Chapter 4

HOW SIMPLICITY HELPS YOU FIND THE TRUTH WITHOUT POINTING AT IT

KEVIN T. KELLY
Department of Philosophy, Carnegie Mellon University, Pittsburg, PA 15213, U.S.A., kk3n@andrew.cmu.edu

Abstract: It seems that a fixed bias toward simplicity should help one find the truth, since scientific theorizing is guided by such a bias. But it also seems that a fixed bias toward simplicity cannot indicate or point at the truth, since an indicator has to be sensitive to what it indicates. I argue that both views are correct. It is demonstrated, for a broad range of cases, that the Ockham strategy of favoring the simplest hypothesis, together with the strategy of never dropping the simplest hypothesis until it is no longer simplest, uniquely minimizes reversals of opinion and the times at which the reversals occur prior to convergence to the truth. Thus, simplicity guides one down the straightest path to the truth, even though that path may involve twists and turns along the way. The proof does not appeal to prior probabilities biased toward simplicity. Instead, it is based upon minimization of worst-case cost bounds over complexity classes of possibilities.

1 THE SIMPLICITY PUZZLE

There are infinitely many alternative hypotheses consistent with any finite amount of experience, so how is one entitled to choose among them? Scientists boldly respond with appeals to "Ockham's razor", which selects the "simplest" hypothesis among them, where simplicity is a vague family of virtues including unity, testability, uniformity of nature, minimal causal entanglement, and minimal ontological commitment. The debate over "scientific realism" in the philosophy of science hinges on the

M. Friend, N. B. Goethe and V. S. Harizanov (eds.), Induction, Algorithmic Learning Theory, and Philosophy, 111–143.

propriety of this response. Scientific realists view simplicity as a legitimate reason for belief and anti-realists do not. More recently, the question has spread to computer science, where the widespread adoption of simplicity-biased learning and data-mining software makes it all the more unavoidable [15].

Scientific realists infer from the rhetorical force of simplicity arguments that the simpler theory is better "confirmed" and, hence, that belief in the simpler theory is better justified [5]. Anti-realists [21] concede the rhetorical force of simplicity arguments, but wonder why they should be so compelling.[1] Presumably, epistemic justification is supposed to direct one toward the truth and away from error. But how could simplicity do any such thing? If you already know that the truth is simple or probably simple, then Ockham's razor is unnecessary, and if you don't already know that the truth is simple or probably simple, then how could a fixed bias toward simplicity steer you toward the true theory? For a fixed bias can no more indicate the truth than a compass whose needle is stuck can indicate direction.

There are answers in the literature, but only irrelevant or circular ones. The most familiar and intuitive argument for realism is that it would be a "miracle" if a complex, disunified theory with many free parameters were true when a unified theory accounts for the same data. But the alleged miracle is only a miracle with respect to one's personal, prior probabilities. At the level of theories, one is urged to be even-handed, so that both the simple theory and its complex competitor carry non-zero prior probability. Then since the complex theory has more free parameters to tweak than the simple theory has, each particular setting of its parameters has lower prior probability than does each of the parameter settings of the simple theory. So the miracle argument amounts to an *a priori* bias in favor of simple parameter settings over complex parameter settings. But that is just how a Bayesian agent implements Ockham's razor; the question under consideration is why one should implement it, so far as finding the true theory is concerned [12].

Another standard argument is that simple explanations are "better" and that one is entitled, somehow, to infer the "best" explanation [7]. But even

[1] Van Fraassen focuses on the problem of theories that are not distinguished even by all the evidence that might ever be collected. There is no question of simplicity guiding you to the truth in such cases, since no method based only on observations possibly could. On the other hand, it is almost always the case that simple and complex theories that disagree about some future observations are compatible with the current data and the simpler one is preferred (e.g., in routine curve-fitting). I focus exclusively on this ubiquitous, local problem of simplicity rather than on the hopelessly global one.

assuming that the simplest explanation is best, that sounds like wishful thinking [21], for one may like strong explanations, but liking them doesn't make them true. The same objection applies to the view that simplicity is just one virtue among many [13]. An apparently more promising idea is that simple or unified theories compatible with the data are more severely tested or probed by the data and, hence, are better "corroborated" [16] or "confirmed" [5]. But if the truth isn't simple, then the truth is less testable than falsehood, so why should one presume that the truth is simple? Either considerations like testability and explanatory power are irrelevant to the question at hand or one must assume, circularly, that the world is simple in order to explain why one is entitled to prefer more testable theories.

Another idea [20] is that if a simple theory is false, future data will lead to its retraction, so a simplicity-biased, rational agent will converge to the truth in the limit of inquiry. But the question at hand is not merely how to overcome one's simplicity bias. If Ockham's razor is truly helpful, as opposed to merely being a defeasible impediment, it should facilitate truth-finding better than competing biases. But since other biases would also be over-ruled by experience eventually, mere convergence to the truth does not explain why simplicity is a better bias than any other, so this approach is irrelevant to the realism debate.

Perhaps the most interesting of the standard arguments in favor of simplicity is based upon the concept of "overfitting" [2]. The idea is that predicting the future by means of an equation with too many free parameters compared to the size of the sample is more likely to produce a prediction far from the true value. But that argument has more to do with the size of the sample than with the nature of reality, for the same argument against overfitting still favors use of a simple theory for prediction from small samples even when you know that the true theory is very complex. So although this argument is sound and compelling, so far as using an equation for predictive purposes is concerned, it is also irrelevant to the question at hand, which concerns finding the true theory rather than using a false theory for predictive purposes.

Taking stock of the standard answers, it appears that the anti-realist's objection is insuperable, for it can only be met by showing how a fixed simplicity bias helps one find the truth even when the truth is complex. That sounds hopeless, for in complex worlds simplicity points in the wrong direction. Nonetheless, it is demonstrated below that simplicity is the best possible advice for a truth-seeker to follow, in a certain sense, no matter how complex the truth might be.

2 THE FREEWAY TO THE TRUTH

It is no fault of simplicity that it fails to point out or indicate the true theory, since nothing possibly could. General theories or models can always be overturned in the future by the discovery of subtle effects missed earlier even by the most diligent probing. So science is not an uneventful voyage along a compass course to the truth. It is more like an impromptu road trip through the mountains, with numerous hairpin twists and detours along the way. Taking this more appropriate metaphor seriously is the key to the simplicity puzzle.

Suppose that, on your way to a distant city, you exit the freeway for a rest stop and become lost in the neighboring town. If you ask for directions, you will be told the shortest route back to the freeway entrance ramp even before you say which city you are headed to, because the freeway is the best route to anywhere a stranger might wish to go (Figure 4.1). That remains true even if the shortest route to the entrance ramp takes you west for a few miles when your ultimate destination is east.

Suppose that you disregard the local resident's advice. You find yourself on small dirt tracks headed nowhere and, after enough of this, you make a U-turn and head back toward the entrance ramp. Your hubris is rewarded by the addition of one gratuitous course reversal to your route before you even begin the real journey on the freeway, with all of its unavoidable curves through the mountains. So even if directions to the freeway take you directly away from your ultimate goal at first, you ought to follow them.

The journey to the truth likewise occasions reversals and detours: revolutions or revisions in which one theory is retracted and replaced by another and the textbooks are rewritten accordingly [13]. Some retractions are unavoidable in principle given that one finds the truth at all, since accepting a general theory always occasions a risk of being surprised by an unanticipated anomaly later. In that case, retracting the theory is not merely excusable but virtuous—the alternative would be dogmatic commitment to error for

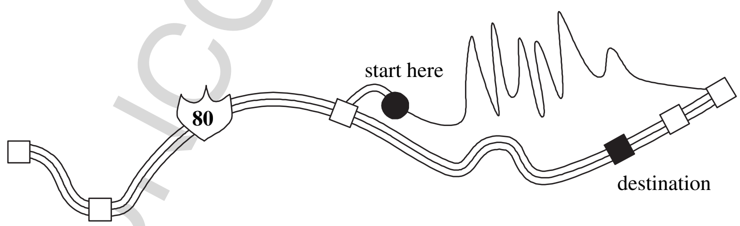

Figure 4.1. Entrance ramp

eternity, as Popper [16] emphasized. But gratuitous reversals in the course of inquiry are another matter entirely: it would be better to avoid them.[2]

Suppose that you violate Ockham's razor by selecting a theory more complex than experience requires. Then the simple experience up to now can be extended for eternity with equally uniform, simple experience, devoid of "effects" whose detection would indicate the need to postulate more causes or free parameters. If you refuse ever to retract to a simple hypothesis, you never arrive at the truth at all, so you have to take the bait, eventually, and fall back to the simplest theory. Now you are essentially where you would have been had you never violated Ockham's razor, except that you have already retracted once; and you are still subject to the future appearance of any number of subtle empirical effects that could not be detected at current sample sizes or using current instrumentation. Each such effect may occur sufficiently late to result in an unavoidable retraction. So you are stuck with an extra retraction at the outset added to all of these. Therefore, always presuming that the world is simple keeps you on the straightest path to the truth even though the truth may be arbitrarily complex! So both the realist and the anti-realist are right, since simplicity keeps one on the straightest path to the truth, but the straightest path may point in the wrong direction for the time being and for any finite number of times in the future as well, assuming that you converge to the truth at all.

3 ILLUSTRATION: COUNTING MARBLES

Suppose that you are studying a marble-emitting device that occasionally emits a marble (a new empirical effect). Your job is to determine how many marbles it will ever emit (how many free parameters the true theory has). You know nothing about when the marbles will be emitted (empirical effects may be arbitrarily small and hard to notice) but you do know on general grounds that at most finitely many marbles will be emitted (every theory under consideration has at most finitely many free parameters). Call the situation just described the *counting problem*.

