

Logically Reliable Inductive Inference

Oliver Schulte

Department of Philosophy and School of Computing Science
Simon Fraser University
Vancouver, Canada

July 20, 2005

Abstract

This paper aims to be a friendly introduction to formal learning theory. I introduce key concepts at a slow pace, comparing and contrasting with other approaches to inductive inference such as confirmation theory. A number of examples are discussed, some in detail, such as Goodman's Riddle of Induction. I outline some important results of formal learning theory that are of philosophical interest. Finally, I discuss recent developments in this approach to inductive inference.

1 Introduction: Convergence to the Truth

The purpose of this article is to provide a brief and friendly introduction to some of the key mathematical concepts of formal learning theory. Understanding these concepts is essential for following the philosophical and mathematical development of the theory. The reader may find further discussion and defence of the basic philosophical ideas in this volume, as well as in sources such as ([20], [30], [9], [11], [36], [38]).

Learning theory addresses the question of how we should draw conclusions based on evidence. Philosophers have noted since antiquity that if we are interested in questions of a general nature, the evidence typically does not logically entail the answer. To start with a well-worn example, any finite number of black ravens is logically consistent with some future raven not being black. In such cases, logical deduction based on the evidence alone does not tell us what general conclusions to draw. The question is what else should govern our inferences. One prominent idea is that we should continue to seek something like a logical argument from evidence as premises to theory as conclusion. Such an argument is not guaranteed to deliver a true conclusion, but something other than truth. For example, we may seek a type of argument to the effect that, given the evidence, the conclusion is probable, confirmed, justified, warranted, rationally acceptable etc.¹

¹For a fairly detailed but brief comparison of formal learning theory with various ways of

Formal learning theory begins with an alternative response to the underdetermination of general conclusions by evidence. Empirical methods should reliably deliver the truth just as logical methods do. But unlike deduction, *inductive inquiry need not terminate with certainty*. In the learning-theoretic conception of inductive success, a method is guaranteed to eventually arrive at the truth, but this does not mean that after some finite time, the method yields certainty about what the right generalization is: An inquirer can be in the possession of the truth without being certain that she is. A philosophical forerunner of this idea is Peirce's notion that science would find the truth "in the limit of inquiry", but need never yield certainty [28]. As his fellow pragmatist William James put it, "no bell tolls" when science has found the right answer [16]. Reichenbach's pragmatic vindication of induction applied this conception of empirical success to the problem of estimating probabilities (interpreted as limits of relative frequencies) [31]. Reichenbach's student Hilary Putnam showed how the idea could be developed into a general framework for inductive inference [29, 30].² The notion of success in the limit of inquiry is subtle and requires some getting used to. I will illustrate it by working through two simple examples.

2 First Example: Black Ravens

Consider the problem of investigating whether all ravens are black. Imagine an ornithologist who tackles this problem by examining one raven after another. There is exactly one observation sequence in which only black ravens are found; all others feature at least one nonblack raven. Figure 1 illustrates the possible observation sequences.

If the world is such that only black ravens are found, we would like the ornithologist to settle on this generalization. (It may be possible that some nonblack ravens remain forever hidden from sight, but even then the generalization "all ravens are black" at least gets the observations right.) If the world is such that eventually a nonblack raven is found, then we would like the ornithologist to arrive at the conclusion that not all ravens are black. This specifies a set of goals of inquiry. For any given inductive method that might represent the ornithologist's disposition to adopt conjectures in the light of the evidence, we can ask whether that method measures up to these goals or not. There are infinitely many possible methods to consider; we will look at just two, a sceptical one and one that boldly generalizes. The bold method conjectures that all ravens are black after seeing that the first raven is black. It hangs on to this conjecture unless some nonblack raven appears. The skeptical method does not go beyond what is entailed by the evidence. So if a nonblack raven is found, the skeptical method concludes that not all ravens are black, but otherwise the method does not make a conjecture one way or another. Figure 2 illustrates both the generalizing and the skeptical method.

cashing out this idea, see [22].

²The cognitive scientist Mark Gold independently developed the same conception of inductive inference as Putnam to analyze language acquisition [12].

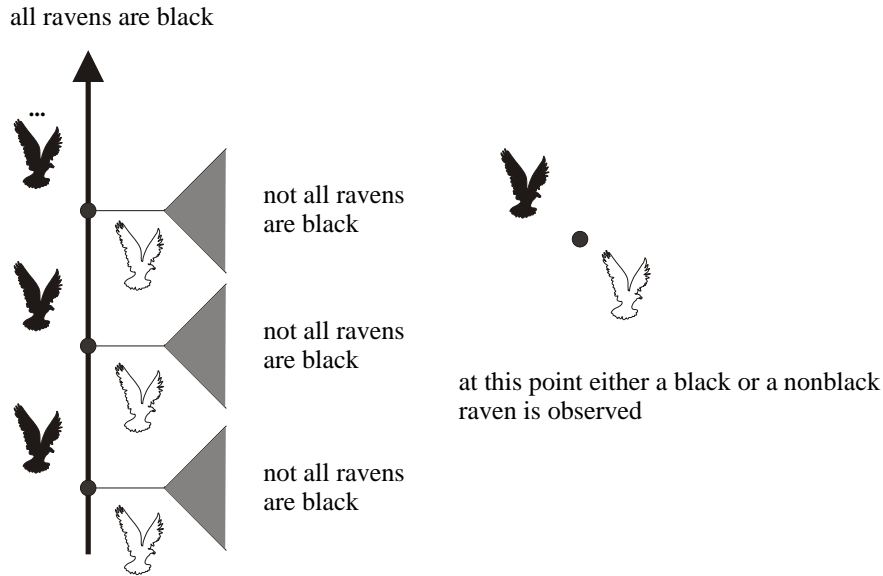


Figure 1: Data Sequences and Alternative Hypotheses for the Raven Color Problem

Do these methods attain the goals we set out? Consider the bold method. There are two possibilities: either all observed ravens are black, or some non-black raven is found. In the first case, the method conjectures that all ravens are black and never abandons this conjecture. In the second case, the method concludes that not all ravens are black as soon as the first nonblack raven is found. Hence no matter how the evidence comes in, eventually the method gives the right answer as to whether all ravens are black and sticks with this answer.

The skeptical method does not measure up so well. If a nonblack raven appears, then the method does arrive at the correct conclusion that not all ravens are black. But if all ravens are black, the skeptic never takes an "inductive leap" to adopt this generalization. So in that case, the skeptic fails to provide the right answer to the question of whether all ravens are black.

