9-24-2001

Learning Theory and Epistemology

Kevin T. Kelly
Carnegie Mellon University

Learning Theory and Epistemology

Kevin T. Kelly
Department of Philosophy
Carnegie Mellon University

September 24, 2001

1 INTRODUCTION

Learning is the acquisition of new knowledge and skills. It spans a range of processes from practice and rote memorization to the invention of entirely novel abilities and scientific theories that extend past experience. Learning is not restricted to humans: machines and animals can learn, social organizations can learn, and a genetic population can learn through natural selection. In this broad sense, learning is adaptive change, whether in behavior or in belief.

Learning can occur through the receipt of unexpected information, as when a detective learns where the suspect resides from an anonymous informant. But it can also be a process whose arrival at a correct result is in some sense guaranteed before the new knowledge is acquired. Such a learning process may be said to be *reliable* at the time it is adopted. *Formal Learning Theory* is an a priori, mathematical investigation of this strategic conception of reliability. It does not examine how people learn or whether people actually know, but rather, how reliable any system, human or otherwise, could possibly be. Thus, learning theory is related to traditional psychological and epistemological issues, but retains its own, distinct emphasis and character.

Reliability is a notoriously vague concept, suggesting a disposition to acquire new knowledge or skill over a broad range of relevantly possible environments. Learning theory deals with the vagueness not by insisting on a single, sharp "explication" of reliability, but by studying a range of possible explications, no one of which is insisted upon. This approach subtly shifts the focus from intractable debates about what reliability *is* to the more objective task of determining which precise sens*es* of reliability are achievable in a given, precisely specified learning *problem*.

A learning problem specifies (1) what is to learned, (2) a range of relevantly possible environments in which the learner must succeed, (3) the kinds of inputs these environments provide to the learner, (4) what it means to learn over a range of relevantly possible environments, and (5) the sorts of learning strategies that will be entertained as solutions. A learning strategy *solves* a learning problem just in case it is admitted as a potential solution by the problem and succeeds in the specified sense over the relevant possibilities. A problem is *solvable* just in case some admissible strategy solves it.

Solvability is the basic question addressed by formal learning theory. To establish a positive solvability result, one must construct an admissible learning strategy and prove that this strategy succeeds in the relevant sense. A negative result requires a general proof that every allowable

learning strategy fails. Thus, the positive results appear "methodological" whereas the negative results look "skeptical". Negative results and positive results lock together to form a whole that is more interesting than the sum of its parts. For example, a learning method may appear unimaginative and pedestrian until it is shown that no method could do better (i.e., no harder problem is solvable). And a notion of success may sound too weak until until it is discovered that some natural problem is solvable in this sense but not in the more ambitious senses we would prefer.

There are so many different parameters in a learning problem that it is common to hold some of them fixed (e.g., the notion of success) and to allow others to vary (e.g., the set of relevantly possible environments). A partial specification of the problem parameters is called a learning *paradigm* and any problem agreeing with these specifications is an *instance* of the paradigm.

The notion of a paradigm raises more general questions. After several solvability and unsolvability results have been established in a paradigm, a pattern begins to emerge and one would like to know what it is about the combinatorial structure of the solvable problems that makes them solvable. A rigorous answer to this question is called a *characterization theorem.*

Many learning theoretic results concern the relative difficulty of two paradigms. Suppose we change a parameter (e.g., success) in one paradigm to produce another paradigm. There will usually remain an obvious correspondence between problems in the two paradigms (e.g., identical sets of serious possibilities). A *reduction* of paradigm P to another paradigm P' transforms a solution to a problem in P' into a solution to the corresponding problem in P. Then we may say that P is *no harder* than P'. Inter-reducible paradigms are *equivalent.* Equivalent paradigms may employ intuitively different standards of success, but the equivalence in difficulty shows that the quality of information provided by the diverse criteria is essentially the same. Paradigm equivalence results may therefore be viewed as epistemic analogues of the conservation principles of physics, closing the door on the temptation to get something (more reliability) for nothing by fiddling with the notion of success.

2 LEARNING IN EPISTEMOLOGY

Epistemology begins with the irritating stimulus of unlearnability arguments. For example, Sextus Empiricus records the classical problem of inductive justification as follows:

> [Dogmatists] claim that the universal is established from the particulars by means of induction. If this is so, they will effect it by reviewing either all the particulars or some of them. But if they review only some, their induction will be unreliable, since it is possible that some of the particulars omitted in the induction may contradict the universal. If, on the other hand, their review is to include all the particulars, theirs will be an inmpossible task, because particulars are infinite and indefinite (Sextus 1985): 105.

This argument may be modelled in the following *data stream paradigm.* A *data stream* is just an infinite sequence e of natural numbers encoding discrete "observations". By stage n of inquiry the learner has seen observations $e(0), e(1), \ldots e(n-1)$. An *empirical proposition* is a proposition whose truth or falsity depends only on the data stream, and hence may be identified with a set

of data streams. A learning strategy *decides* a given empirical proposition *with certainty* just in case in each relevantly possible data stream, it eventually halts and returns the truth value of the proposition.

Let the hypothesis to be assessed be "zeros will be observed forever", which corresponds to the empirical proposition whose only element is the everywhere zero data stream. Let every Boolean-valued data stream be a relevant alternative. To show that no possible learning strategy decides the hypothesis with certainty over these alternatives, we construct a "demonic strategy" for presenting data in response to the successive outputs of an arbitrary learning strategy in such a way that the learner fails to halt with the right answer on the data stream presented. The demon presents the learner with the everywhere zero sequence until the learner halts and returns "true". If this never happens, the learner fails on the everywhere zero data stream. If the learner halts with "true", there is another relevantly possible data stream that agrees with the everywhere zero data stream up to the present and that presents only ones thereafter. The demon then proceeds to present this alternative data stream, on which the learner has already halted with the wrong answer. So whatever the learner's strategy does, it fails on some relevantly possible data stream and hence does not decide the hypothesis with certainty. This is the simplest example of a negative learning theoretic argument.

The argument actually shows something stronger. *Verification* with certainty requires, asymmetrically, that the learner's strategy halt with the output "true" if the hypothesis under assessment is true and that the strategy always say "false" otherwise, possibly without ever halting. The preceding argument shows that the "zeros forever" hypothesis is not verifiable with certainty.

Karl Popper's falsificationist epistemology was originally based on the observation that although universal hypotheses cannot be verified with certainty, they can be *refuted* with certainty, meaning that a method exists that halts with "false" if the hypothesis is false and that always says "true" otherwise. In the "zeros forever" example, the refutation method simply returns "true" until a nonzero value is observed and then halts inquiry with "false".