In this simplistic setting, it seems that when exactly k marbles have been seen so far, k is the simplest answer compatible with experience. First,

[2] Retractions have been studied extensively in computational learning theory. For a review cf. [9]. The first version of the U-turn argument, albeit restricted to problems in which at most k marbles may be seen, is presented in [17]. An infinite ordinal version of the argument, based loosely on ideas in [3] is presented in [10], but that idea still can't handle the marble counting problem described below.

k posits the fewest entities among all answers compatible with experience, which accords with the standard formulation of Ockham's razor. Second, k is satisfied by the most uniform (i.e., eternally marble-free) course of future experience, for alternative answers involve discrete "kinks" in experience (i.e., each time another marble is seen). Third, k has the fewest free parameters (for answer $k + k'$ leaves the appearance time of each of the extra k' posited marbles unspecified). Fourth, k is the best explanation of the data, since $k + k'$ leaves each of the k' appearance times unexplained.[3] Fifth, k is most testable, for if k is false, it is refuted, eventually, but answer $k + k'$ is false but never strictly refuted if the truth is less than $k + k'$.

A *strategy* for solving the counting problem examines the current marble history at each stage and returns either a natural number k indicating the total number of marbles or the skeptical response "?", which indicates a refusal to guess. Such a strategy *solves* the counting problem in the limit if and only if it converges, on increasing data, to the true count k, no matter what the true k happens to be and no matter when the k marbles happen to appear.

Now suppose that you solve the counting problem in the convergent sense just defined. Suppose, further, that no marbles have appeared yet, so the Ockham answer is 0, but you violate Ockham's razor by guessing some k greater than 0 (Figure 4.2). Everything you have seen is consistent with the possibility of never seeing any marbles. Since you converge to the truth, it follows that if the truth is 0, you must eventually converge to 0, so you retract k and revise to 0 at some point. Now it is possible for you to see a marble followed by no more marbles. Since you converge to the truth, you retract 0 eventually and replace it with 1, and so forth. So each answer k is satisfied by a world compatible with the problem's background assumptions

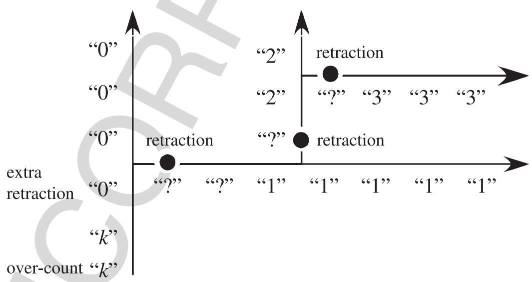

Figure 4.2. U-turn

[3] One might object that if the k marbles have appeared at each stage so far, then one would expect them to continue appearing forever, but that violates the background assumption that they will stop appearing eventually.

in which you retract $k + 1$ times. But had you always produced the Ockham answer at each stage, you would have retracted at most k times in an arbitrary world satisfying answer k. So your worst-case retractions are worse than the Ockham strategy's over each answer. Your initial retraction is analogous to the initial U-turn back to the entrance ramp being added to all the course reversals encountered on one's journey home after getting on the freeway.

Another natural consequence of the U-turn argument is that, after having selected answer 0, you should never retract it until it ceases to be simplest. Call this property *stalwartness*. For suppose that no marbles have been seen and that you follow Ockham's advice by choosing answer 0. Suppose, later, that you retract this answer in spite of the fact that no marble has been observed (for general, skeptical reasons, perhaps). Then if you converge to the truth, this initial retraction gets added to all the others you perform, regardless of which answer is true. So your worst-case retraction bound in answer k is $k + 1$, whereas a stalwart Ockham strategy can converge to the truth with just k retractions in answer k.

As simple as it is, the preceding logic has applications to real scientific questions. For example, consider the case of finding the polynomial degree of the true law, assuming that the law is polynomial. It is plausible to assume that larger samples or improvements in instrumentation allow one to progressively narrow in on the true value of the dependent variable y for any specified, rational value of the independent variable x over some closed, bounded interval as time progresses. Any finite number of such observations for a linear law is compatible with the discovery of a small quadratic effect later. Then any finite amount of such data for a quadratic law is compatible with the discovery of a small cubic effect later, etc.[4] The occasional appearances of these arbitrarily small (i.e., arbitrarily late), higher-order effects are analogous to the occasional appearances of marbles and polynomial degree k is analogous to seeing exactly k marbles for eternity.

4 ITERATING THE ARGUMENT

To this point, the U-turn argument has been applied only in cases in which no marble (anomaly) has yet been detected. But suppose that a marble appears after you say 0 but you stubbornly retain the answer 0 (Figure 4.3). Suppose, further, that when the second marble appears you violate Ockham's

[4] Popper [16] had a similar idea, except that he assumed exact measurements and counted the number of distinct measurements required to refute a given curve. In science, the observations are never exact and the logic is as I have described.

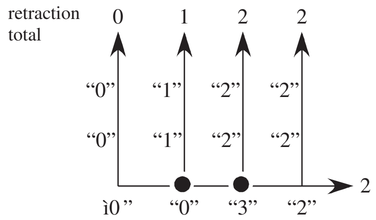

Figure 4.3. Ockham violator who is efficient *ex ante*

razor by producing 3. Thereafter, you follow Ockham’s advice. The U-turn logic rehearsed above does not distinguish your performance from that of the natural strategy that just counts the current marbles, for although guessing 3 opens you to the risk of retracting back to 2 later, that extra retraction is concealed by the retraction you saved by not retracting 0 to 1 earlier. So you converge to the truth and match the Ockham strategy’s performance in terms of overall, worst-case retractions within each answer.

The preceding analysis is carried out at the onset of inquiry (i.e., *ex ante*). The situation changes if your efficiency is assessed *ex post*, at the moment you first violate Ockham’s razor by over-counting; e.g., by saying 3 upon seeing the last entry in input sequence $e = (e_0, \ldots, e_n)$ in which only two marbles are presented. At that very moment, the input data e are already fixed, as is the sequence $b = (B_0, \ldots, B_{n-1})$ of answers you chose at each stage along $e_- = (e_0, \ldots, e_{n-1})$. So only worlds that present e and only strategies that produce b along e_- should count when your efficiency is assessed at the moment e has been presented.

Now the U-turn argument rules out over-counting even after some marbles have been seen (Figure 4.4). For suppose that you over-count for the first time at the end of e. Consider the hybrid strategy σ that agrees with you along e_- and that returns the current count thereafter. Strategy σ

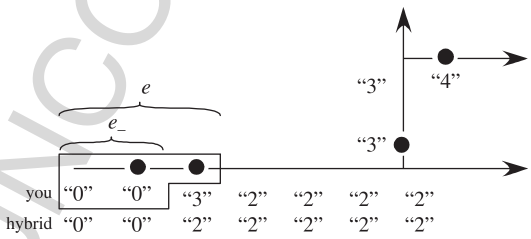

Figure 4.4. inefficiency exposed *ex post*

converges to the right answer (by counting up to it). Like you, strategy σ saves a retraction by not noticing the first marble (which appears in e_{-}), but σ produces the current count k at the end of e rather than the over-count you produce. Moreover, σ never retracts again if the truth is k and, in general, retracts at most k' times after the end of e if the truth is $k + k'$. But if you converge to the truth, you eventually retract your over-count at e to k if the truth is k (the initial U-turn back to the freeway to the truth) and then retract k to $k + 1$ if another marble is presented thereafter, etc., so you retract $k' + 1$ times after the end of e in answer $k + k'$ (the initial U-turn gets added to the k inevitable hairpins along the freeway to destination $k + k'$). Since σ acts just like you along e_{-}, both you and σ retract the same number of times (say r) along e_{-}. Since e is your first over-count, you retract at e, so you retract $r + 1$ times along e, so your worst-case bound over answer $k + k'$ is $r + k' + 2$. Even if σ retracts at e, the worst-case retraction bound for σ over answer $k + k'$ is at most $r + k' + 1$. So if you over-count for the first time at e, then for each answer $k + k'$, your worst-case retraction bound over $k + k'$ exceeds that of σ. So you are *strongly beaten* by σ at e, in the sense that σ agrees with you along e_{-} and over each answer k compatible with e, your worst-case bound over worlds compatible with e in answer k is worse than that of σ. If σ does as well as you in each answer k and worse in some answer, then say that you are *weakly beaten* by σ at e.

The same argument works at each e at which you (a) fail to repeat the answer you produced at the immediately preceding stage e_{-} and (b) choose any answer other than the current count. For in that case you retract at e, do no better than the hybrid method along e_{-}, and do worse in the worst case after e (due to having to retract back to k if no more marbles are seen after e). Say that a *lagged* Ockham strategy is a strategy that only violates Ockham's razor by retaining the answer it selected at the preceding stage. So an arbitrary solution is strongly beaten at *each* violation of the lagged Ockham property.

By a similar argument, if you solve the problem then you are strongly beaten by the hybrid strategy σ at an arbitrary e at which you fail to be stalwart. For if you are not stalwart at e, you drop the answer B you selected at e_{-} even though B is Ockham at e, so your stalwart clone σ also produces B at e_{-} (because it is a clone) and does not drop B at e (by stalwartness). Then, as before, σ retracts no more than you after e in each answer compatible with e, so σ beats you at e.