This illustrates how means-ends analysis can evaluate methods: the bold method meets the goal of reliably arriving at the right answer, whereas the skeptical method does not. Note the character of this argument against the skeptic: The problem, in this view, is not that the skeptic violates some canon of rationality, or fails to appreciate the "uniformity of nature". The learning-theoretic analysis concedes to the skeptic that no matter how many black ravens have been observed in the past, the next one could be white. The issue is that if all observed ravens are indeed black, then the skeptic never answers the question "are all ravens black?". Getting the right answer to that question requires

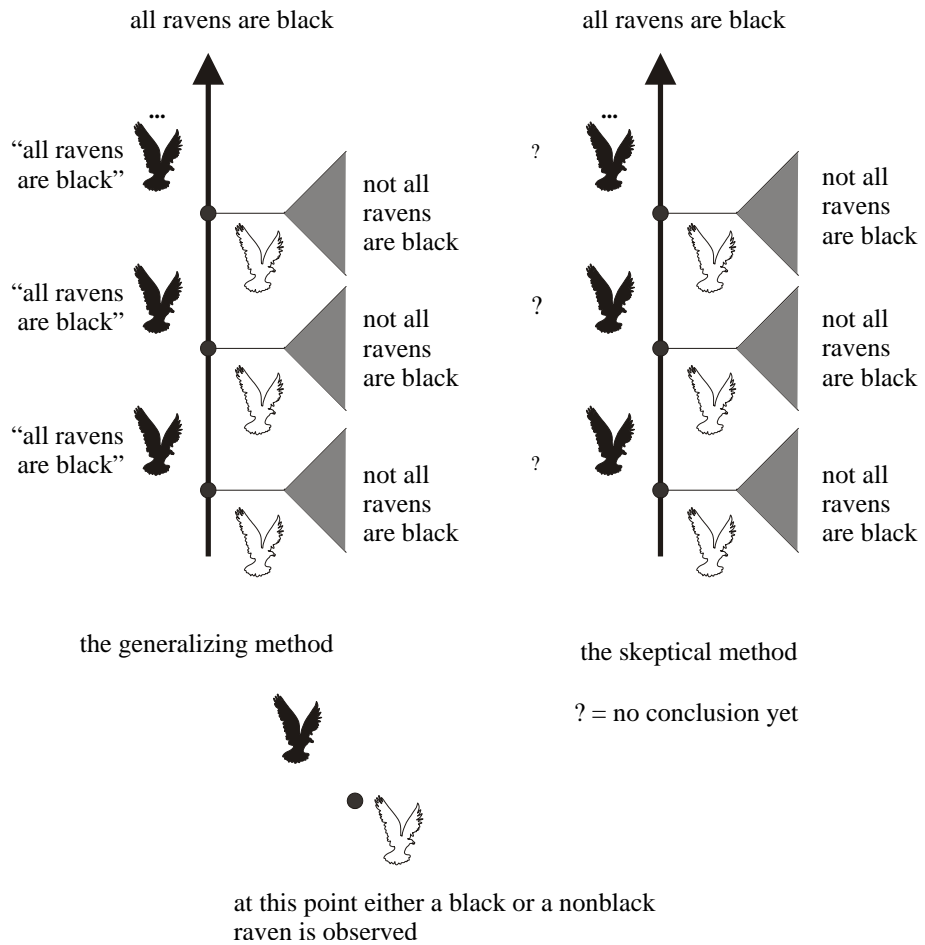


Figure 2: The Generalizer and the Skeptic in the Raven Color Problem

generalizing from the evidence even though the generalization could be wrong.

3 Second Example: The New Riddle of Induction

Let us go through a second example to reinforce the notion of reliable convergence to the right answer.

Nelson Goodman posed a famous puzzle about inductive inference known as the (New) Riddle of Induction [13]. Our next example is inspired by his puzzle. Goodman considered generalizations about emeralds, involving the familiar colors of green and blue, as well as certain unusual ones:

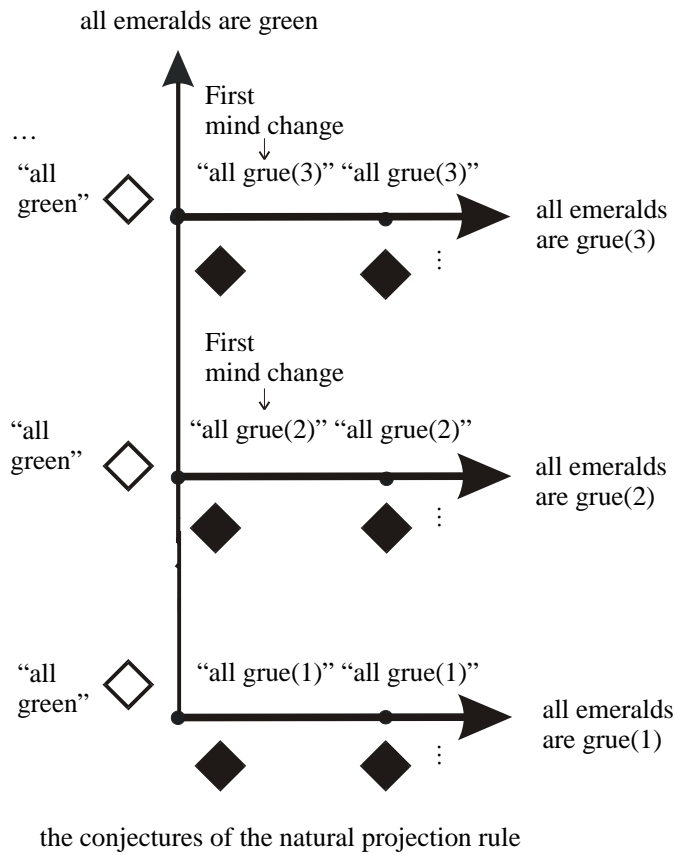
Suppose that all emeralds examined before a certain time t are green... Our evidence statements assert that emerald a is green, that emerald b is green, and so on...

Now let me introduce another predicate less familiar than "green". It is the predicate "grue" and it applies to all things examined before t just in case they are green but to other things just in case they are blue. Then at time t we have, for each evidence statement asserting that a given emerald is green, a parallel evidence statement asserting that emerald is grue.

The question is whether we should conjecture that all emeralds are green rather than that all emeralds are grue when we obtain a sample of green emeralds examined before time t , and if so, why.

