A New Solution to the Puzzle of Simplicity

Kevin T. Kelly[†]

Explaining the connection, if any, between simplicity and truth is among the deepest problems facing the philosophy of science, statistics, and machine learning. Say that an *efficient* truth finding method minimizes worst case costs en route to converging to the true answer to a theory choice problem. Let the costs considered include the number of times a false answer is selected, the number of times opinion is reversed, and the times at which the reversals occur. It is demonstrated that (1) always choosing the simplest theory compatible with experience, and (2) hanging onto it while it remains simplest, is both necessary and sufficient for efficiency.

q1 **1. The Puzzle of Simplicity.** Philosophy of science, statistics, and machine learning all recommend the selection of simple theories or models on the basis of empirical data, where simplicity has something to do with minimizing independent entities, principles, causes, or equational coefficients. This intuitive preference for simplicity is called Ockham's razor, after the fourteenth century theologian and logician William of Ockham. But in spite of its intuitive appeal, how *could* Ockham's razor help us find the true theory? For, in an updated version of Plato's *Meno* paradox, if we already know that the truth is simple, we don't need Ockham's help. And if we don't already know that the truth is simple, what entitles us to assume that it is?

It does not help to say that simplicity is associated with other virtues such as testability (Popper 1968), unity (Friedman 1983), better explanations (Harman 1965), higher "confirmation" (Carnap 1950; Glymour 1980), or minimum description length (Rissanen 1983), since if the truth were not simple, it would not have these nice properties either. To assume otherwise is to engage in wishful thinking (van Fraassen 1981).

q2 Over-fitting arguments (Akaike 1973; Forster and Sober 1994) show that using a simple model for predictive purposes in the presence of ran-

[†] To contact the author, please write to: Department of Philosophy, Carnegie Mellon University, Baker Hall 135, Pittsburgh, PA 15213-3890; e-mail: kk3n@andrew.cmu.edu.

Philosophy of Science, 74 (December 2007) pp. 000–000. 0031-8248/2007/7405-0002$10.00

dom noise can decrease the expected squared error of predictions. But that is still the case when one knows in advance that the truth is complex, so over-fitting arguments concern accuracy of prediction rather than finding the true theory. 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 (Spirtes and Zhang 2003). Finally, such arguments work only in statistical settings, but Ockham's razor seems no less compelling in deterministic ones.

Nor is Ockham's razor explained by a prior probabilistic bias in favor of simple possibilities, for the propriety of a systematic bias in favor of simplicity is precisely what is at issue. The argument remains circular even if complex and simple theories receive equal prior probabilities, for theories with more free parameters can be true in more 'ways', so that 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; that prior bias toward simple possibilities is merely passed through Bayes' theorem (e.g., Rosenkrantz 1983 and the discussion of the Bayes information criterion in Wasserman 2004).

There are noncircular, relevant arguments for Ockham's razor, if one is willing to grant premises far more dubious than the theories Ockham's razor is used to justify. Leibniz ([1714] 1875) appealed to the Creator's taste for elegance. More recently, some "naturalistic" philosophers and machine learning researchers have replaced Providence with an equally vague and optimistic appeal to Evolution (e.g., Giere 1985; Duda et al. 2000, 464–465). But whereas a sufficiently powerful and kind Deity could save us from error in scientific questions never before encountered, it is hardly clear how selective pressures on our hominid ancestors could do so—unless Ockham's razor is invoked to argue that our evolved penchant for simplicity is a reliable guide to the truth in questions never before encountered.