Being strongly beaten is no sin if every solution suffers that fate. To clinch the U-turn argument, each stalwart, lagged Ockham solution σ (e.g., the strategy that always returns the current count) is *efficient* at each e in the sense that over *each* answer compatible with e, solution σ does as well in worst-case retraction performance as an *arbitrary* solution σ' agreeing with

σ along e_-. For let e be given and let σ' be just like σ along e_-. Then both σ and σ' retract the same number of times r along e_- and both produce the current count k at e_-. If σ retracts at e, then since σ is stalwart, it follows that a marble was presented at e and σ produces the current count $k + 1$ at e. So if no more marbles are ever presented, σ' also has to retract to $k + 1$ eventually in order to converge to the truth. So σ' achieves no better retraction bound than σ in answer $k + 1$. Finally, σ retracts no more than k' times after e in answer $k + k'$ and σ' can be forced to retract at least k' times after e in answer $k + k'$ by presenting each of the remaining k marbles and waiting until σ' converges to the current count. So σ does at least as well as σ' in answer $k + k'$.

So the following has been shown.

Proposition 50 *Let σ solve the counting problem. Then for each finite input sequence e:*

1. *if σ violates either the lagged Ockham property or stalwartness at e then σ is strongly beaten in terms of retractions at e;*
2. *if σ satisfies stalwartness and the lagged Ockham property at e, then σ is efficient in terms of retractions at e.*

It is clear from the definitions that being strongly beaten implies being weakly beaten which implies inefficiency, so it follows that:

Corollary 51 *Let σ be a solution to the counting problem and let the cost be retractions. Then the following are equivalent:*

1. *σ is efficient at each e;*
2. *σ is weakly beaten at no e;*
3. *σ is strongly beaten at no e;*
4. *σ is stalwart and has the lagged Ockham property at each e.*

So the set of all solutions to the counting problem is neatly partitioned into the efficient solutions and the strongly beaten solutions, where the former are precisely the stalwart, lagged Ockham solutions. That is hardly obvious from the definitions of efficiency and beating, themselves. It reflects a substantive interaction between the criteria of evaluation and convergence to the truth.

Say that a *method* is a constraint on strategies, so the stalwart, lagged Ockham property is a method. Since violating this method results in being beaten at each violation, it follows that no matter what you did in the past, following the stalwart, lagged Ockham method will always look better at each stage (in terms of worst-case retractions) than violating it (given that you aim to converge to the truth). Thus, one may say that the stalwart, lagged

Book_Friend_9781402061264_Proof1_May 24, 2007

Ockham method is *stably retraction efficient* for agents who wish to converge to the truth in a retraction-efficient manner. Stability is crucial for explaining the history of science, for it has frequently occurred that a complex theory is selected because the simple theory has not yet been conceived or has been rejected on spurious grounds (e.g., Ptolemaic astronomy *vs.* Copernican astronomy or wave optics *vs.* Newtonian optics). If Ockham's razor is to explain the subsequent revision to the simpler theory, the rationale for preferring simpler theories must survive past violations.

The preceding results respond to an additional anti-realist challenge. Suppose that you have already seen $n - 1$ marbles at awkward, distant intervals and that after seeing each marble you came to believe, eventually, that you had seen all the marbles there are. The "negative induction" argument against realism [14] recommends the conclusion that one more marble will appear, since you were fooled each time before. But that policy would risk a gratuitous retraction, according to the preceding argument. So the realist wins, no matter how many times Ockham's razor led to disaster in the past![5]

5 TIMED RETRACTIONS

Retraction efficiency does not prohibit a solution from hanging onto its previous answer in spite of the appearance of new marbles, since no retraction is incurred thereby. Mere consistency with experience rules out under-counting, so consistency together with retraction efficiency entails that one never return a value other than the correct count. But that response is not sufficiently general, for suppose that the question is modified so that if the true number of marbles is even, all you have to say is "even".[6] When the first marble is seen, the right answer seems to be 1 rather than "even", but the lagged Ockham property together with consistency does not imply this conclusion, for "even" is consistent with any possible experience.

Here is a more general and unified explanation. Suppose that you hang on to answer "even" to save a retraction when the first marble is seen. Nature can withhold further marbles until you converge to answer 1. The obvious Ockham strategy would drop "even" immediately and would eventually gain enough confidence to say 1 later, so if the answer is 1, both you and the

[5] On the other hand, enough surprises might push one to rethink the problem by adding the answer "infinitely many marbles will appear". The U-turn argument concerns only the problem as presented, not other possible problems one might take one's self to be solving instead.

[6] Worst-case bounds must still be taken over total marble counts rather than over answer "even". The general theory of simplicity developed below works the same way.

Ockham strategy retract once, but you retract later than the Ockham strategy. That is worse, for one's state after the retraction is more enlightened than one's state prior to it (think of the Newtonians before and after they lost their faith that an ether drift would be detected) and needlessly delaying a retraction allows more subsidiary conclusions to accumulate that must be flushed when it finally occurs.

So instead of simply counting retractions, let the cost of inquiry in a given world w be represented by a possibly empty, finite sequence of ascending natural numbers $(r_1, \ldots, r_k)$ such that the strategy retracts exactly k times in w and for each i from 1 to k, the strategy retracts at moment r_i. It is necessary to rank such cost sequences. It would be unfortunate if Ockham's razor were to depend upon some fussy weighting of time against overall retractions so that, say, $(9) > (1, 2)$. Happily, it suffices in the following argument to restrict attention to weak Pareto dominance with respect to overall retractions and the times of occurrence thereof, which yields only a partial order over cost sequences. Accordingly, if c, c' are both cost vectors, let $c \leq c'$ if and only if there exists a sub-sequence d of c' whose length matches that of c such that the successive entries in d are at least as great as the corresponding entries in c. Then define $c < c'$ if and only if $c \leq c'$ but $c' \nleq c$. For example:

$$(1, 3, 8) < (1, 5, 9) < (1, 2, 5, 9).$$

Refer to the cost concept just defined as *timed retractions*.

Next, consider bounds on sets of timed retraction cost sequences. Recall that ω is the least (infinite) ordinal upper bound on the natural numbers. A potential *timed retraction bound* is the result of substituting ω from some point onward in a cost sequence: e.g., $(1, 2, \omega, \omega)$. If S is a set of cost sequences and b is a potential bound, then b bounds S (written $S \leq g$) if and only if for each c in S, $c \leq b$. Thus, $(1, \omega)$ bounds the set of all sequences $(1, k)$ such that k is an arbitrary natural number.

Finally, say that a strategy is *Ockham* just in case it never chooses an answer other than the current count (or possibly '?'). Then one obtains the following, strengthened result.

Proposition 52 *Let a solution to the counting problem be given. Then:*

1. *if the solution violates either the Ockham property or stalwartness at e, then the solution is strongly beaten in terms of timed retractions at e;*
2. *if the solution satisfies stalwartness and the Ockham property at e, then the solution is efficient in terms of timed retractions at e.*

Proof. Suppose that you over or under count at e, which presents exactly k marbles. As before, let hybrid strategy σ be just like you along e_- and then always return the current count from e onward. Consider answer $k + k'$, where k' is an arbitrary natural number. Suppose that you retract at e if σ does. Then the cost sequence for σ along e is no worse than yours, which is, say, $(c_1, \ldots, c_r)$. Then since σ retracts at most once, for each of the additional marbles that appear after e in answer $k + k'$, the worst-case cost bound for σ over answer $k + k'$ is at most $(c_1, \ldots, c_r, \omega, \ldots, \omega)$, with k' repetitions of ω. Nature can withhold marbles after e until you eventually retract your answer (say, at stage i) in preparation for convergence to k. Furthermore, after you converge to k, nature can continue to withhold marbles until you say k an arbitrary number of times before presenting another marble. Eventually, you drop k in preparation for convergence to $k + 1$, etc. So your bound in answer $k + k'$ is at least $(c_1, \ldots, c_r, i, \omega, \ldots, \omega)$, with k' repetitions of ω. That is worse than the bound for σ because the bound for σ is a proper sub-sequence of your bound.

Now suppose that you don't retract at e but σ does. Then let your cost through e be $(c_1, \ldots, c_r)$, in which case the cost of σ through e is $(c_1, \ldots, c_r, i)$, where i is the length of e. Then since σ retracts at least once, arbitrarily late, for each of the additional marbles that appear after e in answer $k + k'$, the worst-case cost bound for σ over answer $k + k'$ is at most $(c_1, \ldots, c_r, i, \omega, \ldots, \omega)$, with k' repetitions of ω. But since you do not produce k at the end of e, nature can withhold marbles until (say, at stage $i' > i$) you retract your answer at e in preparation for convergence to k. Then nature can exact one retraction out of you, arbitrarily late, for each of the k' marbles that appears after e in answer $k + k'$. Hence, your worst case bound is at least $(c_1, \ldots, c_r, i', \omega, \ldots, \omega)$, where $i' > i$. So your bound is worse than that of σ. The beating argument for stalwartness and the efficiency argument for stalwart, Ockham solutions are similar.[7] ■

So when retraction delays are taken into account, every solution is either efficient, stalwart, and Ockham or strongly beaten. Again, there is no middle ground.

Corollary 53 *Let σ be a solution to the counting problem and let the cost be timed retractions. Then the following are equivalent:*[8]

1. *σ is efficient at each e;*
2. *σ is weakly beaten at no e;*

[7] In any event, more general arguments are provided in the appendix for propositions 56 and 58 below.

[8] Corollary 53 is an instance of corollary 59 below.

3. σ is strongly beaten at no e;
4. σ is stalwart and Ockham at each e.