Clearly we have a family of grue predicates in this problem, corresponding to different "critical times" t ; let's write $grue(t)$ to denote these. Following Goodman, I refer to "projection rules" in discussing this example. A projection rule succeeds in a world just in case it settles on a generalization that is correct in that world. Thus in a world in which all examined emeralds are found to be green, we want our projection rule to converge to the proposition that all emeralds are green. If all examined emeralds are $grue(t)$, we want our projection rule to converge to the proposition that all emeralds are $grue(t)$. Note that this stipulation treats green and grue predicates completely on a par, with no bias towards either. As before, let us consider two rules: the "natural" and the "gruesome" projection rules. The natural projection rule conjectures that all emeralds are green as long as only green emeralds are found; if a blue emerald is found, say at stage n for the first time, the rule conjectures that all emeralds are $grue(n)$. The "gruesome" rule keeps projecting the next grue predicate consistent with the available evidence. Expressed in the green-blue vocabulary, the gruesome projection rule conjectures that after observing some number of n green emeralds, all future ones will be blue. Figures 3 and 4 below illustrate the possible observation sequences and the two methods mentioned in this model of the New Riddle of Induction.

How do these rules measure up to the goal of arriving at a true generalization? Suppose for the sake of the example that the only serious possibilities

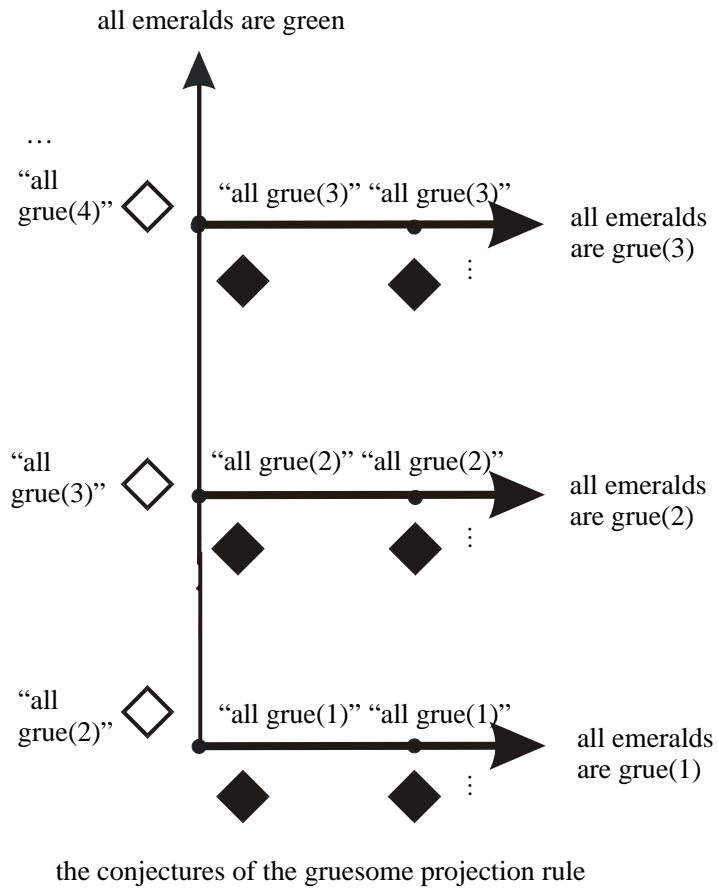


"all grue(t)" = "all emeralds are grue(t)"

"all green" = "all emeralds are green"

\diamond • \blacklozenge At this stage, either a green or a blue emerald may be observed

Figure 3: The Natural Projection Rule in the New Riddle of Induction



“all grue(t)” = “all emeralds are grue(t)”



 At this stage, either a green or a blue emerald may be observed

Figure 4: The Gruesome Projection Rule in the New Riddle of Induction

under consideration are that either all emeralds are green or that all emeralds are *grue*(t) for some critical time t . Then the natural projection rule settles on the correct generalization no matter what the correct generalization is. For if all emeralds are green, the natural projection rule asserts this fact from the beginning. And suppose that all emeralds are *grue*(t) for some critical time t . Then at time t , a blue emerald will be observed. At this point the natural projection rule settles on the conjecture that all emeralds are *grue*(t), which must be correct given our assumption about the possible observation sequences. Thus no matter what evidence is obtained in the course of inquiry—consistent with our background assumptions—the natural projection rule eventually settles on a correct generalization about the color of emeralds.

The gruesome rule does not do as well. For if all emeralds are green, the rule will never conjecture this fact because it keeps projecting *grue* predicates. Hence there is a possible observation sequence—namely those on which all emeralds are green—on which the gruesome rule fails to converge to the right generalization. So means-ends analysis would recommend the natural projection rule over the gruesome rule. Some comments are in order.

(1) As in the previous example, nothing in this argument hinges on arguments to the effect that certain possibilities are not to be taken seriously a priori. In particular, nothing in the argument says that generalizations with *grue* predicates are ill-formed, unlikelike, or in some other way a priori inferior to "all emeralds are green".

(2) The analysis does not depend on the vocabulary in which the evidence and generalizations are framed. For ease of exposition, I have mostly used the green-blue reference frame. However, *grue*-*bleen* speakers would agree that the aim of reliably settling on a correct generalization requires the natural projection rule rather than the gruesome one, even if they would want to express the conjectures of the natural rule in their *grue*-*bleen* language rather than the blue-green language that I have used. (For more on the language-invariance of means-ends analysis see [33, 34].)

(3) Though the analysis does not depend on language, it does depend on assumptions about what the possible observation sequences are. The example as I have described it seems to comprise the possibilities that correspond to the color predicates Goodman himself discussed. But means-ends analysis applies just as much to other sets of possible predicates. Schulte [34] and Chart [6] discuss a number of other versions of the Riddle of Induction, in some of which means-ends analysis favors projecting that all emeralds are *grue* on a sample of all green emeralds.

4 Reliable Convergence to the Truth: General Concepts and Definitions

Now that we have seen two examples of the basic idea, let us encapsulate it more generally in a mathematical definition. I begin with the description of an

inductive or learning problem, which involves a specification of *possible observations*, *alternative hypotheses* and which hypotheses count as *correct* given a total body of evidence. Then I define the concept of an inductive method, and finally specify Putnam’s and Gold’s notion of empirical success for inductive methods.

4.1 Inductive Problems: Observations, Data Streams, Hypotheses, Background Knowledge and Correctness

In both examples, we have a *set of possible hypotheses* that an inquirer could adopt in the course of inquiry. In the ravens example, the set is {“all ravens are black”, “not all ravens are black”}. In the Riddle of Induction, the (infinite) set of hypotheses is {“all emeralds are green”, “all emeralds are *grue*(1)”, “all emeralds are *grue*(2)”, ..., “all emeralds are *grue*(*n*)”, ...}. In realistic examples, the hypotheses may be considerably more complex. For instance, in language learning models the set of alternatives is the set of all grammars that may govern the language spoken in the learner’s (child’s) native environment [27]. In models of scientific inquiry, the alternative theories could be sets of conservation principles [35], or models of cognitive functioning [10], [3].