When reliability demands verification with certainty, there is no tension between the static concept of conclusive justification and the dynamical concept of reliable success, since convergence to the truth occurs precisely when conclusive justification is received. Refutation with certainty severs this tie: the learner reliably stabilizes to the truth value of h but when h is true there is no time at which this guess is certainly justified. The separation of reliablity from complete justification was hailed as a major epistemological innovation by the American Pragmatists.[1] In light of it, one may either try to invent some notion of *partial* empirical justification (e.g., a theory of *confirmation*), or one may, like Popper, side entirely with reliability.[2] Learning theory has nothing to say about whether partial epistemic justification exists or what it might be. Insofar as such notions are entertained at all, they are assessed either as components of reliable learning strategies or as extraneous constraints on admissible strategies that may make

[1] "We may talk of the *empiricist* and the *absolutist* way of believing the truth. The absolutists in this matter say that we not only can attain to knowing truth, but we can know when we have attained to knowing it; while the empiricists think that although we may attain it, we cannot infallibly know when." (James 1948): 95-96.

[2] "Of course theories which we claim to be no more than conjectures or hypothees need no justification (and least of all a justification by a nonexistent 'method of induction', of which nobody has ever given a sensible description)." (Popper 1982): 79.

reliability more difficult or even impossible to achieve. Methodological principles with the latter property are said to be *restrictive*.[3]

"Hypothetico-deductivism" is sometimes viewed as a theory of partial inductive support (Glymour 1980), but it can also been understood as a strategy for *reducing* scientific discovery to hypothesis assessment (Popper 1968, Kemeny 1953, Putnam 1963). Suppose that the relevant possibilities are covered by a countable family of hypotheses, each of which is refutable with certainty and informative enough to be interesting. A *discovery* method produces empirical hypotheses in response to its successive observations. A discovery method *identifies* these hypotheses *in the limit* just in case on each relevantly possible data stream, the method eventually stabilizes to some true hypothesis in the family. Suppose that we have an assessment method that refutes each hypothesis with certainty. The corresponding hypothetico-deductive method is constructed as follows. It enumerates the hypotheses (by "boldness", "abduction", "plausibility", "simplicity", or the order by which they are produced by "creative intuition") and outputs the first hypothesis in the enumeration that is not rejected by the given refutation method. This reduction has occurred to just about everyone who has ever thought about inductive methodology. But things needn't be quite so easy. What if the hypotheses aren't even refutable with certainty? Could enumerating the right hypotheses occasion computational difficulties? These are just the sorts of questions of principle that are amenable to learning theoretic analysis, as will be seen below.

Another example of learning theoretic thinking in the philosophy of science is Hans Reichenbach's "pragmatic vindication" of the "straight rule" of induction (Reichenbach 1938). Reichenbach endorsed Richard Von Mises' frequentist interpretation of probability. The relative frequency of an outcome in a data stream at position n is the number of occurrences of the outcome up to position n divided by n. The *probability* of an outcome in a data stream is the limit of the relative frequencies as n goes to infinity. Thus, a probabilistic statement determines an empirical proposition: the set of all data streams in which the outcome in question has the specified limiting relative frequency.

To discover limiting relative frequencies, Reichenbach recommended using the *straight rule*, whose guess at the probability of an outcome is the currently observed relative frequency of that outcome. It is immediate by definition that if the relevant possibilities include only data streams in which the limiting relative frequency of an event type is defined, then following the straight rule *gradually identifies* the true probability value, in the sense that on each relevantly possible data stream, for each nonzero distance from the probability, the conjectures of the rule eventually stay within that distance.

If the straight rule is altered to output an open interval of probabilities of fixed width centered on the observed relative frequency, then the modified method evidently identifies a true interval in the limit (given that a probability exists). This is the same property that hypothetico-deductive inquiry has over countable collections of refutable hypotheses.

So are probability intervals refutable with certainty? Evidently not, for each finite data sequence is consistent with each limiting relative frequency: simply extend the finite sequence with an infinite data sequence in which the probability claim is true. Is there any interesting sense in which open probability intervals can be reliably assessed? Say that a learner *decides* a

[3]Cf. section 6 below.

hypothesis *in the limit* just in case in each relevantly possible enviroinment, the learner eventually stabilizes to "true" if the hypothesis is true and to "false" if the hypothesis is false. According to this notion of success, the learner is guaranteed to end up with the correct truth value, even though no relevantly possible environment affords certain verification or refutation. But even assuming that some limiting relative frequency exists, open probability intervals are not decidable even in this weak, limiting sense (Kelly 1996). A learner *verifies* a hypothesis *in the limit* just in case on each relevantly possible data stream, she converges to "true" if the hypothesis is true and fails to converge to "true" otherwise. This even weaker notion of success is "one sided", for when the hypothesis is true, it is only guaranteed that "false" is produced infinitely often (possibly at ever longer intervals).[4] Analogously, *refutation in the limit* requires convergence to "false" when the hypothesis is false and anything but convergence to "false" otherwise. It turns out that open probability intervals are verifiable but not decidable in the limit given that some probability (limiting relative frequency) exists.[5]

Thus, identification in the limit is possible even when the possible hypotheses are merely verifiable in the limit. Indeed, identification in the limit is in general reducible to limiting verification, but the requisite reduction is a bit more complicated than the familiar hypothetico-deductive construction. Suppose we have a countable family of hypotheses covering all the relevant possibilities and a limiting verifier for each of these hypotheses. Enumerate the hypotheses so that each hypothesis occurs infinitely often in the enumeration. At a given stage of inquiry, find the first remaining hypothesis whose limiting verifier currently returns "true". If there is no such, output the first hypothesis and go to the next stage of inquiry. If there is one, output it and delete all hypotheses occurring prior to it from the hypothesis enumeration. It is an exercise to check that this method identifies a true hypothesis in the limit. So although limiting verification is an unsatisfying sense of reliable assessment, it sufficees for limiting identification. If the hypotheses form a partition, the limiting verifiability of each cell is also necessary for limiting identification (Kelly 1996). So limiting verification is perhaps more important than it might first have appeared.

Neyman and Pearson justified their theory of statistical testing in terms of the frequentist interpretation of probability:

> It may often be proved that if we behave according to such a rule, then in the long run we shall reject h when it is true not more, say, than once in a hundred times, and in addition we may have evidence that we shall reject h sufficiently often when it is false (Neyman and Pearson 1933): 142.

The significance level of a test is a fixed upper bound on the limiting relative frequency of false rejection of the hypothesis under test over all possible data streams. A test is "useless" if the limiting frequency of mistaken acceptances exceeds one minus the significance, for then we could

[4]If there were any schedule governing the rate at which the the outputs "false" spread apart through time, this schedule could be used to produce a method that decides the hypothesis in the limit: the new rule outputs "false" until the simulated rule produces more "true"s than the schedule allows for. Thus the potential for ever rarer "false" outputs when the hypothesis is false is crucial to the extra lenience of this criterion.