Even if Providence or Evolution did arrange the truth of simple theories after a fashion that may remain eternally obscure, it would surely be nice, in addition, to have a clear, normative 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 q3 such an argument.[1] The idea is that it is hopeless to provide an *a priori*

[1] The approach is based on concepts from computational learning theory. An early appearance of retractions as a fundamental cost of inquiry is in Putnam 1965. An abstract theory of complexity of inductive inference is presented in Daley and Smith 1986. For a survey of results concerning retractions in inductive inference see Jain et

explanation of how simplicity points at the truth immediately, since the truth may depend upon subtle empirical effects that have not yet been observed or even conceived. The best that Ockham's razor could guarantee *a priori* is to keep us on the straightest possible path to the truth, allowing for unavoidable twists and turns along the way as new effects are discovered—and that is just what it does guarantee. Readers who wish to cut to the chase may prefer to peek immediately at Theorem 1 in Section 5 prior to reviewing the relevant definitions.

2. Illustration: Empirical Effects. Suppose that you are interested in the structure S of an unknown polynomial law

$$f(x) = \sum_{i \in S} a_i x^i, \tag{1}$$

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. 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.

So conceived, empirical effects have the property that they never appear if they do not exist but may appear arbitrarily late if they do exist.[3] To reduce the curve fitting problem to its essential elements, let E be a de-

al. 1999. Earlier versions of the following argument may be found in Schulte 1999; Kelly 2002, 2004; Kelly and Glymour 2004; and especially Kelly 2005, 2007, 2008.

[2] In statistics, the situation is analogous: increasing the sample size reduces the interval estimates of the values of the function at each argument.

[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.

numerable set of *potential effects* and assume that at most finitely many of these effects will ever occur. Assume that your laboratory merely reports the finite set of all effects that have been detected so far, so an *input sequence* is an upwardly nested sequence of finite subsets of E that converges to some finite subset S of E. An *input stream* or *empirical world* is an infinite input sequence. Let the effects presented in input sequence e be denoted S_e. The true answer to the effect accounting problem in empirical world w is then just S_w. Call this abstract problem the *effect accounting problem.* The effect accounting problem reflects, approximately, the structure of a number of naturally posed inference problems, such as determining the set of independent variables a dependent variable depends upon, determining quantum numbers from a set of reactions (Schulte 2000), and causal inference (Spirtes et al. 2000), in addition to the polynomial inference problem already mentioned.[4]

A *strategy* for effect accounting responds to an arbitrary input sequence either with a finite set of effects or with '?', indicating a refusal to choose. Strategy M *solves* the effect accounting problem if and only if M converges to the true set of effects S_w in each empirical world w. One obvious solution to the effect accounting problem is the strategy $M_0(e) = S_e$, which guesses exactly the effects it has seen so far. If the possibility of infinitely many effects were admitted, then the effect accounting problem would not be solvable at all, due to a classic result by Gold (1978).

Ockham's razor is the principle that one should never output an informative answer unless that answer is among the simplest answers compatible with experience. In the effect accounting problem, there is a uniquely simplest answer compatible with experience e, namely, the set S_e of effects reported so far along e.[5] Thus, strategy M is *Ockham* at e if and only if M produces either S_e or '?' in response to finite input sequence e.

If the inputs received so far are $e = (e_0, \ldots, e_{n+1})$, then let the immediately preceding evidential state be $e_- = (e_0, \ldots, e_n)$ (where e_- is stipulated to denote the empty sequence if e does). Say that solution M is *stalwart* at e if and only if $M(e) = M_{e_-}$ when $M(e_-) = S_e$—that is, if you are already accepting the simplest answer, don't drop it until it is no longer simplest. One may speak of stalwartness and of Ockham's razor as being satisfied from e onward (i.e., at each extension e' of e compatible with K).

The simplicity puzzle now arises because, although every convergent strategy must agree with an Ockham strategy eventually (since the true

[4] Strictly speaking, the inference of causal structure is more complicated because some finite sets of effects are routinely ruled out a priori. A more general setting of that sort is discussed below.