6 GENERALIZING THE ARGUMENT

In order to argue, in general, that Ockham's razor is necessary for minimizing timed retractions, one must say, in general, what Ockham's razor amounts to. That may seem like a tall order compared to counting marbles. First, simplicity has such manifold characteristics—e.g., uniformity, unity, testability, and reduction of free parameters, causes, or ontological commitments—that one wonders if there is a single notion that underlies them all. Second, it seems that some aspects of simplicity are a mere matter of description. For example, if one describes inputs as marbles or non-marbles, then marble-free worlds are most uniform. But if an "*n*-ble" is a marble at each time other than *n*, when it is a non-marble, then uniformly marble-free experience is not uniformly *n*-ble free experience. Nor can one complain that the definition of *n*-ble is strange, since marbles are *n*-bles at each stage but *n*, when they are non-*n*-bles Goodman 1983 [6]. So with respect to the syntactic complexity of definitions, the situation is entirely symmetrical. These sorts of observations have led to widespread skepticism about the prospects for a general, unified, objective account of simplicity . But the skepticism is premature, for in the marble counting problem, the question at hand concerns marbles rather than *n*-bles and simplicity may depend upon the structure of the problem one is trying to solve. Indeed, if simplicity is to have anything to do with efficiency, it must somehow reflect the structure of the problem one is trying to solve.

In the marble counting problem, answers positing more marbles are more complex. Presumably, then, worlds that present more marbles are more complex, assuming that simpler answers are answers satisfied by simpler worlds. One might plausibly say that each marble is an *anomaly* relative to the counting problem, since the previously simplest (best) explanation is no longer simplest after the marble appears. Some insight is gained into the nature of anomalies by characterizing the occurrence of a marble entirely in terms of the structure of the marble counting problem, itself.[9]

One structural feature of the marble counting problem is that, prior to seeing a third marble, nature can *force* an arbitrary solution to the problem to produce successive answers 2, 3, 4, ... by presenting no marbles until the

[9] For a critique of this idea and a response, see [1] and [19].

solution converges to 2, one marble followed by no more until the method converges to 3, and so forth. But after seeing the third marble, nature can only force the solution to produce successive answers 3, 4, 5, ... So as a working hypothesis, it seems that an anomaly occurs when the sequence of answers nature can force is truncated (from the front). This might be expressed by saying that an anomaly occurs when nature uses up an opportunity to force the scientist to change her mind or, more colorfully, when nature leads the scientist one exit further down the freeway to the truth.

Nature may be capable of taking more than one step down the freeway at a time (e.g., modify the marble counting problem so that several marbles can be emitted at one time), in which case nature takes two steps down the forcible path $(0, 1, 2, \ldots)$ when two marbles are presented at one time, for after these marbles are seen, only $(2, 3, 4, \ldots)$ is forcible.

Also, there may be more than one freeway to the truth, in which case there may be several simplest answers to select among. For example (Figure 4.5), modify the counting problem so that marbles come in two colors, white and black, and you have to determine (i, j), where i is the total number of white marbles and j is the total number of black. If no marbles have been seen so far, then patterns of form $((0, 0), (1, 0), \ldots)$ and $((0, 0), (0, 1), \ldots)$ are forcible. Suppose you now hear a rumble in the machine, which guarantees that another marble is coming, but you don't see the color. Now $(0, 0)$ is no longer forcible (the rumble can't be "taken back") so only patterns of form $((1, 0), \ldots)$ and $((0, 1), \ldots)$ are forcible. That is a step by nature down both possible paths, so the rumble constitutes an anomaly. Suppose that the announced marble is black. Now only patterns of form $((0, 1), \ldots)$ are forcible. No step is taken down path $((0, 1))$, however, so seeing the black marble after hearing the noise is not an anomaly— intuitively, the anticipated marble has to have some color or other. The same is true if a black marble is seen. So no world in which just one more marble is seen presents any anomalies after the sound, so all such worlds are maximally simple in light of

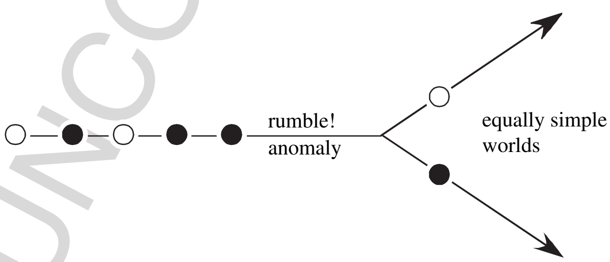

Figure 4.5. nature chooses a path without stepping down it

the sound. Hence, answers $(1, 0)$ and $(0, 1)$ are both simplest after the sound, whereas answers that entail more than one marble are more complex than necessary. That is intuitive, since Ockham's razor seems to govern number rather than color in this example.[10]

As it is usually formulated, Ockham's razor requires that one never presume a more complex hypothesis than necessary, which allows for selection among simplest answers when the noise is heard: e.g., $(3, 3)$ over $(2, 4)$. Answers positing extra marbles—e.g., $(3, 4000)$ are plausibly ruled out. But there is still something odd about guessing one color rather than another before seeing what the color is: after the noise, it seems that one should simply wait to see what color the announced marble happens to be. Indeed, the problem's future structure is entirely symmetrical with respect to color, so there could be no efficiency advantage in favoring one color over another until one sees which color it is. Say that a method has the *symmetry* property at a given stage if it produces no answer other than the uniquely simplest answer at that stage.

In the counting problem, worst-case cost bounds were assessed over possible answers to the question. In the general theory presented below, worst-case bounds are assessed over complexity classes of worlds. One reason for this is to "break up" coarse answers sufficiently to recover the U-turn argument. For example, recall the problem in which you must count the marbles if the total count is odd and must return "even" if the total count is even. In this problem, Ockham violators are not necessarily strongly beaten because the retractions of an arbitrary solution are unbounded in answer "even", both in terms of retractions and in terms of timed retractions. In the general theory, the answer "even" is partitioned into anomaly complexity classes corresponding to each possible even count and retractions are bounded over these complexity classes so that the strong beating arguments rehearsed earlier for the counting problem can be lifted to this coarser problem. This agrees with standard practice in the theory of computational complexity, in which one examines an algorithm's worst-case resource consumption over sets of inputs of equal size [4].

[10] That is because color does not lead to unavoidable retractions in the example under discussion. If each white marble could spontaneously change color, just once, from white to black at an arbitrary time after being emitted, then white would be simpler than black. The same is true if a continuum of gray-tones between white and black is possible and marbles never get brighter. Then Ockham should say "presume no more darkness than necessary".

7 EMPIRICAL SIMPLICITY DEFINED

It remains to state the preceding ideas with mathematical precision. An *empirical problem* is a pair (K, Π), where K is a set of infinite sequences of inputs and Π partitions K. Elements of K are called *worlds* and cells in Π are called *potential answers*. A scientific *strategy* is a mapping from finite sequences of inputs to answers in Π (or to "?", signalling a refusal to choose). A *solution* is a strategy that converges to the true answer in each world in K. Let K_e denote the set of all elements of K that extend finite input sequence e and let Π_e denote the set of all answers A in Π such that A is compatible with e (i.e., such that K_e shares an element with A). Finally, say that e is *compatible* with K just in case some world in K extends e.

All of the following definitions are relative to a given problem (K, Π), which is suppressed to avoid clutter. Say that an *answer pattern* is a finite sequence of answers in which no answer occurs immediately after itself. Let g be an answer pattern. The *g-forcing game* given finite input sequence e compatible with K is played between the scientist and nature as follows. The scientist plays an answer (or "?"), nature plays an input, and so forth, forever.[11] In the limit, the two players produce an infinite play sequence p, of which p_N is the infinite subsequence played by nature and p_S is the infinite subsequence played by the scientist. Let i be the length of e and let $p_S - i$ denote the result of deleting the first i entries from the beginning of p_S. Then nature wins the game if and only if p_n is in K_e and either p_N does not converge to the answer true in p_N or g is a subsequence of $p_S - i$.

Strategies for the scientist have already been defined. A strategy for nature maps finite sequences of answers (or "?") to inputs. A strategy for the scientist paired with a strategy for nature determines a play sequence. A strategy is *winning* for a player if it wins against an arbitrary strategy for the other player. Say that g is *forcible* given e if and only if nature has a winning strategy in the g-forcing game given e. The g-forcing game is *determined* just in case one player or the other has a winning strategy. The assumption of determinacy for forcing games is so useful formally that I will restrict attention to such problems.

Restriction 1 (determinacy of forcing games) *The following results are restricted to problems such that for each pattern g, the g-forcing game is determined.*

[11] Cf. (Kechris 1991) for a general introduction to the pivotal role of infinite games in descriptive set theory.

The restriction turns out not to matter in typical applications, for D. Martin's Borel determinacy theorem (1975) has the following consequence:

Proposition 54 (determinacy of Borel forcing games) *If* (K, Π) *is solvable and if* K *is a Borel set and* e *is a finite input sequence, then for all answer patterns* g*, the* g*-forcing game in* (K, Π) *is determined given* e*.*

Since unsolvable problems are irrelevant to the results that follow, it suffices for determinacy of forcing games to assume that K is Borel. That is weaker than saying that K can be stated with some arbitrary number of quantifiers over observable predicates, which covers just about any empirical problem one might encounter in practice.[12] The antecedent of the proposition is not a necessary condition for the consequent, so the scope of the following results is broader still.