Another part of the specification of a learning problem is a set of *evidence items*. In the raven example, there are two kinds of evidence items “a black raven is observed”, or “a nonblack raven is observed”. In the Riddle of Induction, the set of evidence items is {green emerald, blue emerald}. In more realistic applications, we have many more, even infinitely many, evidence items. For example, an evidence item may be a measurement of a quantity, or set of quantities, in a physical experiment. In studying particle dynamics, the set of evidence items comprises all interactions among elementary particles that we may observe in particle accelerators [35]. In cognitive psychology, an evidence item could be the behavior profile of a subject in an experiment [10].

A *data stream* is an infinite sequence of evidence items. We write ε for a typical data stream, ε_i for the *i*-th datum observed in the data stream ε , and $\varepsilon|n$ for the first *n* data observed along ε . For example, if ε is the data stream along which only green emeralds are observed, then $\varepsilon_i = \text{"green"}$ for all *i*, and $\varepsilon|n$ is $\langle \text{"green"}, \text{"green"}, \dots, \text{"green"} \rangle$ for *n* repetitions of "green". If ε is the data stream on which all emeralds are *grue*(1), then $\varepsilon_1 = \text{"green"}$, $\varepsilon_i = \text{"blue"}$ for all *i* > 1, and $\varepsilon|n = \langle \text{"green"}, \text{"blue"}, \dots, \text{"blue"} \rangle$ with *n* – 1 repetitions of "blue".

An inquirer may have background knowledge relevant to the question under investigation. For example, a particle physicist may assume that all particle reactions satisfy relativity theory. In a language learning problem, we may restrict attention only to languages with computable (total recursive) grammars. In such cases, the inquirer may be willing to rule out certain observations a priori. We can model the inquirer’s background assumptions as a set *K* of data streams that represents the set of all infinite observation sequences that may arise for all the inquirer knows.

Definition 1 (Evidence Items and Empirical Background Knowledge)

Let E be a set of *evidence items*.

1. A data stream ε is an infinite sequence of evidence items. That is, ε_n is a member of E for each n .
2. The initial sequence comprising the first n observed data along ε is denoted by $\varepsilon|n$.
3. The inquirer's background knowledge is represented by a set of data streams K that may occur for all the inquirer knows.

In applications of learning theory, we assume that for every data stream there is a hypothesis that is *correct* for the data stream. For example, the hypothesis “all emeralds are green” is correct for the data stream on which only green emeralds are observed. The hypothesis “not all ravens are black” is correct on any data stream on which some nonwhite raven is observed. The correctness relation between data streams and hypotheses is part of the specification of the inductive problem. Learning theory is agnostic about what correctness is. In the examples we have considered, correctness amounts to empirical adequacy: the goal is to find a generalization that makes the right predictions about what will be observed when. Correct hypotheses may be the true ones, or the simplest true ones, or simply the empirically adequate hypotheses. Another way to put it is that the correctness relation C expresses the inquirer's goals: if the total (infinite) observational data were such and such, as found in a data stream ε , then the inquirer wants to adopt a hypothesis H such that $C(H, \varepsilon)$ holds. Thus learning theory per se does not recommend to an inquirer what hypotheses she should view as correct for a total body of evidence. Rather, the theory helps the inquirer find a correct hypothesis from the partial body of evidence actually available at a given stage of inquiry.

Without going into details, it may be useful to indicate how the model of inquiry I have outlined so far corresponds to the language learning models much studied in formal learning theory. In language learning models [15], the evidence items are called "strings" and the counterpart of a data stream is a "text". The alternative hypotheses are (indices for) "languages"; a language is a set of evidence items, which models the view of a language as a set of strings.

4.2 Inductive Methods and Inductive Success

After observing a finite sequence of evidence items, an inquirer produces a hypothesis—her guess as to the right answer. Mathematically, this corresponds to a function that assigns a hypothesis to a finite data sequence. We also allow an inquirer to refrain from adopting an answer, which is indicated by a ? for “no guess yet”. Such a function is a mathematical representation of an inquirer's disposition to output guesses in response to evidence. Following some philosophical tradition, we refer to such a function as an inductive method, or method for short. Figure 5 illustrates the notion of a method.

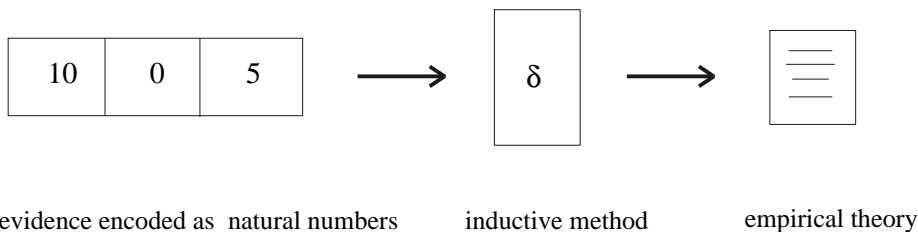


Figure 5: An inductive method takes as input an evidence sequence and outputs a hypothesis. Discrete evidence items can be generically represented by natural numbers (e.g., 0 for "black raven", 1 for "white raven").

Definition 2 (Inductive Methods) *Let E^* denote the set of finite evidence sequences, and let \mathcal{H} be a collection of alternative hypotheses. An inductive method is a function $\delta : E^* \rightarrow \mathcal{H} \cup \{?\}$ such that for each finite data sequence E , the output $\delta(E)$ is either a hypothesis H or the vacuous output $?$.*

Some comments will clarify the concept of a method in relation to other concepts and terminology.

(1) Philosophers often discuss functions from evidence to belief without calling them methods. A fairly common alternative term is “rule”. For example, Goodman discusses “projection rules” for generalizing from observed emeralds. In his analysis of knowledge, Nozick does use the term “method” for a doxastic disposition [26]. Learning-theoretic analysis applies to any disposition that gives rise to belief given evidence, whether such a disposition is called “method” or not. An alternative term for method in learning theory is simply “learner”, and recently the term “scientist” has come into use [15], [25].