[5] Again, this condition is false in some applications and its relaxation is discussed below.

structure S_w in w is eventually the uniquely simplest structure compatible with the data presented along w), convergence is compatible with arbitrarily severe violations of Ockham's razor and stalwartness in the short run; for example, one could start with some complex answer $S \neq \varnothing$ and retract back to $\varnothing$ if $S \neq S_e$ at stage 1000 (Salmon 1967). The trouble is that there are infinitely many ways to converge to the truth in the accounting problem, just as there are infinitely many algorithmic solutions to a solvable computational problem. The nuances of programming practice—the very stuff of textbook computer science—are derived not from solvability itself, but from efficiency or computational complexity (e.g., the time or storage space required to find the right answer). The proposal is that Ockham's razor is similarly grounded in the efficiency of empirical inquiry, rather than in mere convergence (solvability).

3. Costs of Inquiry. An obvious, doxastic cost of inquiry is the total number of times one's strategy produces a false answer prior to convergence to the true answer. Another is the number of times a conclusion is 'taken back' or *retracted* prior to convergence, which corresponds to the degree of 'straightness' of the path followed to the truth.[6] One might also wish to minimize the respective times by which these retractions occur, since there is no point 'living a lie' longer than necessary or allowing subsidiary conclusions to accumulate prior to being 'flushed' when the retraction occurs. Taken together, these costs reflect the directness and timeliness with which one surmounts obstacles on one's way to the truth, and a strategy that minimizes them can be said to have the strongest possible connection with the truth. Insofar as epistemology is distinguishable from 'psychologism' by its regard for truth conduciveness (Bonjour 1985), minimization of retractions, retraction times, and errors is a properly epistemic consideration—indeed, more so than coherence, plausibility, confirmation, or rhetorical force. For a given strategy M and infinite input stream w, let the total *loss* of M in w be represented by the pair

$$\lambda(M, w) = (q, (r_1, \ldots, r_k)), \tag{2}$$

where q is the total number of errors or false answers output by M in w, k is the total number of retractions performed by M in w, and r_i is the stage of inquiry at which the ith retraction occurs.

Happily, it turns out that one need only consider comparisons in which one cost sequence is as good as or better than another in each of the above dimensions (i.e., Pareto comparisons). Accordingly, let $(q,(r_1, \ldots, r_k)) \leq (q',(r'_0, \ldots, r'_{k'}))$ if and only if $q \leq q'$ and there exists a subsequence

[6] Retractions are called *mind-changes* in computational learning theory (cf. Jain et al. 1999) and *contractions* in the literature on belief revision (Gärdenfors 1988).

$(u_0, \ldots, u_k)$ of $(r'_0, \ldots, r'_{k'})$ such that for each i from 1 to k, $r_i \leq u_i$. Then for cost pairs $\mathbf{v}, \mathbf{v}'$, define $\mathbf{v} < \mathbf{v}'$ iff $\mathbf{v} \leq \mathbf{v}'$ but $\mathbf{v}' \leq \mathbf{v}$.

A *potential cost bound* is like a cost pair except that the first infinite ordinal ω may occur. Potential cost bound $\mathbf{b}$ is a *cost bound* on set X of cost pairs if and only if each $\mathbf{v}$ in X is $\leq \mathbf{b}$. If $\mathbf{b}, \mathbf{b}'$ are both potential cost bounds, say that $\mathbf{b} \leq \mathbf{b}'$ if and only if for each cost pair $\mathbf{v}$, if $\mathbf{v} \leq \mathbf{b}$ then $\mathbf{v} \leq \mathbf{b}'$. Then each set X of cost pairs has a unique, least upper cost bound $\sup(X)$ (see Kelly 2007).

4. Empirical Complexity and Efficiency. No solution to the effect accounting problem achieves a nontrivial cost bound over the whole effect accounting problem, since each theory can be overturned by future effects in the arbitrarily remote future. Computational complexity theory (Aho et al. 1974) has long since sidestepped a similar conceptual difficulty by partitioning potential inputs into respective *sizes* (i.e., lengths) and by then examining worst case resource bounds over the finitely many inputs of a given length. In empirical problems, each input stream w has infinite length, but it remains natural to partition potential input streams by *empirical complexity*. After finite input sequence e has been received, let the *conditional empirical complexity* of w *given* e be defined as: $c(w, e) = |S_w| - |S_e|$, where $|S|$ is the cardinality of S, and let the nth *empirical complexity cell* given e be the set $C_e(n)$ of all worlds w 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 section.