Say that answer pattern g is *backwards-maximally forcible* at e if and only if g is forcible given e and for each forcible answer pattern g' given e, if g is a sub-sequence of g' then g is an initial segment of g. Let $\triangle_e$ denote the set of all answer patterns that are backwards-maximally forcible at e. The backwards-maximality property is crucial to the results that follow. The point is to eliminate gaps from all the sequences in $\triangle_e$. For example, in the marble counting problem, if e presents no marbles, then $\triangle_e$ looks like:

()
(0)
(0, 1)
(0, 1, 2)
(0, 1, 2, 3)
⋮

whereas the forcible sequences include all (gappy) sub-sequences of these, such as (4, 7, 9).

It is not necessarily the case that each forcible pattern b at e can be extended to a backwards-maximally forcible pattering at e. For example, suppose that tomorrow you may see any number of marbles and that any of the marbles may disappear at any time thereafter. At the outset, each finite, descending sequence of marble counts is forcible, so each forcible pattern can be extended at the beginning to a forcible pattern. The following formal development is simplified by, frankly, ignoring such problems.

[12] A typical sort of K (e.g., for marble counting and for inferring polynomial degree) says that there exists a stage such that for each later stage, no further empirical effects are encountered. That involves only two quantifiers, so the restriction is easily satisfied.

Book_Friend_9781402061264_Proof1_May 24, 2007

Restriction 2 (well-foundedness of forcibility) *If pattern b is forcible at e, then there exists pattern b′ of which b is a sub-pattern such that b′ is in* $\triangle_e$.

One would expect that if (A, B, C) is in $\triangle_e$, then there should be further experience e' such that (B, C) is in $\triangle_e$; but that is not necessarily the case.[13] It simplifies the following theory to ignore those cases as well. Let $*$ denote concatenation.

Restriction 3 (graceful decrementation) *If* $A * B * c$ *is in* $\triangle_e$, *then there exists proper extension e′ of e compatible with K such that* $B * c$ *is in* $\triangle_{e'}$ *and exactly one anomaly occurs along e′ properly after the end of e.*

If g is an answer pattern, let $g * \triangle_e$ denote the set of all $g * g'$ such that g' is an element of $\triangle_e$. Say that an *anomaly* occurs at finite, non-empty input sequence e compatible with K if and only if there exists a non-empty, finite answer pattern $A * g$ such that:

1. $A * g * \triangle_e \subseteq \triangle_{e\_}$;
2. no g' in $\triangle_e$ begins with answer A.

Suppose that two marbles are seen simultaneously at stage e in the counting problem. This anomaly is represented in figure (Figure 4.6).The fact that answer pattern $A * g$ is non-empty ensures that nature moves down some path in $\triangle_e$. Thus, seeing a black marble is not an anomaly after the noise that announces it.

If w is a world in K, then let $c(w, e)$ denote the number of anomalies that occur along w properly after e. If A is an answer, let $c(A, e)$ denote the least $c(w, e)$ such that w is in $K_e \cap A$. Call $c(w, e)$ the *conditional anomaly complexity* of w (or of A) given e, and similarly for $c(A, e)$.

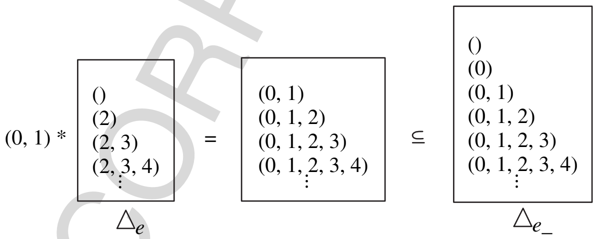

Figure 4.6. simultaneous observation of two marbles

[13] Suppose that you have to determine the total number of marbles and the time of the first marble if the total count is 2. If no marbles appear in e yet, then we have that (0, 1, 3) is in $\triangle_e$. But upon seeing the first marble at stage k in e', $(1, (2, k), 3)$ is in $\triangle_{e'}$, so (1, 3) is not in $\triangle_{e'}$.

Then let *unconditional* anomaly complexity be given by $c(w) = c(w, ())$ and $c(A) = c(A, ())$, where () is the empty input sequence.

Marbles are still anomalies in the marble-counting problem, but the preceding definitions don't see the marbles; they see only the structural "shadow" each marble occurrence casts against the branching topology of the marble counting problem. The noise announcing a marble is anomalous, but seeing a marble after the noise is not. Seeing two marbles after the noise is anomalous, however. If several marbles are visible and some of them might disappear permanently at any time, then disappearances of marbles count as anomalies and simple worlds have more marbles than complex ones. Refutations of lower polynomial degrees and the discovery that a linear function depends upon an independent variable also count as anomalies in the corresponding problems (assuming that the data consist of ever-tighter open intervals around the dependent variable).

8 OCKHAM'S RAZOR, SYMMETRY, AND STALWARTNESS

Answer A is *simplest* at e if and only if

$$c(A, e) = \min_{B \in \Pi_{\mathrm{P}i}} c(B, e).^{14}$$

A method satisfies *Ockham's razor* at e just in case the answer output by the method at e is '?' or is simplest at e. *Symmetry* at e requires that the method output at e either "?" or the unique answer that minimizes $c(A, e)$. *Stalwartness* at e requires that if the scientist's output A at e_- is uniquely simplest at e, then the scientist produces A also at e.

Ockham's razor may be defined in terms of simplicity rather than complexity, using a standard rescaling trick familiar from information theory. Define *conditional simplicity* as:

$$s(A, e) = \exp(-c(A, e)).$$

This definition reveals an interesting connection between Ockham's razor and Bayesian updating, for it follows immediately from the definition of $c(A, e)$ that:

$$c(A, e) = c(A \cap K_e) - c(K_e).$$

[14] In light of lemma 67 in the appendix, this condition is equivalent to $c(A, e) = 0$.

Applying the definition of $s(A, e)$ to both sides of the preceding equation yields:

$$s(A, e) = \frac{s(A \cap K_e)}{s(K_e)},$$

which is the usual definition of Bayesian updating. Then Ockham's razor requires that one choose the uniquely simplest hypothesis, where simplicity degree is updated by conditionalization. Nothing about coherence or probability has been presupposed, however, so Bayesians who seek Ockham's razor in prior probabilities updated by conditioning put the arbitrary cart before the essential horse.

9 SYMMETRICAL SOLVABILITY

Not every problem has a symmetrical solution. For example, suppose that the problem is to say not only how many marbles appear, but when each of them appears. In this problem, every answer compatible with e is simplest at e, since only patterns of unit length are forcible. That may seem counter-intuitive, since particle counts are analogous to free parameters and times of appearance are analogous to settings of those parameters, so it would seem that answers involving more free parameters are more complex. But it must be kept in mind that the same possibilities could be parameterized in different ways, and simplicity depends upon which parametrization the question asks about. If the problem is to count marbles, then worlds with more marbles are more complex, whenever the marbles arrive. If it is to count n-bles, then worlds with more n-bles are more complex, regardless of when the marbles arrive. If the problem is to identify particular worlds, the parametric structure of the problem disappears and complexity is flattened. Examples of the latter sort are excluded from consideration by the following restriction.

Restriction 4 (symmetrical solvability) *Only problems with symmetrical solutions are considered in the results that follow.*

In typical applications, restriction 4 can be sidestepped by coarsening or refining the question in a manner that disambiguates the intended parametrization. It is also worth mentioning that restrictions 2 and 4 are logically independent given restriction 1.[15]

[15] The problem of identifying individual worlds in which at most finitely many marbles occur satisfies the determinacy assumption (restriction 1) and the well-foundedness assumption (restriction 2) but not the symmetrical-solvability assumption (restriction 4), whereas

10 EFFICIENCY DEFINED

Let $C_e(n)$ denote the set of all worlds in K_e such that $c(w, e) = n$. Refer to $C_e(n)$ as the nth *anomaly complexity class* at e.[16] Complexity classes depend only on the structure of the problem to be solved, so they are not mere matters of description.

Let σ be a solution to (K, Π) and let e be compatible with K. Let the worst-case timed retractions over $C_e(i)$ be the supremum of the timed retraction costs incurred by σ over worlds in $C_e(i)$. As mentioned above, the idea is to examine worst-case bounds over anomaly complexity classes rather than over answers. Accordingly, define:

1. solution σ is *efficient* at e with respect to a given cost if and only if for each solution σ' that agrees with σ along e_- and for each n, the worst case cost bound of σ over $C_e(n)$ is less than or equal to that of σ';
2. solution σ is *strongly beaten* at e with respect to a given cost if and only if there exists solution σ' that agrees with σ along e_- such that for each n such that $C_e(n)$ is non-empty, the worst case cost bound of σ over $C_e(n)$ is greater than that of σ';
3. solution σ is *weakly beaten* at e with respect to a given cost if and only if there exists solution σ' that agrees with σ along e_- such that for each n, the worst case cost bound of σ' over $C_e(n)$ is less than or equal to that of σ and there exists n such that the worst-case cost bound of σ' over $C_e(n)$ is less than that of σ.

Notice that there is no imposed bias or weighting, probabilistic or otherwise, in favor of lower complexity classes or simple worlds in the preceding definitions. There are just dominance relations over worst-case bounds on structurally motivated complexity classes. That is as it must be if the efficiency argument for Ockham's razor is to avoid the narrow circularity of standard, Bayesian explanations.

11 NESTED PROBLEMS

The marble counting problem and the problem of finding the true polynomial degree of a curve both have the attractive feature that there exists

the disappearing marble example described earlier satisfies restrictions 1 and 4 but not restriction 2, for a symmetrical solution could simply wait until tomorrow to see how many marbles there are and could then guess the current number of marbles at each stage.