(2) The notion of method as given in Definition 2 is neutral about the interpretation of adopting a hypothesis: “outputting” a hypothesis can model various epistemic attitudes that an inquirer may take towards her theory, such as belief, full belief, posit, acceptance, entertaining, etc. In fact, learning theory is even more agnostic about the concept of belief than Definition 2 suggests because the framework can accommodate just about any concept of belief, including degrees of belief as in a probabilistic theory, or degrees of confirmation as in confirmation theory. For example, Putnam investigated whether Carnap’s inductive methods (his “c-functions”) arrive at the right answer, in the sense that whatever the true generalization is, eventually the true generalization always receives degree of confirmation greater than 1/2 [29]. Or we can ask whether the degree of belief of a Bayesian agent in the true generalization will come arbitrarily close to 1 [17], [7, Ch.9.6], [25, Sec.3.6.9].

In general, to apply learning theory it suffices to have a notion of an (epistemic) state s and a correctness relation $Correct(\varepsilon, s)$ that specifies the correct states for the agent to be in, given that the total observational facts are as described by the data stream ε . The point is that learning theory does not

presuppose, and hence does not depend on, a particular analysis of belief or epistemic attitudes. Rather, the theory addresses the question of how best to change one’s belief, however understood.

(3) The notion of method as given in Definition 2 is agnostic about internal facts concerning how the agent arrives at her hypothesis. In effect, the definition views a method as a black box, as suggested in Figure 5. Learning theory focuses on the behavior and performance of epistemic dispositions, not on their internal structure. As a consequence, learning theoretic analysis applies to any recommendation for how we should reason from evidence to theory: whether the proposal is to follow a certain style of argument (e.g., probabilistic), seek a certain kind of conformation (e.g., Carnap’s c-functions [4] or Glymour’s bootstrap confirmation [13]), or adopt some set of normative criteria for rational belief formation: we can always ask whether those ways of producing belief would lead an inquirer to the correct hypothesis (cf. [36]).

With Definitions 1 and 2 in hand, we are ready to define Putnam’s and Gold’s conception of empirical success.

Definition 3 (Reliable Convergence to the Correct Hypothesis) *Let E be a set of evidence items, \mathcal{H} a set of alternative hypotheses, C a correctness relation that specifies which hypotheses are correct for each data stream ε comprising observations drawn from E .*

1. *A method δ converges to, or identifies, a correct hypothesis H on a data stream $\varepsilon \iff H$ is correct for ε and there is a stage n such that $\delta(\varepsilon|n') = H$ for all stages $n' \geq n$.*
2. *A method δ is reliable for, or identifies, \mathcal{H} given background knowledge $K \iff$ for all data streams ε consistent with K (i.e., ε in K), the method δ converges to a correct hypothesis on ε .*

To illustrate this definition, we verified in Section 3 that the natural projection rule reliably identifies a true generalization about emerald colors given the set of alternatives {"all green", "all *grue*(1)", ...}. The gruesome method that keeps predicting that the next emerald is blue fails to converge to “all emeralds are green” on the data stream featuring only green emeralds. Definition 3 envisions a method converging to a single hypothesis; in algorithmic learning theory, this corresponds to "EX-learning"—see the introductory chapter in this volume.

Part of the traditional concept of a method, for example in Mill and arguably in Aristotle, is that a method should be a step-by-step reasoning procedure. The definition above does not require that a method should be easy to follow. In modern terms, a step-by-step procedure of the sort sought by traditional philosophers corresponds to an *algorithm* which by Church’s thesis can be implemented on a Turing machine. It is therefore natural to require that methods should be algorithmic or computable. Such an algorithm provides a step-by-step procedure for following the method. Much of formal learning theory studies algorithmic methods, so much so that the subject is often referred to

as algorithmic learning theory (as in the title of this volume) or computational learning theory.

Some striking results in algorithmic learning theory examine what norms of inductive reasoning help agents with bounded cognitive powers and which hinder them in attaining the aims of inquiry. The point is not the trivial one that the deductive abilities of agents limited to the reasoning powers of a Turing Machine fall short of ideal logical omniscience. Rather, it turns out that computable inquiry sometimes requires fundamentally different strategies than inquiry by idealized agents. In such cases, trying to approximate or "get as close as possible" to the ideal norm can be a bad strategy for computable agents seeking to identify a correct hypothesis.

For example, consider the seemingly banal *consistency principle*: do not accept a hypothesis that is inconsistent with the data (see for example Hempel's "conditions of adequacy" for a definition of scientific confirmation [14, Ch.I.1.8]). Kelly and Schulte describe an inductive problem with an empirical hypothesis H such that a step-by-step reasoning procedure can reliably identify in the limit whether or not H is correct, but inductive methods even with infinitely uncomputable reasoning powers cannot do so—if they are required to satisfy the consistency principle. (For another restrictiveness result along these lines, see [25, Prop. 60].) Intuitively, the main reason why the consistency principle restricts the potential of computable inquiry is that an agent with bounded logical powers cannot immediately recognize when a hypothesis is inconsistent with the data, but must first gather *more data*. The consistency principle rules out this inductive strategy because it mandates that an agent should reject a hypothesis as soon as it is refuted. For further discussion of the differences between methodology for logically omniscient agents and those with bounded deductive abilities, see [19], [20, Ch.6,7,10], [25].

5 Additional Epistemic Goals: Fast and Stable Convergence to the Truth

The seminal work of Putnam and Gold focused on reliable convergence to a correct hypothesis. A major extension of their approach is to consider cognitive desiderata *in addition to finding a correct hypothesis* (such desiderata are called "identification criteria" in the computer science literature [5]). In this section, I consider two epistemic aims that have received considerable attention from learning theorists: stable and fast convergence to a correct theory.

The motivation for examining convergence speed is that other things being equal, we would like our methods to arrive at a correct theory sooner rather than later. A venerable philosophical tradition supports the idea that stable belief is a significant epistemic good. Since Plato's *Meno*, philosophers are familiar with the idea that stable true belief is better than unstable true belief, and epistemologists such as Sklar [39] have advocated similar principles of "epistemic conservatism". Kuhn tells us that a major reason for conservatism in paradigm

debates is the cost of changing scientific beliefs [23]. In this spirit, learning theorists have examined methods that minimize the number of times that they change their theories before settling on their final conjecture.

As it turns out, the idea of adding cognitive goals in addition to finding a correct hypothesis addresses a long-standing objection to identification in the limit. Reichenbach's student Salmon criticized his teacher's pragmatic vindication of induction on the grounds that the vindication, even if successful, leaves belief underdetermined in the short run [32]. The reason is that while Reichenbach's straight rule is guaranteed to approach the true probability of an event, so are infinitely many other rules. For example, consider a rule δ that estimates the probability of a coin coming up heads to be 1 for 1,000 tosses no matter what the outcome of the tosses is. After 1,000 tosses, δ switches to following the straight rule. Thus in the limit of inquiry, the rule δ converges to the same answer as the straight rule does.