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_- (recall that e_- denotes the result of deleting the last entry of e). Say that such a strategy *agrees with* M along e_- (abbreviated $M \equiv M'$).

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

$$M \underset{e}{\leq} M' \text{ iff } (\forall\, n)\ \lambda_e(M, n) \leq \lambda_e(M', n);$$

$$M \underset{e}{<} M' \text{ iff } M \underset{e}{\leq} M' \text{ and } M' \underset{e}{\nleq} M;$$

$$M \underset{e}{\prec} M' \text{ iff } (\forall\, n)\ C_e(n) \neq \varnothing \Rightarrow \lambda_e(M, n) < \lambda_e(M', n).$$

These comparisons give rise to two natural properties of strategies:

$$M \text{ is } \textit{strongly beaten} \text{ at } e \text{ iff } (\exists \text{ solution } M' \underset{e_-}{\equiv} M)\ M' \underset{e}{\prec} M;$$

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

$$M \text{ is } \textit{efficient} \text{ at } e \text{ iff } (\forall \text{ solution } M' \underset{e_-}{\equiv} M)\ M' \underset{e}{\geq} M.$$

A solution that is strongly beaten does worse than some alternative solution in worst case performance in *each* nonempty, empirical complexity cell. A solution that is beaten does worse than some solution in some complexity cell and no better in the rest of the cells. An efficient solution is as good as an arbitrary solution in worst case performance in *each* empirical complexity cell. One may speak of being efficient from *e* onward. Being strongly beaten implies being beaten, which implies inefficiency.

5. The New Solution. Here is the proposed efficiency argument for Ockham's razor. The proof is in the appendix.

Theorem 1 (Ockham efficiency characterization). Let M solve the effect accounting problem. Let e be a finite input sequence. 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 strongly beaten from *e* onward.

So the set of all solutions to the effect accounting problem is cleanly partitioned given *e* into two groups: the solutions that are stalwart, Ockham, and efficient from *e* onward and the solutions that are strongly beaten at some stage $e' \geq e$ due to future violations of the stalwart, Ockham property. As promised, the argument is *a priori*, normative, truth directed, and yet noncircular. The argument presumes no prior probabilistic bias, so there is no question of a circular appeal to such a bias. 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, the Synthetic a Priori, Convention, or Evolution. There is nothing built into the argument other than a question, simplicity relative to the question, and efficient convergence to the true answer to the question.

Furthermore, the argument is *stable* in the sense that born again Ockhamism strongly beats recidivism at each contemplated violation, so past violations, no matter how severe, do not undermine the normative force of the argument at each moment. That is important, for Ockham violations are practically unavoidable in real science, either due to a failure to think of the simplest answer in time or due to spurious, auxiliary objections that are resolved only later.

The argument does not accomplish the impossible. Ockham's razor cannot be shown, without circularity, to point at or track the truth immediately, for some effects may be arbitrarily hard to detect given current technologies and sample sizes, in which case all possible, convergent strategies—Ockham strategies included—can be forced to retract their opinions any finite number of times. Nor can one demand a stronger notion of efficiency with respect to retractions and errors. (1) One cannot establish weak dominance for Ockham methods with respect to all problem instances jointly, because anticipation of unseen effects might be vindicated immediately, saving retractions that the Ockham method would have to perform when the effects appear. (2) Nor can one show that Ockham's razor does best in terms of a global worst case bound over all problem instances (minimax theory), for such worst case bounds on errors and retractions are trivially infinite for all methods at every stage. (3) Nor can one show a decisive advantage for Ockham's razor in terms of expected retractions. For example, if the question is whether one will see at least one effect, then the expected retractions of the obvious strategy $M(e) = S_e$ are less than those of an arbitrary Ockham violator only if the prior probability of the simpler answer is at least one half, so that if more than one complex world carries nonzero probability, no complex world is as probable as the simplest world, which begs the question in favor of simplicity.[7] If the prior probability of the simple hypothesis drops below 0.5, the advantage lies not only with violating Ockham's razor, but with violating it more rather than less. So Bayesians must either beg the question or rule strongly against Ockham.