[5]Conjecturing "true" while the observed frequency is in the interval and "false" otherwise does suffice unless we exclude possible data streams in which the limiting relative frequency approaches its limit from one side, for all but finitely many satages along the data stream. A reliable method is presented in (Kelly 1996).

have done better at reducing the limiting relative frequency of error by ignoring the data and flipping a coin biased according to the significance level. "Useful" testability can be viewed as a learning paradigm over data streams. How does it relate to the "qualitative" paradigms just discussed? It turns out that the existence of a useful test for a hypothesis is equivalent to the hypothesis being either verifiable or refutable in the limit (Kelly 1996). This is an example of a paradigm equivalence theorem, showing that useful statistical tests provide essentially no more "information" than limiting verification or refutation procedures, assuming the frequentist interpretation of probability.

It is standard to assume in statistical studies that the relevant probabilities exist, but is there a sense in which this claim could be reliably assessed? Demonic arguments reveal the existence of a limiting relative frequency to be neither verifiable in the limit nor refutable in the limit over arbitrary data streams. But this hypothesis is *gradually verifiable* in the sense that there is a method that outputs numbers in the unit interval such that these numbers approach one just if the hypothesis is true (Kelly 1996). A demonic argument shows that the existence of a limiting relative frequency is not *gradually refutable*, in the sense of producing a sequence of numbers approaching zero just in case the hypothesis is false.

Gradual decidability requires that the learner's outputs gradually converge to the truth value of the hypothesis whatever this truth value happens to be. Unlike gradual verification and refutation, which we have just seen to be weaker than their limiting analogues, gradual decision is inter-reducible with limiting decision: simply choose a cutoff value (e.g. 0.5) and output "true" if the current output is less than 0.5 and "false"otherwise. Gradual decision is familiar as the sense of success invoked in Bayesian convergence arguments. Since Bayesian updating by conditionalization can never retract a zero or a one on data of nonzero probablity, these outputs indicate certainty (inquiry may as well be halted), so limiting decision may only be accomplished gradually.

This short discussion illustrates how familiar epistemological issues as diverse as the problem of induction, Popper's falsificationism, Reichenbach's vindication of the straight rule, statistical testability, and Bayesian convergence all fit within a single, graduated system of learnability concepts.

3 COMPUTABLE LEARNING

The preceding discussion framed traditional epistemological topics in learning theoretic terms. But despite its ancient pedigree, the focus of formal learning theory on computational issues anchors it squarely in the twentieth century.

One of the earliest examples of a computationally driven unlearnability argument was presented by Hilary Putnam in 1963 in an article criticizing Rudolph Carnap's (1950) approach to inductive logic. Following suggestions by Wittgenstein, Carnap viewed inductive logic as a theory of "partial entailment", in which the conditional probability of the hypothesis given the data is interpreted as the proportion of logical possibilities satisfying the "premise" that also satisfy the intended "conclusion".

An inductive logic determines a *prediction* function: given the data encountered so far, output the most probable guess at the next datum to be seen. If there is a tie, we interpret

this as a refusal to select a prediction and view it as a failure at this round. Since the relevant probabilities are computable in Carnap's inductive logic, so is the induced prediction function.

In the *extrapolation paradigm*, the goal in each relevantly possible data stream is to eventually produce only correct predictions. Putnam showed that no computable prediction function can extrapolate the set of all total computable data streams, from which it follows that Carnap's inductive logic cannot extrapolate the computable data streams. Let an arbitrary, computable prediction strategy be given. At each stage, the demon calculates the computable prediction strategy's next prediction in light of the data already presented. If the prediction is one or greater, the demon presens a zero. If the prediction is zero, the demon presents a one. Evidently, every prediction made by the computable extrapolator along the resulting data stream is wrong. Since both the demon's strategy and the learner's strategy are computable, this data stream is computable and hence relevantly possible.[6]

On the other hand, the problem is solved by the obvious, but noncomputable, hypothetico-deductive method. Enumerate a set of computer programs computing all and only the total computable functions (i.e., no programs that go into infinite loops are included). Each such program is computably refutable with certainty by calculating its prediction for the current stage of inquiry and rejecting it if this prediction does not agree with what is observed. This method identifies a correct program in the limit. To turn it into a reliable extrapolator, just compute what the currently output hypothesis says will happen at the next stage (another example of a paradigm reduction).

The only part of this procedure that is not computable is enumerating a collection of programs covering exactly the total computable functions. Since the prediction problem is computably unsolvable, it follows immediately that no such program enumeration is computable. So computable predictors fail on this problem "because" they cannot enumerate the right collection of hypotheses.[7]

The *computable function identification* paradigm poses the closely related problem of identifying in the limit a computer program correctly predicting each position in the data stream. The preceding hypothetico-deductive method noncomputably identifies the computable data streams in this sense, but in a seminal paper, the computer scientist E. M. Gold (1965) showed that the problem is not computably solvable. The computable demonic construction employed in the proof of this result is more subtle than in the extrapolation case, because it is a nontrivial matter for a computable demon to figure out what the computable learner's current hypothesis predicts the next datum to be. For all the demon knows, the prediction may be undefined (i.e., the hypothesis may go into an infinite loop).

The demon proceeds in stages as follows:[8] At a given stage, some data points have already been presented to the learner. The demon employs a fixed, computable enumeration of all the

[6]Putnam's actual argument was more complicated.

[7]Putnam concluded that a scientific method should always be equipped with an extra input slot into which hypotheses that occur to us during the course of inquiry can be inserted. But such an "open minded" method must hope that the external hypothesis source (e.g., "creative intuition") does not suggest any programs that go into infinite loops, since the inability to distinguish such programs from "good" ones is what restricted the reliability of computable predictors to begin with!

[8]This construction (Case and Smith 1983) is a bit stronger than Gold's. It produces a data stream on which infinitely many outputs of the learner are wrong. Gold's construction merely forces the learner to vacillate forever (possibly among correct conjectures).

ordered pairs of natural numbers. He then seeks the first pair (i, j) in the enumeration such that after reading the current data followed by i zeros, the learner outputs a program that halts in j steps of computation with a prediction of zero for the next datum. If the search terminates with some such pair (i, j), then the demon adds i zeros to the data presented so far, and then presents a one (falsifying the hypothesis output by the learner after seeing the last zero). Otherwise, the demon continues searching forever and never proceeds to the next stage.

Suppose the demon's construction runs through infinitely many stages. Then the search for a pair always terminates, so the resulting data stream falsifies the learner's conjecture infinitely often. The data stream is computable because it is produced by the interaction of two computable strategies. Suppose, then, that the demon's construction eventually gets stuck at a given stage. Then the demon's search for a pair fails. So on the data stream consisting of the data presented so far followed by all zeros, the learner never produces a hypothesis that correctly predicts the next zero. This data stream is also computable: use a finite lookup table to handle the data presented so far and output zero thereafter. So in either case, the demon never identifies a correct program along some relevantly possible data stream.