[16] The complexity classes are sets (subsets of K).

a uniquely simplest answer for each possible evidential circumstance e. But there may be more than one maximally simple answer, as in the black and white marble counting problem when the noise is heard. Accordingly, say that a problem is *nested* if there exists a uniquely simplest answer at each e compatible with K. Nested problems allow for branching paths, but have the property that there is a uniquely simplest answer at each stage of inquiry, as in the two-color counting problem when no noise is heard prior to seeing the marble. In that case, nature can choose which color to present at each stage, but the current count is always the uniquely simplest answer. Other familiar scientific questions with this structure include finding the set of all independent variables a linear equation depends upon and the inference of conservation laws in particle physics [18].

12 THE MAIN RESULTS

For brevity, these assumptions govern all the results that follow. All proofs are presented in the appendix.

1. (K, Π) is a problem satisfying restrictions 1–4;
2. the cost under consideration is timed retractions;
3. e is a finite input sequence compatible with K.

The main result is that, in general, every deviation from Ockham's razor incurs a strong beating. Hence, the argument for Ockham's razor is stable, in the sense that you always have a motive to return to Ockham's fold no matter how prodigal you have been in the past.

Proposition 55 (efficiency stably implies Ockham's razor) *If solution σ violates Ockham's razor at e, then σ is strongly beaten in terms of timed retractions at e.*

The same is true of stalwartness.

Proposition 56 (efficiency stably implies stalwartness) *If solution σ violates stalwartness at e, then σ is strongly beaten in terms of timed retractions at e.*

Symmetry is a stronger principle than Ockham's razor and its general vindication is correspondingly weaker: violating symmetry results in a weak beating at the first violation rather than a strong beating at each violation.[17]

[17] For example, suppose at e that a curtain will be opened tomorrow that reveals either a marble emitter or nothing at all. The question is whether there is an emitter behind the

Proposition 57 (efficiency implies symmetry) *If solution σ violates symmetry at e, then σ is weakly beaten at the first moment e' along e at which symmetry is violated.*

Again, being beaten is no sin if every solution is beaten. To clinch the argument, stalwart, symmetrical solutions are efficient. That amounts to an existence proof, given that the problem is symmetrically-solvable, since every symmetrically solvable problem is solvable by a stalwart, symmetrical method.[18] The efficiency is also stable if the problem under consideration is nested.

Proposition 58 (symmetry and stalwartness imply efficiency)

1. *If the problem is nested and σ is a stalwart, Ockham solution from e onward, then σ is efficient at e.*
2. *If σ is a stalwart, symmetrical (and, hence, Ockham) solution at every stage, then σ is efficient at every stage.*

In nested problems, all solutions are partitioned into the strongly beaten ones and the stalwart Ockham ones. This duplicates the situation in the counting problem.

Corollary 59 *If the problem is nested and σ is a solution, then the following statements are equivalent:*

1. *σ is efficient at each e;*
2. *σ is weakly beaten at no e;*
3. *σ is strongly beaten at no e;*
4. *σ is stalwart and Ockham at each e.*

More generally, the possibility of weakly beaten, non-symmetrical methods must be allowed.

curtain and if so, how many marbles it will emit. The no-emitter world and the marble-free emitter world are both simplest in this example, so symmetry requires that one suspend judgment between the corresponding answers until the curtain is opened. Suppose that you flout symmetry and guess that you are in the marble-free emitter world. Had you refrained from choosing, you would have had no retractions in complexity class $C_e(0)$, but you have incurred at least one retraction in class $C_e(0)$, so you are weakly beaten (every solution, including you, retracts at least k times after e in the worst case in class $C_e(k)$). You are not strongly beaten, however, because you do as well as possible in each class $C_e(k)$ such that k exceeds zero.

[18] For a symmetrical solution converges to the uniquely simplest answer in each world and is not prevented from doing so by hanging onto a uniquely simplest answer until it is no longer uniquely simplest.

Corollary 60 *If σ is a solution, then the following statements are equivalent.*

1. *σ is efficient at each e;*
2. *σ is weakly beaten at no e;*
3. *σ is stalwart and symmetrical (and, hence, Ockham) at each e.*

13 CONCLUSION AND PROSPECTS

A very general, structural theory of simplicity and of Ockham's razor has been presented, according to which Ockham's razor does not point at the truth, but keeps one on the most direct route thereto. Indeed, choosing only the uniquely simplest hypothesis compatible with experience and hanging onto it until its uniquely simple status is undermined is demonstrably equivalent to minimizing timed retractions prior to convergence to the truth. This result provides a relevant, non-circular connection between simplicity and finding the true theory. No standard, alternative account of simplicity does so.

The results suggest that the scientific realism debate is not a genuine debate. The anti-realist is correct that simplicity cannot function as a magical divining rod for truth. The realist is correct that simplicity, nonetheless, provides the best possible advice for finding the truth, because it keeps one on the straightest possible path thereto. The results also provide some solace for scientists who employ off-the-shelf data-mining procedures that employ a wired-in prior bias toward simplicity. Such methods really are more efficient at finding the truth, even though they cannot be said to divine or point at the truth.[19] Finally, the results reverse the common impression that convergence considerations impose no constraints on the course of inquiry in the short run. It has been demonstrated that timed retraction efficiency leaves just one choice open to a convergent scientist: how long to wait for evidence to accumulate before leaping to the uniquely simplest hypothesis in light of the data. Which answer to choose and when to drop it are both uniquely determined.

Like all new ideas, the proposed account of Ockham's razor suggests a range of potential improvements and generalizations. (1) Efficiency with respect to total number of erroneous answers produced prior to convergence is equivalent to the symmetry principle and, hence, entails Ockham's razor.

[19] Simulation studies suggesting the contrary notwithstanding. When an Ockham procedure seems to have a higher chance of producing the true answer in a randomly chosen example than non-Ockham procedures, the underlying sampling distribution over worlds is biased toward simple worlds. That is just a motorized version of the circular Bayesian argument.

The same is true if efficiency is defined in terms of weak Pareto-dominance with respect to timed retractions and errors jointly. Other combinations of costs can be considered. (2) Penalizing total retracted content rather than just retractions yields the intuitive result that one should only retract to "one black or one white" when the noise announcing a new marble is heard. (3) It remains to apply the preceding ideas with equal rigor and generality to statistical and causal inference (see [12] for some preliminary ideas). (4) It also remains to explore realistic recommendations when finding the Ockham hypothesis is computationally infeasible (see [11] for more preliminary ideas). (6) Finally, the symmetrical solvability and well-foundedness restrictions can and should be weakened.

14 APPENDIX

In the following results, (K, Π) is assumed to be an empirical problem satisfying restrictions 1–4, and e, e' range over finite input sequences. Also, let $\omega[k]$ denote the sequence $(\omega, \ldots, \omega)$ in which ordinal ω is repeated exactly k times.

Proof of proposition 54. Let p be a play sequence in the g-forcing game in problem (K, Π) at e. Let p_S be the sub-sequence consisting of the scientist's plays, and let p_N be the corresponding sub-sequence for nature. Let W be the winning condition for nature. In light of Martin's (1975) theorem, it suffices to show that W is a Borel set. Then $p \in W$ if and only if:

1. $p_N \in K_e$ and
2. (a) $\neg((\exists n)(\forall m \geq n)\ p_S(m) \neq$ '?' and $p_N \in p_S(m))$ or
   (b) g is a sub-sequence of p_S.

Condition $p_N \in K_e$ is Borel because K is assumed to be Borel and the condition of extending e is clopen. Condition $p_S(m) \neq$ '?' is clopen. Since (K, Π) is solvable, each cell in Π is Σ^0_2, since w is in answer A if and only if there exists a time such that for each later time the solution converges to A. Hence, the condition that $p_N \in p_S(m)$ is Σ^0_2 Borel. Finally, the condition that g is a sub-sequence of p_S is open. Borel conditions are preserved under first-order quantification and Boolean connectives, so W is Borel. ■

Proof of proposition 55. Let σ be a solution that violates Ockham's razor at e (which need not be the first violation). So $\sigma(e) = A$, where A is not a simplest answer compatible with e. Let σ' agree with σ along e and then produce the simplest answer compatible with e' if it exists and '?'

otherwise, for each e' properly extending e. Since (K, Π) is symmetrically solvable (restriction 4), σ' solves (K, Π), because σ' converges, in each world, to whatever the assumed symmetrical solution converges to in that world. Let r be the timed retraction cost common to both methods σ and σ' along e_- (recall that r is a finite, ascending sequence of natural numbers).