[7] Let M_i be a non-Ockham strategy that starts by guessing answer ≥ 1 until no effect is seen by stage i, at which point M_i returns 0. If the effect is ever seen, M returns answer ≥ 1. Consider the competing Ockham method M that always guesses 0 until the effect is seen, at which time M returns answer ≥ 1. Consider probabilities at stage 0. Let a denote the probability that no effect occurs, let b denote the probability that an effect occurs no later than stage i and let c denote the probability that an effect occurs after stage i. Then, *a priori*, the expected retractions of M_i are given by $a + 2c$, whereas the expected retractions of M are $b + c$. So the Ockham strategy M does better when $a + c > b$. Since $a + c + b = 1$, this is true if and only if $b < 0.5$. By increasing i, one can drive c arbitrarily small (by countable additivity), so if the Ockham strategy is to beat the expected retractions of an arbitrary M_i, then $a \geq b$. That implies that each of the several (complex) possibilities over which mass b is distributed receives less probability than the simple world carrying probability b. This bias increases with i and with the number of ways the complex theory can be true.

Some applications, like the search for causal structure (Spirtes et al. 2000), imply *a priori* restrictions on the possible, finite sets of effects that correspond to possible answers. Let Γ be the set of a priori possible, finite sets of effects that nature might reveal for eternity. Let Γ_e denote the subset of Γ whose elements are all consistent with e (where S is consistent with e if $S_e \subseteq S$). A *directed path* in Γ_e is just an upwardly nested, finite sequence of elements of Γ_e. Now define the conditional empirical complexity $c(S, e)$ of world $S \in \Gamma_e$ given e as one less than the length of a longest path in Γ_e terminating in S and let $c(w, e) = c(S_w, e)$. Theorem 1 extends to such cases (cf. Kelly 2008), except that the beating incurred by Ockham violators may fail to be strong when there is more than one simplest answer compatible with e.

The preceding approach still assumes that the theorist is fed pre-digested empirical effects, rather than raw experience itself. Here is a very general definition of empirical complexity that agrees with the preceding account when applied to pre-digested problems (cf. Kelly 2008). In general, an *empirical problem* $\mathcal{P}$ consists of a set K of possible *input streams* or *worlds* and an *empirical question* Π, which is just a partition of K into *potential answers*. No objectionable pre-digestion is assumed here: the successive inputs presented by $w \in K$ could be boolean bits in a highly 'gruified' coding scheme with an ocean of information irrelevant to the question Π thrown in. If e is a finite input sequence, let K_e denote the restriction of K to input streams extending finite input sequence e. 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_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. Restrict attention to the natural problems $\mathcal{P}$ in which $\lim_{i\to\infty} S_{w|i}$ exists, for each $w \in K$, and let S_w denote this limit. Let Γ_e denote the set of all S_w such that $w \in K_e$. If $S, S' \in \Gamma_e$, say that $S \leq S'$ if and only if for each e' extending e such that $S = S_{e'}$, there exists e'' extending e' such that $S' = S_{e''}$. Now define $c(w, e)$ in terms of longest $\leq$-path length in Γ_e to S_e, just as in the preceding paragraph. This definition of simplicity depends only upon the (semantic) structures of K, Π, so it is invariant under arbitrary, grue-like (Goodman 1983) recodings of the inputs (which leave the semantics of the problem intact). Moreover, if $\mathcal{P}_\Gamma$ is the (pre-digested) sort of problem discussed in the preceding paragraph, then it can be shown that the complexity degree assignment $c(w, e)$ just defined is identical to the one defined in the preceding paragraph. Finally, applying this definition to problems that look, intuitively, like effect accounting problems (e.g., polynomial structure, causal structure, or conservation laws) identifies what intuition would point out as the empirical effects relevant to the question Π given. So careful attention to truth finding efficiency provides not only a novel explanation of Ockham's razor, but also a fresh perspective on the nature of simplicity, itself.