Since the demon makes the learner's conjecture false infinitely often, his strategy wins even if we weaken the criterion of success to *unstable* identification in the limit, according to which the learner must eventually output only true hypotheses but need not stabilize to a particular hypothesis.[9]

Each total computer program is computably refutable with certainty (compute its successive predictions and compare them to the data), so we now know that computable refutability with certainty reduces neither computable extrapolation nor computable limiting identification. Does computable identification in the limit reduce computable extrapolation? One might suppose so: just compute the prediction of the limiting identifier's current conjecture, which must eventually be right since the identifier's conjectures are eventually correct. But although the limiting identifier eventually produces programs without infinite loops, nothing prevents it from producing defective programs in the short run. If a computer attempts to derive predictions from these conjectures in the manner just described, it may get caught in an infinite loop and hang for eternity.

Blum and Blum (1975) constructed a learning problem that is computably identifiable in the limit but not computably extrapolable for just this reason. Consider a problem in which an unknown Turing machine without infinite loops is hidden in a box and the successive data are the (finite) runtimes of this program on successive inputs. The learner's job is to guess some computer program whose runtimes match the observed runtimes for each input (a task suggestive of fitting a computational model to psychological reaction time data). In this problem, every program is computably refutable with certainty: simulate it and see if it halts precisely when the data say it should. Infinite loops are no problem, for one will observe in finite time that the program doesn't halt when it should have. Since the set of all programs is computably enumerable (we needn't restrict the enumeration to *total* programs this time), a computable implementation of the hypothetico-deductive strategy identifies a correct hypothesis in the limit.

[9]Cf. the preceding footnote. In the learning theoretic literature unstable identification is called BC identification for "behaviorally correct", whereas stable identification is called EX identification for "explanatory". Osherson et. al. (1986) call stable identification "intensional" and unstable identification "extensional".

Nonetheless, computable extrapolation of runtimes is not possible. Let a computable extrapolator be given. The demon is a procedure that wastes computational cycles in response to the computable predictor's last prediction. So at a given stage, the demonic program simulates the learner's program on the successive runtimes of the demonic program on earlier inputs. Whatever the prediction is, the demon goes into a wasteful subroutine that uses at least one more step of computation than the predictor expected.

Another question raised by the preceding discussion is whether stable identification is equivalent to or harder than unstable identification for computable learners in the computable function identification paradigm. This question is answered affirmatively by Case and Smith (1983). To see why the answer might be positive, consider the function identification problem in which the relevant possibilities are the "almost self-describing data streams". A unit variant of a data stream is a partial computable function that is just like the data stream except that it may disagree or be undefined in at most one position. A data stream is *almost self-describing* just in case it is a unit variant of the function computed by the the program whose index (according to a fixed, effective encoding of Turing programs into natural numbers) occurs in the data stream's first position. In other words, an "almost self-describing" data stream "gives away" a nearly correct hypothesis, but it doesn't say where the possible mismatch might be. An unstable learner can succeed by continually patching the "given away" program with ever larger lookup tables specifying what has been seen so far, since eventually the lookup table corrects the mistake in the "given away" program. But a stable learner would have to know *when* to stop patching, and this information was not given away.

In the problem just described, it is trivial to stably identify an almost correct program (just output the first datum) whereas no computable learner can stably identify an exactly correct program. Indeed, for each finite number of allowed errors there is a learning problem that is computably solvable under that error allowance but not with one fewer error (Case and Smith 83). This result, known as the *anomaly hierarchy theorem*, can be established by means of functions that are self-describing up to n possible errors.

There are many more sophisticated results of the kind just presented, all of which share the following points in common. (1) Uncomputability is taken just as seriously as the problem of induction from the very outset of the analysis. This is different from the approach of traditional epistemology, in which idealized logics of justification are proposed and passed along to experts in computation for advice on how to satisfy them (e.g., Levi 1991). (2) When computability is taken seriously, the *halting problem* (the *formal* problem of determining whether a computer program is in an infinite loop on a given input) is very similar to the classical problem of induction: for as soon as one is sure that a computation will never end, it might, for all the simulator knows *a priori*, halt at the next stage. (3) Thus, computable learners fail when ideal ones succeed because computable solvability requires the learner to solve an *internalized* problem of induction (Kelly and Schulte 1997).

4 SOME OTHER PARADIGMS

E. M. Gold's *language learnability* paradigm (1967) was intended to model child language acquition. In this setting, a *language* is just a computably enumerable set and a hypothesis is a

code number (index) of a procedure that *accepts* all and only the members of the set.[10] Different kinds of relevantly possible environments are considered. An *informant* for a language is an enumeration of all possible strings labelled as positive or negative examples of the language. A *text* for a language is an enumeration of the elements of the language, and hence provides only positive information about membership.

Gold showed a number of results that attracted wide attention from cognitive scientists. The results for informant are similar to those for computable function identification. For example, (1) the obvious hypothetico-deductive method (non-computably) identifies all languages and (2) even the set of all computably decidable languages is not computably identifiable in the limit (the proof is similar to the one showing that the total computable functions are not identifiable in the limit). But the results for text are much weaker. For example, no collection of languages containing one infinite language and all finite subsets of that language is identifiable in the limit, even by non-computable learners.[11] Since children seem to learn language with few negative examples or corrections (Brown and Hanlon 1970), there have been attempts to obtain stronger positive results. For example, Wexler and Culicover (1980) modelled the environment as a presentation of context-utterance pairs, exchanging language learning from positive examples for the easier problem of computable function identification. Many other variations of the language learnability paradigm have been examined.[12]

The special difficulty with learning from text is "over-generalization", or leaping to a language that properly extends the actual language, for then no further data will correct the error. If there is no way to avoid positioning a language prior to one of its proper subsets (e.g., an infinite language must occur prior to all but finitely many of its finite subsets), hypothetico-deductivism must fail, since it will converge to the large language when one of its subsets is true. What is required is a way to use evidence to avoid overgeneralizing. This can be accomplished if (†) each possible language has a finite, characteristic sample such that once that sample is seen, the language can be produced without risk of overgeneralization. Then one may proceed by enumerating the relevantly possible grammars and conjecturing the first in the enumeration that is consistent with the data and whose characteristic sample has been observed. If no such grammar exists, stick with the preceding conjecture. Condition (†) is both necessary and sufficient for a collection of languages to be identifiable in the limit from text (Angluin 1980, Osherson et. al. 1996), providing our first example of a learning theoretic *characterization theorem*. Computable identification from text is characterized by the existence of a procedure that enumerates the characteristic sample for a language when provided with the index of a formal verification program for that language.

The *logical paradigm* (Shapiro 1981, Osherson and Weinstein 1986, 1989, Kelly and Glymour 1989, 1990), situates learning theoretic ideas in a more traditional epsistemological setting. In this paradigm, there is a first-order language in which to frame hypotheses and the underlying world is a countable relational structure interpreting this language. An environment consists of such a structure together with a variable assignment onto the domain of the structure and an

[10] I.e., the procedure halts on members of the set (indicating acceptance) and not on any other inputs.