From this example it is easy to see the general pattern: Suppose that δ is a reliable method; let e be any evidence sequence, and H be any hypothesis. Then there is a method δ' that outputs H on e and follows the reliable method δ on any other evidence. So δ' converges to the same hypothesis as δ and thus δ' is reliable. This shows that any conjecture H on any evidence e is consistent with long-run reliability.

The situation changes drastically if we take into account other aspects of empirical success. Several general recent results show that maximizing stable belief, or minimizing mind changes, strongly constrains the conjectures of optimal inductive methods in the short run. I will illustrate the power of additional epistemic goals in the two simple traditional examples already considered.

First, we need to define what it is for an inductive method to succeed with respect to an epistemic goal. For a given epistemic desideratum, a method may perform well in some circumstances but not in others. To compare the performance of methods with regard to a range of possible ways the world might be—more precisely, with regard to all the data streams consistent with background knowledge—we may apply two familiar principles from decision theory: **admissibility** and **minimax**. A method is *admissible* iff it is not *dominated*. In general, an act A dominates another act A' if A necessarily yields results at least as good as those of A' , and possibly better ones, where a given collection of “possible states of the world” determines the relevant sense of necessity and possibility. An act A *minimizes* if the worst possible outcome from A is as good as the worst possible outcome from any other act.

For the two epistemic desiderata of minimizing time-to-truth and reversals of opinion, applying the two decision-theoretic criteria of admissibility and minimax yields $2 \times 2 = 4$ identification criteria. It turns out that two of these, admissibility for mind changes and minimaxing convergence time, are feasible only for empirical questions that pose no genuine problem of induction; more precisely, they are feasible only if the data are eventually guaranteed to entail which hypothesis is correct. (For the details see [33].) Thus learning theorists have focused on minimaxing theory changes and admissibility with respect to convergence time. I will discuss minimizing reversals of opinion in the remain-

der of this section and the next and then return to admissibility with respect to time-to-truth.

5.1 Stable Convergence to a Correct Hypothesis

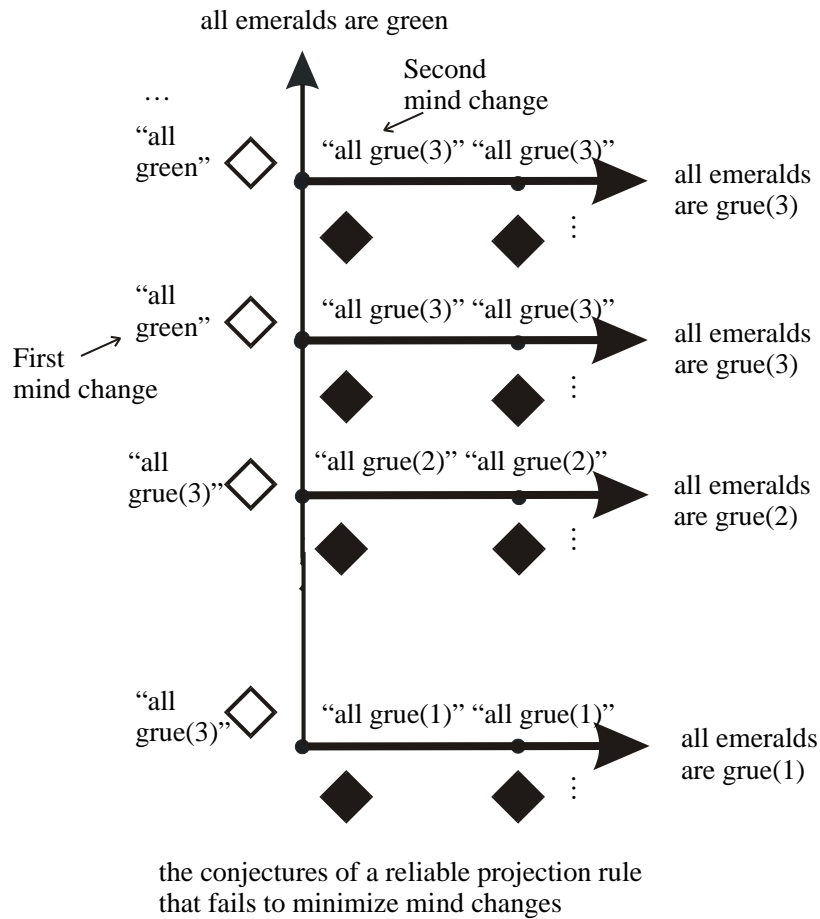
We say that a method δ **changes its mind** on a data sequence e_1, \dots, e_n, e_{n+1} if the method's output on the previous data e_1, \dots, e_n is not? (i.e., $\delta(e_1, \dots, e_n) \neq ?$) and differs from its output at stage $n + 1$ (i.e., $\delta(e_1, \dots, e_n) \neq \delta(e_1, \dots, e_n, e_{n+1})$). No mind changes occur on the empty data sequence.

Definition 4 (Stable Belief: Minimizing Mind Changes) *Suppose that δ is a reliable discovery method for alternative hypotheses \mathcal{H} given background knowledge K .*

1. *The number of mind changes of δ on data stream ε is given by $MC(\delta, \varepsilon) \equiv |\{n : \delta \text{ changes its mind on } \varepsilon|n\}|$.*
2. *The method δ succeeds with at most n mind changes given $K \iff MC(\delta, \varepsilon) \leq n$ for all data streams ε consistent with K .*
3. *The method δ **minimizes mind changes** given hypotheses \mathcal{H} , background knowledge $K \iff$ there is no other reliable method δ' for \mathcal{H} such that the maximum number of times that δ might change its mind, given background knowledge K , is greater than the same maximum for δ' .*

The New Riddle of Induction turns out to be a nice illustration of minimizing mind changes. Consider the natural projection rule (conjecture that all emeralds are green on a sample of green emeralds). If all emeralds are green, this rule never changes its conjecture. And if all emeralds are *grue*(t) for some critical time t , then the natural projection rule abandons its conjecture "all emeralds are green" at time t —one mind change—and thereafter correctly projects "all emeralds are *grue*(t)". Hence the natural projection rule changes its mind at most once in the New Riddle of Induction (see Figure 3). Remarkably, rules that project *grue* rather than green do not do as well. For example, consider a rule that conjectures that all emeralds are *grue*(3) after observing one green emerald. If two more green emeralds are observed, the rule's conjecture is falsified and it must eventually change its mind, say to conjecture that all emeralds are green (suppose that green emeralds continue to be found). But then at that point, a blue emerald may appear, forcing a second mind change. This argument can be generalized to show that the aim of minimizing mind changes allows only the green predicate to be projected on a sample of all green emeralds [33, Prop. 11]. Figure 6 illustrates in a typical case how an unnatural projection rule may have to change its mind twice or more. From the insight illustrated in Figure 6, we can establish the optimality of the natural projection rule.