Appendix: Proof of Theorem 1.

q4

(2 ⇒ 3), is immediate from the definitions. For (3 ⇒ 1), suppose that M violates Ockham's razor or stalwartness at finite input sequence e. Let M be a solution that is stalwart and Ockham from e' onward. Let $e \geq e'$ have length j. Then M is Ockham and stalwart from e onward. Let M' be an arbitrary solution such that $M' \equiv M$. Let $r_1, \ldots, r_k$ be the retraction times for both M and M' along e_-. Let^{e}q denote the number of times M produces an answer other than S_e along e_-. Consider the hard case in which M retracts at e. Let $w \in C_e(0)$. In w, M retracts at e but never retracts after e and M produces only the true answer S_e after e. Hence:

$$\lambda_e(M, 0) \leq (q, (r_1, \ldots, r_k, j)). \qquad \text{(A1)}$$

There exists $w_0 \in C_e(0)$ (just extend e by repeating S_e forever). Then $M(e_-) = M'(e_-)$ is false in w_0. So since M' is a solution, M' converges to the true answer S_e in w_0 at some point after e_-, which implies a retraction at some point no sooner than e. Hence:

$$\lambda_e(M', 0) \geq (q, (r_1, \ldots, r_k, j)) \geq \lambda_e(M, 0). \qquad \text{(A2)}$$

If $C_e(n+1) = \emptyset$, then every method succeeds under the trivial bound $(0, ())$, so suppose that $C_e(n+1) \neq \emptyset$. Since M is a stalwart, Ockham solution, M retracts at most once at each new effect, so

$$\lambda_e(M, n+1) \leq (\omega, (r_1, \ldots, r_k, j, \underbrace{\omega, \ldots, \omega}_{n+1 \text{ times}})). \qquad \text{(A3)}$$

Let arbitrary natural number i be given. Since M' is a solution, M' eventually converges to $A_0 = S_e$ in w_0, so there exists e_0 such that $e \leq e_0 < w_0$ by which M' has retracted the false answer $M'(e_-)$ and has produced the true answer A_0 successively at least i times after the end of e, so M' retracts at least as late as e in e_0. Then there exists $w_1 \in C_e(1)$ such that $e_0 < w_1$ (since $C_e(n+1) \neq \emptyset$, nature can choose some $x_0 \in E - A_0$ and extend e_0 forever with answer $A_1 = A_0 \cup \{x_0\}$). Again, M' must converge to A_1 in w_1 and, therefore, produces A_1 successively at least i times by some initial segment e_1 of w that extends e_0. Continuing in this manner, construct $w_{n+1} \in C_e(n+1)$. Then

$$\lambda_e(M', w_{n+1}) \geq (i, (r_1, \ldots, r_k, j, j + 1i, j + 2i, \ldots, j + (n+1)i)). \qquad \text{(A4)}$$

Since i is arbitrary,

$$\lambda_e(M', n+1) \geq (\omega, (r_1, \ldots, r_k, j, \underbrace{\omega, \ldots, \omega}_{n+1\text{times}})) \geq \lambda_e(M, n+1). \tag{A5}$$

Now consider the easy case in which M does not retract at e. Then the argument is similar to that in the preceding case except that the retraction at j is dropped from all the bounds.

For the proof of ($1 \Rightarrow 2$), let M be a solution that violates either Ockham's razor or stalwartness at e of length j. Let M' return $S_{e'}$ at each $e' \in K_{\text{fin}}$ such that $e' \geq e$ and let M' agree with M otherwise. Then $M' \underset{e_-}{\equiv} M$ by construction and M' is evidently a solution. Let $r_1, \ldots, r_k$ be the retraction times for both M and M' along e up to but not including the last entry in e.