[11] The demon presents a text for the infinite language until the learner outputs a grammar for it, then keeps repeating the preceding datum until the learner produces a grammar for the data presented so far, then starts presenting the text from where he left off last, etc.

[12] A systematic compendium of results on language learnability is (Osherson et. al. 1986).

enumeration of the set of all quantifier-free formulas true under that assignment.[13] The relevant possibilities are all the environments presenting models of some theory representing the learner's background knowledge.

An hypothesis assessment method tries to guess the truth value of a particular sentence or theory in light of the increasing information provided by the environment, and successful assessment can be interpreted in any of the senses introduced above. So for example, the dense order postulate (each pair of points has a point between them) is refutable but not verifiable in the limit given as background the theory of total orders with endpoints (Osherson and Weinstein 1989).

The characterization theorem for this paradigm explains the grain of truth in the positivist's program of linking "cognitive significance" to logical form. An hypothesis is refutable (respectively, verifiable) with certainty given background theory K just in case the hypothesis is equivalent in K to a sentence in prenex normal form[14] with a purely universal (respectively, existential) quantifier prefix. Similarly, an hypothesis is refutable (respectively, verifiable) in the limit given K just in case it is equivalent in K to a prenex sentence with a prefix of form $\forall\exists$ (respectively, $\exists\forall$) (Osherson and Weinstein 1989, Kelly and Glymour 1990). As one might expect, decision with certainty is possible just in case the hypothesis is equivalent to a quantifier-free sentence in K and decision in the limit (and hence gradual decision) is possible just in case the hypothesis is equivalent in K to a finite Boolean combination of purely universal and existential sentences.

A discovery method outputs theories in response to the information provided. As the goal of discovery, one can require that the method converge to the complete true theory in some fragment of the language (e.g., the purely universal sentences). *Uniform* theory identification requires that after some time the outputs of the method are true and entail the complete theory of the required fragment. For example, the complete truth is uniformly identifiable in the limit in a language with only unary predicates, but if there is a binary predicate or a unary predicate and a function symbol in the language, then neither the the purely universal nor the purely existential fragment of the complete truth is identifiable in the limit (Kelly and Glymour 1989, Kelly 1996). *Nonuniform* or *pointwise* theory identification requires only that each true sentence in the specified fragment is eventually always entailed by the scientist's successive conjectures and each false sentence is eventually never entailed. The theory of all true Boolean combinations of universal and existential sentences is identifiable in the limit in this sense. Thus, nonuniform theory identification provides a logical conception of scientific progress that, unlike Popper's "deductivist" epistemology, treats verifiable and refutable hypotheses symmetrically.

Nonuniform theory identification bears on another Popperian difficulty. Popper held that hypothetico-deductivism leads us ever closer to the truth in the limit. David Miller (1974) argued that "closeness" to the truth is not a semantic notion since it is not preserved under translation. Thomas Mormann (1988) traced the difficulty to mathematics: translation is a type of topological equivalence, but topological equivalence permits "stretching" and hence does not preserve distance (e.g., verisimilitude). Nonuniform identification is a topological rather than a metrical notion, and hence is preserved under translation, thereby avoiding Miller-style

[13] The "onto" assumption can be dropped if empirical adequacy rather than truth is the goal (Lauth 1993).

[14] I.e., the sentence has the form of a quantifier-free sentence preceded by a sequence of quantifiers.

objections. Nonetheless it constitutes a nontrivial account of scientific progress toward the complete truth that does not imply that any future theory produced by science will be literally true.

5 RELIABILITY AND COMPLEXITY

Learnability is a matter of how the possible futures making different hypotheses correct branch off from one another through time. The more complex the temporal entanglement of the futures satisfying incompatible hypotheses, the more difficult learning will be. Learnability is governed by the *topological* complexity of the possible hypotheses and computable learnability depends on their *computational* complexity.[15]

Data streams can be topologized in an epistemologically relevant manner as follows. A *fan* of data streams is the set of all data streams extending some finite data sequence, which we may call the *handle* of the fan. A fan with a given handle is just the empirical proposition asserting that the handle has occurred in the data. An empirical proposition is *open* just in case it is a union of fans and is *closed* just in case its complement is open.[16] Then we have the following characterization: an empirical proposition is verifiable with certainty just in case it is open, is refutable with certainty just in case it is closed, and is decidable with certainty just in case it is both closed and open. For suppose that a hypothesis is open. To verify it with certainty, just wait until the observed data sequence is the handle of a fan contained in the hypothesis and halt inquiry with "true". Conversely, if a given method verifies a hypothesis with certainty, the hypothesis can be expressed as the union of all fans whose handles are finite data sequences on which the method halts with "true".

To characterize limiting and gradual success, topological generalizations of the open and closed propositions are required. Call the open and closed propositions the Σ_1 and Π_1 propositions, respectively. For each n, the Σ_{n+1} propositions are countable unions of Π_n propositions and the Π_{n+1} propositions are countable intersections of Σ_n propositions. At each level n, a proposition is Δ_n just in case it is both Π_n and Σ_n. These are known as the *finite Borel* complexity classes, which have been familiar in functional analysis since early in this century (Hinman 1978). Then it can be shown that limiting verifiability, refutability, and decidability are characterized by Σ_2, Π_2, and Δ_2, respectively and that gradual verifiability, refutability, and decidability are characterized by Π_3, Σ_3 and Δ_2, respectively. It can also be shown that when the hypotheses are mutually incompatible, stable identification in the limit is characterized by each hypothesis being Σ_2.[17]

In computable inquiry, attaching hypotheses to propositions is a nontrivial matter, so instead of bounding the complexity of empirical propositions, we must consider the overall *correctness*

[15]The computational versions of these ideas are in (Gold 1965, Putnam 1965, Kugel 1977). The topological space is introduced in (Osherson et. al. 1986) and the characterizations are developed in (Kelly 1992, 1996) A logical versions of the characterizations are developed in (Osherson and Weinstein 1991) and (Kelly and Glymour 1990).

[16]These are, in fact, the open sets of an extensively studied topological space known as the *Baire space* (Hinman 1978).

[17]Necessity of the condition fails if the hypotheses are mutually compatible or if we drop the stability requirement.

relation $C(e, h)$ indicating that hypothesis h is correct in environment e. In computable function identification, for example, correctness requires that h be the index of a computer program that computes e. In language learning from text, h must be the index of a positive test procedure for the range of e. By suitable coding conventions, language learning from informant and logical learning can also modelled with correctness relations in the data stream paradigm. Computational analogs of the Borel complexity classes can be defined for correctness relations, in which case analogous characterization theorems hold for computable inquiry (Kelly 1996).

The moral of this discussion is that the problem of induction, or empirical underdeterination, comes in degrees corresponding to standard topological and computational complexity classes, which determine the objective sense in which reliable inquiry is possible.