Proposition 5 *Let δ be any projection rule (inductive method) that reliably identifies a true generalization about emerald colors in the Riddle of Induction*



“all grue(t)” = “all emeralds are grue(t)”

“all green” = “all emeralds are green”

◊ ● At this stage, either a green or a blue emerald may be observed

Figure 6: A reliable projection rule that projects a grue predicate on an all green sample of emeralds can be forced to change its mind twice.

and changes its conjecture at most once. Let e be any finite sequence featuring only green emeralds (i.e., e is of the form $\langle \text{green emerald, green emerald, } \dots \rangle$). Then either $\delta(e) = ?$ – the method makes no guess – or $\delta(e) = \text{“all emeralds are green”}$.

Less formally, the proposition says that after observing a sequence of emeralds consistent with “all emeralds are green”, an optimal method must conjecture “all emeralds are green” or else withhold opinion. The criteria of reliable convergence to the truth and stable belief do not determine how many instances exactly are required for inference to “build up enough confidence” and “take an inductive leap”. These goals do determine that (1) a reliable method must eventually take an inductive leap, and (2) when the method does adopt a universal generalization in the Riddle of Induction, on a sample of all green emeralds that generalization must be “all emeralds are green”.

In the ravens example, the results of the analysis are similar. A reliable method that minimizes retractions may withhold opinion on a sample of all black ravens, but if it does generalize beyond the data, it must conjecture that all ravens are black rather than that some nonblack raven will appear in the future. Our two examples illustrate the typical pattern for methods that achieve as much stable belief as possible: minimizing mind changes determines the what of inductive generalizations, but not the when. (For more precise statements and proofs of this principle, see [24], [21]).

5.2 Fast Convergence to a Correct Hypothesis

Let us return to the idea of minimizing time-to-truth. Formally, we may develop this success criterion as follows. Define the **convergence point** of a method δ on a data stream ε to be the time at which the method starts to converge to an answer. That is, $CP(\delta, \varepsilon) \equiv$ the least n such that $\delta(\varepsilon|n) = \delta(\varepsilon|n')$ for all $n' \geq n$. For a set of alternative hypotheses \mathcal{H} and given background knowledge K , an inductive method δ **dominates** another inductive method δ' with respect to convergence time \iff

1. background knowledge K entails that δ converges no later than δ' does (i.e., $CP(\delta, \varepsilon) \leq CP(\delta', \varepsilon)$ for all $\varepsilon \in K$), and
2. there is some data stream, consistent with background knowledge K , on which δ converges before δ' does (i.e., there is $\varepsilon \in K$ such that $CP(\delta, \varepsilon) < CP(\delta', \varepsilon)$).

A method δ is data-minimal given K if no other reliable method for \mathcal{H} dominates δ with respect to convergence time ([20, Ch.4.8]; see also [25, Def.28]).

There is a theorem that characterizes the properties of data-minimal methods [34, Th.8], [25, Ex.39]. A consequence of the theorem is that data-minimal methods always adopt a definite belief—that is, they never output “?”. Intuitively, suspending belief loses time, because the method could have begun converging to a true belief instead. For our examples, it follows that the natural

projection rule in the Riddle of Induction and the bold generalizer in the ravens problem are the only reliable data-minimal methods that minimax retractions.

6 Further Extensions and Applications

This section indicates some further extensions and developments of the theory of reliable inquiry with additional epistemic values.

(1) Many problems do not allow a finite bound on mind changes, although there is still an intuitive sense that some methods achieve stable belief more than others. Freivalds showed how the notion of a finite mind change bound can be extended to an ordinal or transfinite bound [8]. This well-studied criterion considerably enhances the range of inductive problem in which the goal of minimizing mind changes is feasible [15]. An even more general formulation of the idea has been very recently developed by Kelly [21].

(2) Although problems such as the Riddle of Induction and generalizing about black ravens may appear very different on the surface, there is a common structure to problems that can be solved with at most 1 mind change, as Figures 1 and 3 suggest. This holds true for any finite and even transfinite mind change bounds. The common deep structure of problems solvable with a given mind change bound can be explicated in terms of point-set topology (cf. [34], [20, Ch.4], [24, Sec.3]). For language and function learning problems, which are commonly studied in Computational Learning Theory, the mind change complexity of an inductive problem is characterized by Cantor’s classic concept of accumulation order ([2], [24, Th.1]).

The fact that the goals of true and stable belief place such strong constraints on inductive inference allows us to evaluate specific inference methods with respect to how well they serve these goals. Pursuing this question almost always leads to insights into the inductive problem under investigation, increases our understanding of known learning methods, and can lead to the development of new methods. I conclude this introduction with some brief illustrations of applying this kind of learning-theoretic analysis in some fairly realistic inference problems.

(1) An inductive problem that arises in particle physics is to find a set of conservation laws that correctly predict which reactions among elementary particles are possible [35]. A prominent type of conservation law consists of additive conservation laws, also known as selection rules. It can be shown that there is a unique optimal method for inferring selection rules [35]. It turns out that the standard set of laws that particle physicists have actually adopted makes exactly the same predictions as the output of the learning-theoretically optimal method [37].

(2) Angluin introduced the well-known concept of a “pattern” for describing a set of strings [1]. For example, the pattern $0xx1$ describes such strings as 0001, 0111, 000001, 011111. A one-variable pattern is a pattern that contains at most one distinct variable, such as $0xx1$. Angluin provided an inference algorithm for identifying a one-variable pattern in the limit that does not, however, minimize

mind changes [24, Sec.5]. Luo and Schulte describe a different algorithm that is mind change optimal (moreover, their algorithm requires time only linear in the length of a data sequence e to produce a conjecture for e).