Consider the case in which M violates Ockham's razor at e. So for some $A \subseteq E$, $M(e) = A \neq S_e$. Let $w \in C_e(0)$. Then A is false in w and S_e is true in w. Let q denote the number of times both M and M' produce an answer other than S_e along e_-. Since M' produces the true answer at e in w and continues to produce it thereafter:

$$\lambda_e(M', 0) \leq (q, (r_1, \ldots, r_k, j)).$$

There exists w_0 in $C_e(0)$ (just extend e forever with S_e). Since A is false in w_0 and M is a solution, M retracts A in w_0 at some stage greater than j, so

$$\lambda_e(M, 0) \geq \lambda(M, w_0) \geq (q+1, (r_1, \ldots, r_k, j+1)) > \lambda_e(M', 0). \tag{A6}$$

As in the proof of ($3 \Rightarrow 1$), it suffices to consider the case in which $C_e(n+1) \neq \varnothing$. Since M' produces $S_{e'}$ at each $e' \geq e$,

$$\lambda_e(M', n+1) \leq (\omega, (r_1, \ldots, r_k, j, \underbrace{\omega, \ldots, \omega}_{n+1 \text{ times}})). \tag{A7}$$

Let $i \in \omega$. Answer $A = M(e)$ is false in w_0, so since M is a solution, M eventually converges to $A_0 = S_e$ in w_0, so there exists e_0 properly extending e by which M has produced A_0 successively at least i times after the end of e and M retracts A back to A_0 no sooner than stage $j+1$. Now continue according to the recipe described in the proof of ($3 \Rightarrow 1$) to construct $w_{n+1} \in C_e(n+1)$ such that:

$$\lambda(M, w_{n+1}) \geq (i, (r_1, \ldots, r_k, j+1, j+1i, j+2i, \ldots, j+(n+1)i)). \tag{A8}$$

Since i is arbitrary,

$$\lambda_e(M, n+1) \geq (\omega, (r_1, \ldots, r_k, j+1, \underbrace{\omega, \ldots, \omega}_{n+1 \text{ times}})) > \lambda_e(M', n+1). \tag{A9}$$

Next, consider the case in which M violates stalwartness at e. So $M(e_{-}) = S_e$ but $M(e) \neq S_e$. Let $w \in C_e(0)$. Let q denote the number of errors committed in w by both M and M' along e_{-}. Since $M'(e_{-}) = S_e$, it follows that M' does not retract in w from j onward, so:

$$\lambda_e(M', 0) \leq (q, (r_1, \ldots, r_k)). \tag{A10}$$

Again, there exists w_0 in $C_e(0)$. Since M retracts at j,

$$\lambda_e(M, 0) \geq (q, (r_1, \ldots, r_k, j)) > \lambda_e(M', 0). \tag{A11}$$

Let $C_e(n + 1) \neq \emptyset$. Since M' produces $S_{e'}$ at each $e' \geq e$,

$$\lambda_e(M', n + 1) \leq (\omega, (r_1, \ldots, r_k, \underbrace{\omega, \ldots, \omega}_{n+1 \ \text{times}})). \tag{A12}$$

Let arbitrary natural number i be given. Since M retracts at j, one may continue according to the recipe described in the proof of ($3 \Rightarrow 1$) to construct w_{n+1} extending e in $C_e(n + 1)$ such that:

$$\lambda(M, w_{n+1}) \geq (i, (r_1, \ldots, r_k, j, j + 1i, j + 2i, \ldots, j + (n + 1)i)). \tag{A13}$$

Since i is arbitrary,

$$\lambda_e(M, n + 1) \geq (\omega, (r_1, \ldots, r_k, j, \underbrace{\omega, \ldots, \omega}_{n+1 \ \text{times}})) > \lambda_e(M', n + 1). \tag{A14}$$

REFERENCES

Aho, A., J. Hopcroft, and J. Ullman (1974), *The Design and Analysis of Computer Algorithms*. New York: Addison-Wesley.

Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle", in B. N. Petrov and F. Csaki (eds.), *The Second International Symposium on Information Theory*. Budapest: Akadémiai Kiadó, 267–281.

Bonjour, L. (1985), *The Structure of Empirical Knowledge*. Cambridge, MA: Harvard University Press.

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

Daley, R., and C. Smith, (1986), "On the Complexity of Inductive Inference", *Information and Control* 69:12–40.

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

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

Friedman, M. (1983), *Foundations of Space-Time Theories: Relativistic Physics and Philosophy of Science*. Princeton, NJ: Princeton University Press.

Gärdenfors, P. (1988), *Knowledge in Flux*. Cambridge, MA: MIT Press.

Giere, R. (1985), "Philosophy of Science Naturalized," *Philosophy of Science* 52: 331–356.

Gold, E. (1978), "Language Identification in the Limit", *Information and Control* 10: 447–474.

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

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

Harman, G. (1965), “The Inference to the Best Explanation”, *Philosophical Review* 74: 88–95.

Jain, S., D. Osherson, J. Royer, and A. Sharma (1999), *Systems That Learn*. Cambridge, MA: MIT Press.

Kelly, K. (2002), “Efficient Convergence Implies Ockham’s Razor”, paper delivered at the 2002 International Workshop on Computational Models of Scientific Reasoning and Applications, Las Vegas, June 24–27.

——— (2004), “Justification as Truth-Finding Efficiency: How Ockham’s Razor Works”, *Minds and Machines* 14: 485-505.

——— (2007), “Ockham’s Razor, Empirical Complexity, and Truth-Finding Efficiency”, *Theoretical Computer Science*, 270–289.

——— (2008)“Ockham’s Razor, Truth, and Information,” in J. Van Benthem and P. Adriaans (eds.), *Philosophy of Information*. Amsterdam: Elsevier, forthcoming.

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

Leibniz, G. W. ([1714] 1875), *Monadologie*, in L. E. Loemker (ed.), *Die Philosophischen Schriften von G. W. Leibniz*, vol. 4. Berlin: Gerhardt, 607–623.

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

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

Rissanen, J. (1983), “A Universal Prior for Integers and Estimation by Minimum Description Length”, *Annals of Statistics* 11: 416–431.

Rosenkrantz, R. (1983), “Why Glymour is a Bayesian”, in J. Earman (ed.), *Testing Scientific Theories*. Minneapolis: University of Minnesota Press, 69–98.

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

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

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

Spirtes, P., C. Glymour, and R. Scheines (2000), *Causation, Prediction, and Search*. Cambridge, MA: MIT Press.

Spirtes, P., and J. Zhang (2003), “Strong Faithfulness and Uniform Consistency in Causal Inference”, in Christopher Meek and Uffe Kjærulff (eds.), *Proceedings of the 19th Conference in Uncertainty in Artificial Intelligence*. San Mateo, CA: Kaufmann, 632–639.

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

Wasserman, L. (2004), *All of Statistics: A Concise Course in Statistical Inference*. New York: Springer.

QUERIES TO THE AUTHOR

1 Au: In the references, I changed: Kelly, K. (2002), Efficient Convergence Implies Ockhams Razor, paper delivered at the 2002 International Workshop on Computational Models of Scientific Reasoning and Applications, Las Vegas, June 24–27. Correct as modified?

2 Au: For the Akaike reference, I changed the final o to an accented ó in the publisher name "Akadémiai Kiadó"; there was a stray "l" after the comma, which I have removed. Does this now appear to be correct?

3 Au: Please check the author date citations for Kelly in n. 2 very carefully. Should Kelly be cited twice in the last sentence of this note? Please note that Kelly 2005 is not in the reference list. For the Kelly 2007 ref. list entry, please add vol. number for this journal, if available.

4 Au: There is no square in the text to show where the proof ends; should a square be added?