6 A FOOLISH CONSISTENCY

A *consistent* learner never produces an output that is incorrect of every relevantly possible data stream extending the current data sequence. For non-computable learners, consistency makes a great deal of sense: why should someone who aims to find the truth say what has to be wrong? On the other hand, we have seen that formal relations can pose an "internal" problem of induction for computable learners. Since we do not require omniscience on the empirical side, why should we do so on the formal side when the underlying structure of the problem of induction is the same on both sides?

This raises an interesting question. Could insistence on computationally achievable consistency *preclude* computationally achievable empirical reliability? The answer is striking. One can construct an empirical proposition with the following properties. (1) The proposition is computably refutable with certainty. (2) Some computable, consistent method exists for the proposition (the method that always says "false" suffices since the proposition is never verified). But (3) No consistent, computable method of even a highly idealized, uncomputable kind[18] can even gradually decide the hypothesis. Thus, where traditional epistemology sees consistency as a *means* for finding the truth sooner, enforcing *achievable* consistency may prevent computable learners from finding truths they could otherwise have reliably found. So if the aim of inquiry is to find the truth, inconsistency may be an epistemic *obligation* (rather than a merely forgivable lapse) for computable agents. Such results exemplify the sharp difference in emphasis between computational learning theory and traditional, justificationist epistemology.[19]

7 GAMBLING WITH SUCCESS

Suppose that each learning problem comes equipped with an assignment of probabilities to empirical propositions. More precisely, suppose that the probability assignment is defined on the set of all *Borel propositions* (i.e., the least set that contains all the open (Σ_1) propositions and that is closed under countable union and complementation). A *probability assignment* on the Borel propositions is a function taking values in the unit interval that assigns unity to the

[18] i.e., hyperarithmetically definable

[19] (Osherson et. al. 1986) contains many restrictiveness results carrying a similar moral. Also, see (Osherson and Weinstein 1988).

vacuous proposition and that is *finitely additive* in the sense that the probability of a finite union of mutually incompatible Borel propositions is the sum of the probabilities of the propositions the union is taken over. *Countable additivity* extends finite additivity to countable, disjoint unions. While Kolmogorov's familiar mathematical theory of probability assumes countable additivity as a postulate, limiting relative frequencies do not satisfy it and the usual foundations of Bayesian probability theory do not entail it (e.g., DeFinetti 1990, Savage 1972).

Say that an hypothesis is gradually decidable *with probability* r just in case there exists some empirical proposition of probability r over which the hypothesis is gradually decidable in the usual sense, and similarly for the other assessment criteria. Probabilistic success can be much easier to achieve than success in each relevant possibility. If the probability assignment is countably additive, then, remarkably, every Borel hypothesis is (1) decidable in the limit with unit probability and (2) decidable with certainty with arbitrarily high but non-unit probability. (1) can be improved to the result that the method of updating the given probability measure by conditionalization gradually decides the hypothesis with unit prior probability (e.g., Halmos 1970). This is a very general version of the familiar Bayesian claim that prior probabilities are eventually "swamped" by the data.

Compared with the purely topological analysis of section 5, these probabilistic results seem almost too good to be true, since Borel propositions can be infinitely more complex than Δ_2 propositions (Hinman 1978). What accounts for the dramatic difference? Suppose we want to decide the "zeros forever" hypothesis with a given, nonzero probability r. The negation of this hypothesis is the countable, disjoint union of the hypotheses h_i = "the first nonzero occurs at position i". So by countable additivity, the probability that the "zeros forever" hypothesis is false is the sum of the probabilities of the propositions h_i. Since the infinite sum converges to a finite value, there is some position n such that the sum of the probabilities of $h_n, h_{n+1}, \ldots$ is less than r. So our probability of failure is less than r if we halt with "true" at stage n if no nonzero datum has been seen by position n and halt with "false" as soon as a nonzero datum is seen. In other words, countable additivity asserts that when a high prior probability of successful learning suffices, only finitely many of the demon's opportunities to make the hypothesis false matter.

Without countable additivity, it is possible that the probability that the hypothesis is false exceeds the mass distributed over the h_n, say by a value of r. Since this "residual" probability mass is not distributed over the propositions h_i, the learner never "gets past" it, so whenever the learner halts inquiry with "true", the probability that this conclusion was in error remains at least as high as r. The residual probability reflects the demon's *inexhaustible* opportunities to falsify the hypothesis in the infinite future, providing a probabilistic model of Sextus' demonic argument. In fact, both (1) and (2) can fail when countable additivity is dropped (Kelly 1996), highlighting the pivotal epistemological significance of this questionable and somewhat "technical" looking assumption.

8 CONCEPT LEARNING AND THE PAC PARADIGM

In the *Meno*, Plato outlined what has come to be known as the *concept learning paradigm*, which has captured the imagination of philosophers, psychologists, and artificial intelligence researchers

ever since. A concept learning problem specifies a domain of *examples* described as vectors of *values* (e.g., blue, five kilos) of a corresponding set of *attributes* (e.g., color, weight), together with a set of possible *target concepts*, which are sets of examples. The learner is somehow presented with examples labelled either as positive or as negative examples of the concept to be learned, and the learner's task is to converge in some specified sense to a correct definition. In contemporary artificial intelligence and cognitive science, the "concepts" to be learned are defined by neural networks, logic circuits, and finite state automata, but the underlying paradigm would still be familiar to Socrates.

Socrates ridiculed students who proposed disjunctive concept definitions, which suggests that he admitted only conjunctively definable concepts as relevant possibilities. Socrates' solution to the problem was to have the environment "give away" the answer in a mystical flash of insight. But J. S. Mill's (i.e., Francis Bacon's) well-known inductive methods need no mystical help to identify conjunctive concepts with certainty: the first conjecture is the first positive example sampled. On each successive positive example in the sample, delete from the current conjecture each conjunct that disagrees with the corresponding attribute value of the example (the "method of difference"). On each successive negative example that agrees with the current conjecture everywhere except on one attribute, underline the value of that attribute in the current conjecture (the "method of similarity"). When all conjuncts in the current conjecture are underlined, halt inquiry.

Boolean concepts are also identifiable with certainty over a finite set of attribute values: wait for all possible examples to come in and then disjoin the positive ones. Bacon's methods sound plausible in the conjunctive case, but this "jerrymandering" procedure for learning Boolean concepts sounds hopeless (it is, in fact, just what Socrates ridiculed). Yet both procedures identify the truth with certainty since the set of examples is finite. The PAC (Probably Approximately Correct) paradigm distinguishes such "small" problems in terms of *tractable* rather than merely *computable* inquiry.[20]

In the PAC paradigm, examples are sampled with replacement from an urn in which the probability of selecting an example is unknown. There is a collection of relevantly possible concepts and also a collection of hypotheses specifying the possible forms in which the learner is permitted to define a relevantly possible concept. Say that a hypothesis is ϵ-accurate just in case the sampling probability that a single sampled individual is a counterexample is less than ϵ. The learner is given a *confidence* parameter δ and an *error* parameter ϵ. From these parameters, the learner specifies a sample size and upon inspecting the resulting sample, she outputs a hypothesis. A learning strategy is *probably approximately correct* (PAC) just in case for each probability distribution on the urn and for each ϵ, δ exceeding zero, the strategy has a probability of at least $1 - \epsilon$ of producing an ϵ-accurate hypothesis.

