused retractions* at θ given by:

$$e(\theta) = \sup_{\theta'} n_{\theta'}(1 - \rho_{\theta,\theta'}),$$

where the supremum ranges over focus points θ' and $n(\theta')$ is the maximum number of retractions for which θ' is a focus point. Now choose acceptance zones to mininize unexcused retractions at every possible parameter value. This theory is a natural extension of the deterministic theory described above, but in this case no fixed notion of consistency with the evidence is presupposed and yet some objective balance between simplicity and fit is implied. Nor is there anything in the account that could be accused of a prior bias in credence toward simple worlds or answers. The focus points presuppose only convergence to the truth

and the metric for defining excused retractions reflects only how distinguishable two sampling distributions are, which has nothing to do about which is simpler. The linchpin of the argument is an invariant feature of converging to the truth in a model selection problem, namely, that the focus points for larger numbers of retractions are simpler worlds.

Solving for the essential balance between simplicity and fit in problems of genuine interest, such as the inference of causal networks (Spirtes et al. 2000) is not trivial analytically, but it remains possible to run computer simulations of model selection techniques on large numbers of samples at increasing sample sizes and to examine the histogram of theory choices at each sample size. Such histograms do not purport to show the impossible: that the method reliably finds the true model no matter how the parameters of the various models are adjusted, but rather, that the methods do or do not approach the theoretically optimum performance of one retraction in chance per free parameter. To the extent that a strategy falls short of the optimum, there is pressure to improve it, as in the preceding example.

References Cited

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

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

Churchland, P. (1982), "The anti-realist epistemology of van Fraassen's The Scientific Image," *Pacific Philosophical Quarterly* 63: 226-235.

Dretske, F. (1981) *Knowledge and the Flow of Information*, Cambridge: M.I.T. Press.

Duda, R., D. Stork, and P. Hart (2000), *Pattern Classification*, 2nd. ed., v. 1. New York: Wiley.

Forster, M. R (2007) "A Philosopher's Guide to Empirical Success", forthcoming, *Philosophy of Science.*

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.

Ford, K. (1963) *The World of Elementary Particles*, New York: Blaisdell.

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

Giere, R. (1985), "Philosophy of Science Naturalized," *Philosophy of Science*, 52: 331-56.

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

Goldman, A. (1986) *Epistemology and Cognition*, Cambridge: Harvard University Press.

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

Harman, G. (1965) The inference to the best explanation, *Philosophical Review* 74: 88-95.

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. (2006a) "Ockham's Razor, Empirical Complexity, and Truth-finding Efficiency," forthcoming, *Theoretical Computer Science*.

Kelly, K. (2006b) "Simplicity, Truth, and Information' forthcoming, *Philsosophy of Information*, J. Van Benthem and P. Adriaans, eds., Dordrecht: Kluwer.

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.

Kitcher, P. (1981) "Explanatory Unification," *Philosophy of Science*, 48: 507-31.

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

Kuhn, T. (1957) *The Copernican Revolution*, Cambridge: Harvard University Press.

Mayo, D. (1996) *Error and the Growth of Experimental Knowledge*, Chicago: University of Chicago Press.

Leibniz, G. W. (1714) *Monadologie*, in *Die Philosophischen Schriften von G. W. Leibniz*, vol. IV. Berlin: C. J. Gerhardt, 1875, 607-23.

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

Malament, D. (1977) "Causal Theories of Time and the Conventionality of Simultaneity," *Nous* 11: 293-300.

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

Morrison, M. (2000) *Unifying Scientific Theories: Physical Concepts and Mathematical Structures*, Cambridge: Cambridge University Press.

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.

Reichenbach, H. (1949) *The Theory of Probability*, London: Cambridge University Press.

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

Robins, J., Scheines, R., Spirtes, P., and Wasserman, L. (1999) "Uniform Consistency in Causal Inference," *Biometrika* 90:491-515.

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.

Schwarz, G. (1978) "Estimating the Dimension of a Model," *The Annals of Statistics*, 6: 461-464.

Silva, R., Scheines, R., Glymour, C. and Spirtes, P. (2006) "Learning the Structure of Linear Latent Variable Models," *Journal of Machine Learning Research* 7: 191-246.

Sklar, L. (1974) *Space, Time, and Spacetime*, Berkeley CA: University of California Press, 1974.

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

Valdez-Perez, R. and Zytkow, J. (1996) "Systematic Generation of Constituent Models of Particle Families, " *Physical Review*, 54: 2102-2110.

Van Benthem, J. (2006) "Epistemic Logic and Epistemology, the state of their affairs," *Philosophical Studies* 128: 49 - 76.

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

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

Wolpert, D. H. and MacReady, W. G. (1997) "No Free Lunch Theorems for Optimization," *IEEE Transactions on Evolutionary Computation* 1: 67-82.

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

Whewell, W. (1840) *The Philosophy of the Inductive Sciences, Founded Upon Their History*, London.

Zhang, J. and Spirtes, P. (2003) "Strong Faithfulness and Uniform Consistency in Causal Inference," in *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence*, 632-639. Morgan Kaufmann.