Suppose that $C_e(k)$ is non-empty. There exists a pattern $B * b$ of length at least $k + 1$ in $\triangle_e$ (by lemma 63). Since A is not a simplest answer, $B \neq A$ (by lemma 65). There exists w in $B \cap K_e$ along which $B * b$ remains forcible after e (by lemma 64). Since σ is a solution, σ retracts A after e along w, say at e' of length j. Now $B * b$ is still forcible given e', so there exists w' in $C_{e'}(k)$ along which σ can be made to repeat each successive entry in $B * b$ an arbitrary number of times (by lemma 69). Since $B * b$ is forcible at e' and $B * b$ is in $\triangle_e$, no anomaly occurs along e' after e (by lemma 61). Hence, w' is in $C_e(k)$. So the worst-case timed retraction bound for σ over $C_e(k)$ is at least $r * j * \omega[k]$, where it will be recalled that $\omega[k]$ denotes the sequence $(\omega, \ldots, \omega)$, with ω repeated k times and $*$ indicates concatenation. But since σ' retracts after e only at anomalies (by lemma 68), the worst-case timed retraction bound for σ' over $C_e(k)$ is at most $r * i * \omega[k]$, where $i < j$ is the length of e. Since $r * i * \omega[k] < r * j * \omega[k]$ and C_k is an arbitrary, non-empty complexity class, σ' strongly beats σ at e in terms of timed retractions. ■

Proof of proposition 56. Let σ be a solution that violates stalwartness at e (which need not be the first violation). So for some answer A that is uniquely simplest at e, $\sigma(e_-) = A$ but $\sigma(e) \neq A$. Let σ' be a solution constructed as in the proof of proposition 55, and let r be the timed retraction cost incurred along e_- by both σ and σ'. Let i be the length of e. Then σ incurs timed retraction cost $r * i$ along e, but σ' incurs only r. Let $C_e(k)$ be non-empty. So there exists a pattern b in $\triangle_e$ of length at least $k + 1$ (by lemma 63). There exists w in $C_e(k)$ along which σ can be made to repeat each successive entry in b an arbitrary number of times (by lemma 69). So the worst-case timed retraction bound for σ over $C_e(k)$ is at least $r * i * \omega[k]$. Since $\sigma'(e_-) = A$ and σ' is stalwart at e and A is simplest at e, $\sigma'(e) = A$, so the timed retraction cost of σ' along e is just r. Since σ' retracts after e only at anomalies (by lemma 68), the worst-case timed retraction bound for σ' at e is at most $r * \omega[k]$. Since $r * \omega[k] < r * i * \omega[k]$ and C_k is an arbitrary, non-empty complexity class, σ' strongly beats σ at e in terms of timed retractions. ■

Proof of proposition 57. Suppose that σ is a solution that violates the symmetry principle (somewhere). Then there exists finite input sequence

e compatible with K such that σ violates symmetry at e, but not at any proper sub-sequence of e. So $\sigma(e) = A$, where A is not the uniquely simplest answer compatible with e. Let σ', r, and $\omega[k]$ be as in the proof of proposition 55.

Since A is not uniquely simplest at e, there exists world w in $C_0(e)$ such that w satisfies some answer $B \neq A$ (by lemma 67). Since σ is a solution, σ converges to B in w, so there exists some e' properly extending e and extended by w such that $\sigma(e') \neq A$. So the timed retractions of σ along e' are at least $r * j$, where j is the length of e'. So the worst case timed retractions of σ over $C_e(0)$ are at least $r * j$. Let w' be an arbitrary element of $C_e(0)$. Then σ' never retracts in w after e (by lemma 68). It is possible that σ' retracts at e. So the worst case timed retractions of σ' over $C_e(0)$ are less than or equal to $r * i$, where $i < j$ is the length of e. Observe that $r * i < r * j$.

Now consider non-empty complexity class $C_e(k)$, for arbitrary $k \geq 0$ and let w be in $C_e(k)$. Then there exists pattern b in $\triangle_e$ of length at least $k + 1$ (by lemma 63).

Case A: σ retracts at e if σ' does. Then the worst case timed retractions of both methods along e are exactly the same, say r', and the worst-case timed retraction bound for σ' over $C_e(k)$ is no worse than $r' * \omega[k]$. Also, there exists w' in $C_e(k)$ along which σ produces the successive entries along b after e with arbitrarily many repetitions (by lemma 69). Hence, the worst-case timed retractions of σ after e are at least as bad as $\omega[k]$, so the worst-case timed retraction bound for σ over $C_e(k)$ is at least $r' * \omega[k]$. But since σ' retracts after e only at anomalies (by lemma 68), the worst-case timed retraction bound for σ' over $C_e(k)$ is at most $r' * \omega[k]$.

Case B: σ' retracts at e and σ does not. Since e is the first symmetry violation by σ and $\sigma(e_-) = \sigma(e)$, answer $A = \sigma(e)$ is uniquely simplest at e_- but not at e. So there exists w in $C_e(0) - A$ such that w is not in $C_{e_-}(0)$ (by lemma 67). So $c(w, e) = 0$ but $c(w, e_-) > 0$. Hence, e is an anomaly. So there exists pattern $B * d$ such that no pattern in $\triangle_e$ begins with B and $B * d * \triangle_e \subseteq \triangle_{e_-}$. Since the uniquely simplest hypothesis A at e_- begins each forcible sequence in $\triangle_{e_-}$ (by lemma 66), $B = A$, so no pattern in $\triangle_e$ begins with A. So pattern b begins with some answer $D \neq A$. So there exists world $w' \in D \cap K_e$ such that for each e' extending e and extended by w', b is forcible at e' (by lemma 62). Since σ is a solution, σ converges to D in w' and, hence, retracts A at some e' properly extending e and extended by w'. Let j be the length of e', so $j > i$, where i is the length of e. Then b is still forcible at e', so there exists w'' in $D \cap C_{e'}(k)$ along which the successive entries in b are produced with arbitrary repetitions (by lemma 69). Since b is still forcible at e' and b is in $\triangle_e$, no anomalies occur after e along e' (by lemma 61), so w'' is also in $C_e(k)$. Hence, the worst-case timed retraction bound for σ over $C_e(k)$ is at least $r * j * \omega[k]$. But since σ' retracts after e only at anomalies (by lemma 68), the worst-case timed retraction bound for σ' over $C_e(k)$ is at most $r * i * \omega[k] < r * j * \omega[k]$. ■

Proof of proposition 58.1. Let σ' be a solution to a nested problem that is Ockham and stalwart from e onward. Since the problem is nested, σ' is also symmetrical from e onward. Let σ agree with σ' along e_{-}. Suppose that $C_e(k)$ is non-empty. There exists a pattern b of length at least $k + 1$ in $\triangle_e$ (by lemma 63).

Case A: σ retracts at e if σ does. Then let r denote the identical costs of σ and σ' along e. Since σ' is symmetrical and stalwart from e onward, the worst-case timed retraction bound for σ' over $C_e(0)$ is less than or equal to $r * \omega[k]$ (by lemma 68). There exists w' in $C_e(k)$ along which σ can be made to repeat each successive entry in b an arbitrary number of times (by lemma 69), so the worst-case timed retraction bound for σ over $C_e(0)$ is at least $r * \omega[k]$.

Case B: σ' retracts at e and σ does not. Since (K, Π) is nested, there exists a uniquely simplest answer B at e. So every pattern in $\triangle_e$ begins with B (by lemma 66), so b begins with B. Let i be the length of e. Then since σ' retracts only at anomalies after e (by lemma 68), the worst-case timed retraction bound for σ' over $C_e(k)$ is less than or equal to $r * i * \omega[k]$. Since σ' is stalwart at e and retracts at e, answer $A = \sigma'(e_{-}) = \sigma(e_{-})$ is not uniquely simplest at e, so $A \neq B$. There exists w in $B \cap K_e$ along which b remains forcible after e (by lemma 64). Since σ is a solution, σ must retract A in w after e, say by e'. Now b is still forcible given e', so there exists w' in $C_{e'}(k)$ along which σ can be made to repeat each successive entry in b an arbitrary number of times (by lemma 69). Since b is forcible at e', no anomaly occurs along e' after e (by lemma 61). Hence, w' is in $C_e(k)$. So letting $j > i$ be the length of e', the worst-case timed retraction bound for σ over $C_e(k)$ is at least $r * j * \omega[k] > r * i * \omega[k]$. ■

Proof of proposition 58.2. Let σ' be a stalwart, symmetrical solution at every e. Let σ agree with σ along e_{-}. Now consider non-empty complexity class $C_e(k)$, for arbitrary $k > 0$ and let w be in $C_e(k)$. Then there exists pattern b in $\triangle_e$ of length at least $k + 1$ (by lemma 63).

Case A: σ retracts at e if σ' does. Follow the argument for case A in the proof of proposition 57, observing that a stalwart, symmetrical solution retracts only at anomalies (by lemma 68).

Case B: σ' retracts at e and σ does not. Since σ' is always symmetrical and stalwart and σ' retracts at e, answer $A = \sigma'(e_{-}) = \sigma(e_{-})$ is uniquely simplest at e_{-} but not at e. Pick up from here in case B of the proof of proposition 57, again observing that a stalwart, symmetrical solution retracts only at anomalies (by lemma 68). ■

Proof of corollary 59. (1) implies (2) implies (3) by definition. (3) implies (4) by propositions 55 and 56. (4) implies symmetry and stalwartness since the problem is nested. Symmetry and stalwartness imply (1) by proposition 58.1. ■

Proof of corollary 60. (1) implies (2) by definition. (2) implies (3) by propositions 57 and 56. (3) implies (1) by proposition 58.2. ■

Lemma 61 (anomaly freedom) *Let b be in $\triangle_e$ and let b be forcible at e' properly extending e. Then for all e'' properly extending e and extended by e':*

1. *b is in $\triangle_{e''}$ and*
2. *e'' is not an anomaly.*

Proof. Suppose that b is in $\triangle_e$ and b is forcible at e' properly extending e. Let e'' properly extend e and be extended by e'. Then b is forcible at e'' since b is still forcible at e'. Suppose for contradiction that b is not in $\triangle_{e''}$. Then since b is forcible at e'', there exists b' forcible at e'' such that b is a subsequence of b' but b is not an initial segment of b'. But then b' is forcible at e, so b is not in $\triangle_e$. Contradiction. So b is in $\triangle_{e'}$. Again, let e'' be an arbitrary input sequence properly extending e and extended by e'. Then it has just been shown that b is in both $\triangle_{e''}$ and $\triangle_{e''_-}$. Suppose that e'' is an anomaly. Then there exists $A * g$ such that $A * g * \triangle_{e''} \subseteq \triangle_{e''_-}$ and no element of $\triangle_{e''}$ begins with A. So $A * g * b$ is in $\triangle_e$. But b does not begin with A, so b is not an initial segment of $A * g * b$. Hence, b is not in $\triangle_e$. Contradiction. ■