(3) Kelly has generalized the idea of reliable inference with bounded mind changes to settings of statistical inference concerned with statistical theories that determine the distribution of observed variables [18, Sec.11]. In that setting Kelly shows that the standard practice of statistical hypothesis testing is mind change optimal: take as the null hypothesis a point estimate (e.g., the mean of the distribution is 0) and neither accept nor reject (corresponding to "?") unless and until the null hypothesis is rejected. Another application of Kelly's analysis are problems of causal inference. In causal inference, a basic problem is to find which variables are directly causally linked to each other (e.g., there is a direct connection between "tar content in lung" and "lung cancer" which mediates the indirect connection between "smoking" and "lung cancer"). Standard methods for causal inference conjecture that there is no direct link between two variables unless and until a direct connection is conclusively verified (by statistical tests). Kelly shows that this inference method is mind change optimal [18, Sec.11].

In conclusion, formal learning theory provides a rich set of concepts for analyzing the complexity of inductive problems and the performance of inductive methods. In applications, these analytical tools have yielded insights into the learning problem, validated existing learning methods and led to the development of new ones. One goal of this article was to lay out some of the basic concepts and techniques that underlie learning-theoretic analysis to invite the development of further applications.

References

- [1] D Angluin. Finding patterns common to a set of strings. *J. Comput. Syst. Sci.*, 21(1):46–62, 1980.
- [2] K. Apsitis. Derived sets and inductive inference. In S. Arikawa and K. P. Jantke, editors, *Proceedings of ALT 1994*, pages 26–39. Springer, Berlin, Heidelberg, 1994.
- [3] J. Bub. Testing models of cognition through the analysis of brain-damaged performance. *British Journal for the Philosophy of Science*, 45:837–855, 1994.
- [4] Rudolf Carnap. *The Continuum of Inductive Methods*. University of Chicago Press, Chicago, 1952.
- [5] J. Case and C. Smith. Comparison of identification criteria for machine inductive inference. *Theoretical Computer Science*, 25:193–220, 1983.
- [6] David Chart. Discussion: Schulte and goodman's riddle. *British Journal for the Philosophy of Science*, 51:147–149, 2000.

- [7] J. Earman. *Bayes or Bust?* MIT Press., Cambridge, Mass., 1992.
- [8] R. Freivalds and C. H. Smith. On the role of procrastination in machine learning. *Inf. Comput.*, 107(2):237–271, 1993.
- [9] C. Glymour. The hierarchies of knowledge and the mathematics of discovery. *Minds and Machines*, 1:75–95, 1991.
- [10] C. Glymour. On the methods of cognitive neuropsychology. *British Journal for the Philosophy of Science*, 45:815–835, 1994.
- [11] C. Glymour and K. Kelly. Thoroughly modern meno. In John Earman, editor, *Inference, Explanation and Other Frustrations*. University of California Press, 1992.
- [12] E. Mark Gold. Language identification in the limit. *Information and Control*, 10(5):447–474, 1967.
- [13] N. Goodman. *Fact, Fiction and Forecast*. Harvard University Press, Cambridge, MA, 1983.
- [14] Carl Gustav Hempel. *Aspects of Scientific Explanation*. Free Press, New York, 1965.
- [15] S. Jain, D. Osherson, J. S. Royer, and A. Sharma. *Systems That Learn*. M.I.T. Press, 2 edition, 1999.
- [16] W. James. The will to believe. In H.S. Thayer, editor, *Pragmatism*. Hackett, Indianapolis, 1982.
- [17] C. Juhl. Objectively reliable subjective probabilities. *Synthese*, 109:293–309, 1997.
- [18] K. Kelly. Justification as truth-finding efficiency: How ockham’s razor works. *Minds and Machines*, 14(4):485–505, 2004.
- [19] K. Kelly and O. Schulte. Church’s thesis and hume’s problem. In *Proceedings of the IX International Joint Congress for Logic, Methodology and the Philosophy of Science*, Dordrecht, 1995. Kluwer.
- [20] Kevin T. Kelly. *The Logic of Reliable Inquiry*. Oxford University Press, Oxford, 1996.
- [21] Kevin T. Kelly. Simplicity, truth and the unending game of science. In *Foundations of the Formal Sciences V*, 2005. In Press.
- [22] Kevin T. Kelly, Oliver Schulte, and Cory Juhl. Learning theory and the philosophy of science. *Philosophy of Science*, 64:245–267, 1997.
- [23] Thomas Kuhn. *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago, 1970.

- [24] Wei Luo and Oliver Schulte. Mind change efficient learning. In *18th Annual Conference On Learning Theory (COLT 2005)*, number 3559 in LNAI, pages 398–412, Bertinoro, Italy, June 27-30 2005. Springer.
- [25] Eric Martin and Daniel Osherson. *Elements of Scientific Discovery*. MIT Press, Cambridge, Mass., 1998.
- [26] R. Nozick. *Philosophical Explanations*. Harvard University Press, Cambridge, 1981.
- [27] D. N. Osherson, M. Stob, and S. Weinstein. *Systems that learn: an introduction to learning theory for cognitive and computer scientists*. MIT Press, 1986.
- [28] C.S. Peirce. How to make our ideas clear. In N. Houser and C. Kloesel, editors, *The Essential Peirce*, volume 1, pages 124–141. Indiana University Press, Bloomington, 1878/1992.
- [29] Hilary Putnam. 'degree of confirmation' and inductive logic. In A. Schilpp, editor, *The Philosophy of Rudolf Carnap*. Open Court, La Salle, Ill., 1963.
- [30] Hilary Putnam. Probability and confirmation. In *Mathematics, Matter and Method, Philosophical Papers*, volume 1. Cambridge University Press, London, 1975.
- [31] H. Reichenbach. *The Theory of Probability*. Cambridge University Press, London, 1949.
- [32] Wesley Salmon. Hans reichenbach's vindication of induction. *Erkenntnis*, 35:99–122, 1991.
- [33] Oliver Schulte. Means-ends epistemology. *The British Journal for the Philosophy of Science*, 79(1):141–147, 1996.
- [34] Oliver Schulte. The logic of reliable and efficient inquiry. *The Journal of Philosophical Logic*, 28:399–438, 1999.
- [35] Oliver Schulte. Inferring conservation laws in particle physics: A case study in the problem of induction. *The British Journal for the Philosophy of Science*, 51:771–806, 2000.
- [36] Oliver Schulte. Review of martin and osherson's 'elements of scientific inquiry'. *The British Journal for the Philosophy of Science*, 51:347–352, 2000.
- [37] Oliver Schulte. Automated discovery of conservation principles and new particles in particle physics. Manuscript Under Review, 2005.
- [38] Oliver Schulte. Formal learning theory. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. The Metaphysics Research Lab, Stanford University, summer 2005 edition edition, 2005.

- [39] Lawrence Sklar. Methodological conservatism. *Philosophical Review*, LXXXIV:374–400, 1975.