It remains to specify what it means for a PAC learning strategy to be *efficient*. Computational complexity is usually analyzed in terms of asymptotic growth *rate* over an infinite sequence of "similar" but "ever larger" examples of the problem. Tractability is understood as resource consumption bounded almost everywhere by some polynomial function of problem size. The size of a concept learning problem is determined by (1) the number of attributes (2) the size

[20] An excellent source presenting all of the results mentioned here is (Kearns and Vazirani 1994), which provides detailed descriptions and bibliographic notes for all the results mentioned below.

of the smallest definition of the target concept, (3) the reciprocal of the confidence parameter, and (4) the reciprocal of the error parameter (higher accuracy and reliability requirements make for a "bigger" inference problem). A *data efficient* PAC learner takes a sample in each problem whose size is bounded by a polynomial in these four arguments.

There is an elegant combinatorial characterization of how large the sample required for PAC learning should be. Say that a concept class *shatters* a set S of examples just in case each subset of S is the intersection of S with some concept in the class. The *Vapnik-Chervonenkis* (VC) dimension of the concept class is the cardinality of the largest set of instances shattered by the class. There exists a fixed constant c such that if the VC dimension of the concept class is d, it suffices for PAC learnability that a sample of size s be taken, where

$$s \geq c\left(\frac{1}{\epsilon}\log\frac{1}{\delta} + \frac{d}{\epsilon}\log\frac{1}{\epsilon}\right).$$

For example, the VC dimension of the conjunctive concepts over n Boolean attributes is $2n$ (and in fact is just n if $n > 1$) so the problem is data-efficiently solvable by setting the sample size according to the above formula and then using any method producing conjectures consistent with the data (e.g., Bacon's method of similarity). Calculating the VC dimension of the concepts decidable by neural networks reveals that they are also data-efficiently learnable.

On the negative side, it can be shown that if the VC dimension of a concept class is d, then on some concept and in some sampling distribution, a sample size of at least d/ϵ is required. Since the VC dimension of the Boolean concepts over n Boolean attributes is 2^n, exponentially large samples will sometimes be required. Thus, any algorithm that takes a sample whose size depends only on the problem and not the size of the (unknown) target concept itself will be data-inefficient (since the sample size grows non-polynomially when concept size is held fixed at the minimum value).

A *computationally efficient* PAC learner is a PAC learner whose runtime is bounded by a polynomial of the sort described in the definition of data efficiency. Since scanning a sampled instance takes time, computational efficiency implies data efficiency. Since Bacon's method is computationally trivial and requires small samples, it is a computationally efficient PAC learner. This method can be generalized to efficiently PAC learn k-CNF concepts (i.e., conjunctions of k-ary disjunctions of atomic or negated atomic sentences), for fixed k.

Sometimes computational difficulties arise entirely because it is hard for the learner to frame her conjecture in the required hypothesis language. It is known, for example, that the k-term DNF concepts (i.e., disjunctions of k purely conjunctive concepts) are not efficiently PAC learnable using k-term DNF hypotheses (when $k \geq 2$),[21] whereas they are efficiently PAC learnable using k-CNF hypotheses

For some time it was not known whether there exist efficiently solvable PAC problems that are unsolvable neither due to sample-size compexity nor due to output representation. It turns out (Kearns and Valiant 1994) that under a standard cryptographic hypothesis,[22] the Boolean concepts of length polynomial in the number of attributes have this property, as does the neural network training problem.

[21] This negative result holds only under the familiar complexity-theoretic hypothesis that $P \neq NP$.

[22] I.e, that computing discrete cube roots is intractable even for random algorithms.

An alternative way to obtain more refined results in a non-probabilistic context is to permit the learner to ask questions. A *membership oracle* accepts an example from the learner and returns "in" or "out" to indicate whether it is a positive or a negative example. A *Socratic oracle* responds to an input conjecture with a counterexample, if there is one.[23] One such result is that Socratic and membership queries suffice for identification of finite state automata with certainty in polynomial time (Angluin 1987).

9 LEARNING THEORY AND EPISTEMOLOGY

To coherentists, learning theory looks like a naive form of foundationalism, in which incorrigible beliefs are the fulcrum driving inquiry to the truth. But foundationalists are also disappointed because positive learning theoretic results depend on substantial, contingent assumptions such as the nature of the signals from the environment, the structure of time, and the range of relevant possibilities. Externalists would prefer to investigate *our* reliability directly, instead of taking a mathematical detour into possible methods and problems. And contextualists will object to the fixity of truth through time, ignoring the possibility of meaning shifts due to conceptual change.

But on a more careful examination, learning theory reinforces recent epistemological trends. The search for incorrigible foundations for knowledge is no longer considered a serious option, so the fact that reliability depends on contingent assumptions is hardly a penetrating objection. Indeed, it can be shown by learning theoretic means that if some background knowledge is necessary for reliability, this knowledge cannot be reliably assessed according to the same standard, blocking any attempt at an entirely reliability-based foundationalism.

Externalist epistemologies sidestep the foundational demand that the conditions for reliability be known by requiring only that we be reliable, without necessarily being aware of this fact. Knowledge attributions are then empirical hypotheses that can be studied by ordinary empirical means. But empirical science is not the same as behavioristic science. Mature empirical investigations are always focused by general mathematical constraints on what is possible. Accordingly, learning theoretic results constrain naturalistic epistemology by specifying how reliable an arbitrary system, whether computable or otherwise, could possibly be in various learning situations.

Externalism has encountered the objection (Lehrer 1990) that reliability is insufficient for knowledge if one is not justified in believing that one is reliable (e.g., someone has a thermometer implanted in her brain that suddenly begins to produce true beliefs about the local temperature). The intended point of such objections is that reliable belief-forming processes should be embedded in a coherent belief system incorporating beliefs about the agent's own situation and reliability therein. Learning theory may then be viewed as defining the crucial relation of *methodological coherence* between epistemic situations, ambitions, and means. Unlearnability arguments isolate methodological incoherence and positive arguments suggest methods, background assumptions, or compromised ambitions which, if adopted, could bring a system of beliefs into methodological coherence.

Incorporating learning theoretic structure into the concept of coherence addresses what some coherentists take to be the chief objection to their position.

[23] In the learning theoretic literature, Socratic queries are referred to as "equivalence" queries.

> . . . [A]lthough any adequate epistemological theory must confront the task of bridging the gap between justification and truth, the adoption of a nonstandard conception of truth, such as a coherence theory of truth, will do no good unless that conception is independently motivated. Therefore, it seems that a coherence theory of justification has no acceptable way of establishing the essential connection with truth (Bonjour 1985): 110.