Lemma 62 (forcibility is asymptotic) *Let $A * a$ be forcible given e. Then there exists a world w in $K_e \cap A$ extending e such that for each finite initial segment e' of w, $A * a$ is forcible given e'.*

Proof. Suppose $A * b$ is forcible given e. Suppose for contradiction that the consequent of the lemma is false. Then for each w in $A \cap K_e$ there exists e' extending e and extended by w such that $A * b$ is not forcible given e'. For each w in $A \cap K_e$, let e_w be the shortest such e'. For each e_w, $A * b$ is not forcible at e_w, so since the forcing games in (K, Π) are all determined (by restriction 1), there exists a solution σ_w for (K_{e_w}, Π_{e_w}) that never produces $A * b$ after e_w. Let σ solve (K, Π) and let σ^* be just like σ except that control is shifted permanently to σ_w when e_w is encountered. So σ^* is a solution that never produces $A * b$ after seeing some e_w. Let $\sigma^\dagger$ be like σ^* except that

$\sigma^\dagger$ produces "?" along each e_w and at each e not extended by some e_w such that σ returns A at e. Then $\sigma^\dagger$ is still a solution, since σ^* converges to the truth over $K_e \cap A$ (the question marks eventually end in each w in $K_e \cap A$) and over $K_e - A$ (σ does not converge to A in any such world, so again, the question marks end eventually in each w in $K_e - A$). But $\sigma^\dagger$ doesn't produce $A * b$ after e along any e' extending e. So $A * b$ is not forcible given e. Contradiction. ■

Lemma 63 (forcible pattern existence) *Suppose that $C_e(n)$ is non-empty. Then there exists a finite pattern in $\triangle_e$ of length at least $n + 1$.*

Proof. Let w be in $C_e(0)$. In the base case, nature can force the answer A true in w from an arbitrary solution. For induction, suppose that w is in $C_e(n + 1)$. Let e' be the first anomaly along w after e. So there are n anomalies occurring in w after e'. By the induction hypothesis, there exists pattern a in $\triangle_{e'}$ of length at least $n + 1$. Since e' is an anomaly, there exists pattern $A * b$ such that $A * b * a$ is a pattern in $\triangle_{e'\_}$. Hence, $A * b * a$ has length at least $n + 2$. Since $A * b * a$ is forcible at $e'\_$, $A * b * a$ is forcible at e as well. So there exists some pattern d in $\triangle_e$ of which $A * b * a$ is a sub-pattern (by restriction 2), so d has length at least $n + 2$. ■

Lemma 64 (nature's starting point) *Let $A * a$ be in $\triangle_e$. Then there exists a world w in $C_e(0) \cap A$ such that for each finite initial segment e' of w that extends e, $A * a$ is in $\triangle_{e'}$.*

Proof. Let $A * a$ be in $\triangle_e$. So $A * a$ is forcible given e. By lemma 62, there exists w in $K_e \cap A$ such that $A * a$ is forcible along each initial segment of w extending e. Let e' properly extend e and be extended by w. Then $A * a$ is in $\triangle_{e'}$ and e' is not an anomaly (by lemma 61). Hence, w is in $C_e(0)$. ■

Lemma 65 (simplest answer forcible first) *Let answer A be the first entry in some pattern in $\triangle_e$. Then A is a simplest answer.*

Proof. Suppose that $A * b \in \triangle_e$. Then there exists w in $A \cap C_e(0)$ (by lemma 64). So $c(A, e) = 0$. ■

Lemma 66 (uniquely simplest answer and forcibility) *Let answer A be uniquely simplest at e. Then each pattern in $\triangle_e$ begins with A.*

Proof. Suppose that for some answer $B \neq A$, pattern $B * a$ is in $\triangle_e$. Then by lemma 65, B is simplest at e. So A is not uniquely simplest. ■

Lemma 67 (simple world existence) *Let K_e be non-empty. Then there exists a world w in $C_e(0)$.*

Proof. Suppose there exists w in K_e. If $c(w, e) = 0$, we are done. So suppose $c(w, e) = k > 0$. Then (by lemma 63) there exists $A * a$ in $\triangle_e$ of length $k + 1$. So there exists w' in $A \cap C_e(0)$ (by lemma 64). ■

Lemma 68 (simplest answer defeated only by anomalies) *Let K_e be non-empty, let e be non-empty, and let A be an answer in Π such that A is uniquely simplest at e_- and A is not uniquely simplest at e. Then e is an anomaly.*

Proof. Let K_e, e be non-empty. Then K_{e_-} is non-empty, so by lemma 67, $C_{e_-}(0)$, $C_e(0)$ are non-empty. So since A is uniquely simplest at e_- but not at e, we have $C_{e_-}(0) \subseteq A$ but $C_e(0) \nsubseteq A$. So there exists w in $C_e(0) - C_{e_-}(0)$. Hence, $c(w, e_-) > 0$ and $c(w, e) = 0$, so e is an anomaly. ■

Lemma 69 (forcing lemma) *Let σ be a solution and let pattern a of length at least $k + 1$ be in $\triangle_e$ and let m be a natural number. Then there exists w in $C_e(k)$ such that after e, σ produces a_0 successively for m times and then a_1 successively for m, times, ... and finally a_k successively for m times.*

Proof. Let natural number m be given. In the base case, let pattern (A) be in $\triangle_e$. Then there exists world $w \in A \cap C_e(0)$ such that (A) remains in w from e onward (by lemma 64). Since A is true in w and σ is a solution, σ converges to A in w, so σ produces A at least m times in succession after e in w.

For induction, let $A * a$, be forcible at e, where a is a finite answer pattern of length $k + 1$. There exists a world w in $A \cap K_e$ such that $A * a$ is in $\triangle_{e'}$, for each finite, initial segment e' of w extending e (by lemma 64). Since σ is a solution, σ converges to A in w. Nature can wait m steps after the onset of convergence until σ produces A at least m times after e in w. Let e' extend e such that a is in $\triangle_{e'}$ and exactly one anomaly occurs along e' after e (by restriction 3). So by the induction hypothesis, there exists w' in $C_e(k)$ such that, after e, σ produces a_0 successively for m times and then a_1 successively for m, times, ... and finally a_k successively for m times. Hence, σ produces A successively for m times followed by a_0 for m times, etc. Since exactly one anomaly occurs along e' after the end of e and k anomalies occur along w after e', w' is in $C_e(k + 1)$. ■

REFERENCES

[1] Chart, D. (2000). "Discussion: Schulte and Goodman's Riddle", *The British Journal for the Philosophy of Science* 51, 147–149.

[2] Forster, M. 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.
[3] Freivalds, R. and Smith, C. (1993). "On the Role of Procrastination in Machine Learning", *Information and Computation* 107, 237–271.
[4] Garey, M. and Johnson, D. (1979). *Computers and Intractability*, New York: Freeman.
[5] Glymour, C. (1980). *Theory and Evidence*, Princeton: Princeton University Press.
[6] Goodman, N. (1983). *Fact, Fiction, and Forecast*, 4th ed., Cambridge (Mass.): Harvard University Press.
[7] Harman, G. (1965). "The Inference to the Best Explanation", *Philosophical Review* 74, 88–95.
[8] Hitchcock, C. (ed.) (2004). *Contemporary Debates in the Philosophy of Science*, Oxford: Blackwell.
[9] Jain, S., Osherson, D., Royer, J.S. and Sharma A. (1999). *Systems That Learn: An Introduction to Learning Theory*, 2nd ed., Cambridge (Mass.): MIT Press.
[10] Kelly, K. (2002). "Efficient Convergence Implies Ockham's Razor", in *Proceedings of the 2002 International Workshop on Computational Models of Scientific Reasoning and Applications* (CMSRA 2002), Las Vegas, USA, June 24–27.
[11] Kelly, K. (2004). "Uncomputability: The Problem of Induction Internalized", *Theoretical Computer Science* 317, 227–249.
[12] Kelly, K. and Glymour, C. (2004). "Why Probability Does Not Capture the Logic of Scientific Justification", in Hitchcock, C. [8], 94–114.
[13] Kuhn, T.S. (1970). *The Structure of Scientific Revolutions*, Chicago: University of Chicago Press.
[14] Laudan, L. (1981). "A Confutation of Convergent Realism", *Philosophy of Science* 48, 19–48.
[15] Mitchell, T. (1997). *Machine Learning*, New York: McGraw-Hill.
[16] Popper, K. (1968). *The Logic of Scientific Discovery*, New York: Harper and Row.
[17] Schulte, O. (1999). "Means-Ends Epistemology", *The British Journal for the Philosophy of Science* 50, 1–31.
[18] Schulte, O. (2000a). "Inferring Conservation Principles in Particle Physics: A Case Study in the Problem of Induction", *The British Journal for the Philosophy of Science* 51, 771–806.
[19] Schulte, O. (2000b). "What to Believe and What to Take Seriously: A Reply to David Chart concerning the Riddle of Induction", *The British Journal for the Philosophy of Science* 51, 151–153.
[20] Sklar, L. (1977). *Space, Time, and Spacetime*, Berkeley: University of California Press.
[21] Van Fraassen, B. (1981). *The Scientific Image*, Oxford: Clarendon Press.