Whether a methodological principle guarantees or prevents reliable convergence to the truth is, of course, the unshakable focus of learning theoretic analysis. Where coherence is at issue, one must consider a multitude of possible interpretations of reliability and of one's epistemic situation, backing and filling until the analysis seems apt and fits with the rest of one's beliefs. This pluralistic attitude is reflected in the wide variety of success criteria, paradigms and problems considered in the learning theoretic literature.

Contextualists may also find some value in learning theoretic results. The first moral of the subject is that reliability is highly sensitive to the finest details of the data presentation, the range of possible alternatives, the kinds of hypotheses or skills at issue, the learner's cognitive powers and resources, and the methodological principles to which she is committed. Reliable methodology is unavoidably piece-meal, contextual methodology, optimized to the special features of the problem at hand.

A remaining contextualist objection is that learning theory presupposes a fixed "conceptual scheme" in which truth is a fixed target, whereas in light of conceptual revolutions, meaning and hence truth changes as the beliefs of the learner change through time. This objection does apply to the usual learning theoretic paradigms, but the concept of reliability is flexible enough to accommodate it. If truth feints as inquiry lunges, then success can be defined as a methodological *fixed point* in which the beliefs of the learner are eventually true *with respect to themselves* (Kelly 1996, Kelly and Glymour 1992). Unlike norms of justification, which may change through time, convergence to the *relative* truth provides a strategic aim that plausibly survives successive changes in the underlying scientific tradition.

10 Bibliography

Angluin, D. (1987) "Learning Regular Sets from Queries and Counterexamples" *Information and Computation* 75: 87-106.

Angluin, D. (1989) "Inductvive Inference of Formal Languages from Positive Data," *Information and Control* 49: 117-135.

Blum, M. and L. Blum (1975) "Toward a Mathematical Theory of Inductive Inference", *Information and Control* 28: 125-155.

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

Brown, R. and C. Hanlon (1970) "Derivational Complexity and the Order of Acquisition of Child Speech", in *Cognition and the Development of Language* ed. J. Hayes. New York: Wiley.

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

Case, J. and C. Smith (1983) "Comparison of Identification Criteria for Machine Inductive Inference," *Theoretical Computer Science* 24: 193-220.

DeFinetti (1990) *The Theory of Probability*. New York: Wiley.

Glymour, C. (1980) *Theory and Evidence*. Cambridge: M.I.T. Press.

Gold, E. M. (1965) "Limiting Recursion", *Journal of Symbolic Logic* 30: 27-48.

Gold, E. M. (1967) "Language Identification in the Limit", *Information and Control* 10: 447-474.

Halmos, P. (1974) *Measure Theory*. New York: Springer.

Hinman, P. (1978) *Recursion Theoretic Hierarchies*. New York: Springer.

James, W. (1948) "The Will to Believe", in *Essays in Pragmatism*, ed. A. Castell. New York: Collier Macmillan.

Kearns, M. and L. Valiant (1994) "Cryptographic limitations on learning boolean formulae and finite automata". *Journal of the ACM*, 41: 57-95.

Kearns, M. and Vazirani, U. (1994) *An Introduction to Computational Learning Theory*. Cambridge: M.I.T. Press.

Kelly, K. (1992) *Learning Theory and Descriptive Set Theory*, *Logic and Computation* 3: 27-45.

Kelly, K. (1996) *The Logic of Reliable Inquiry*. New York: Oxford University Press.

Kelly, K. and C. Glymour (1989) "Convergence to the Truth and Nothing But the Truth", *Philosophy of Science* 56: 185-220.

Kelly, K. and C. Glymour (1990) "Theory Discovery from Data with Mixed Quantifiers", *Journal of Philosophical Logic* 19: 1-33.

Kelly, K. and C. Glymour (1992) "Inductive Inference from Theory-Laden Data", *Journal of Philosophical Logic* 21: 391-444.

Kelly, K. and O. Schulte (1995) "The Computable Testability of Theories Making Uncomputable Predictions", *Erkenntnis* 43: 29-66.

Kelly K. and O. Schulte (1997) "Church's Thesis and Hume's Problem". *Logic and Scientific Methods*, M. L. Dalla Chiara et. al. eds, Dordrecht: Kluwer.

Kemeny, J. (1953) "The Use of Simplicity in Induction", *Philosophical Review* 62: 391-408.

Kugel, P. (1977) "Induction Pure and Simple", *Information and Control*, 33: 236-336.

Lehrer, K. (1990) *Theory of Knowledge*. San Francisco: Westview.

Levi, I. (1991) *The Fixation of Belief and its Undoing*. Cambridge: Cambridge University Press.

Neyman, J. and E. Pearson (1933) "On the Problem of the Most Efficient Tests of Statistical Hypotheses", *Philosohical Transactions of the Royal Society* 231 A: 289-337.

Lauth, B. (1993) "Inductive Inference in the Limit for First-Order Sentences," *Studia Logica* 52: 491-517.

Miller, D. (1974) "On Popper's Definitions of Verisimilitude", *British Journal of the Philosophy of Science* 25: 155-188.

Mormann, T. (1988) "Are All False Theories Equally False?", *British Journal for the Philosophy of Science* 39: 505-519.

Osherson, D., S. Weinstein (1986) *Systems that Learn.* Cambridge: M.I.T. Press.

Osherson, D. and S. Weinstein (1989) "Paradigms of Truth Detection", *Journal of Philosophical Logic* 18: 1-41.

Osherson, D. and S. Weinstein (1988) "Mechanical Learners Pay a Price for Bayesianism", *Journal of Symbolic Logic* 56: 661-672.

Osherson, D. and S. Weinstein (1989) "Identification in the Limit of First Order Structures", *Journal of Philosophical Logic* 15: 55-81.

Osherson, D. and S., Weinstein (1991) "A Universal Inductive Inference Machine", *Journal of Symbolic Logic* 56: 661-672.

Popper, K. (1982) *Unended Quest: an Intellectual Autobiography*, LaSalle: Open Court.

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

Putnam, H. (1963) "'Degree of confirmation' and inductive logic", in *The Philosophy of Rudolph Carnap*, ed. A. Schilpp. LaSalle: Open Court.

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

Reichenbach, H. (1938) *Experience and Prediction.* Chicago: University of Chicago Press.

Savage, L. (1972) *The Foundations of Statistics.* New York: Dover.

Sextus Empiricus (1985) *Selections from the MajorWritings on Scepticism, Man and God*, ed. P. Hallie, trans. S. Etheridge. Indianapolis: Hackett.

Shapiro, E. (1981) "Inductive Inference of Theories from Facts", Report YLU 192. New Haven: Department of Computer Science, Yale University.

Wexler, K. and P. Culicover (1980) *Formal Principles of Language Acquisition.* Cambridge: M.I.T. Press.