Special Issue III
2009
ISSN 1807-9792

abstracta

Linguagem, Mente & Ação

ABSTRACTA

Linguagem, Mente e Ação

ISSN 1807-9792

Special Issue III
2009

Editors

André Abath

Leonardo de Mello Ribeiro

abstracta
2004-2009

TABLE OF CONTENTS

abstracta
Linguagem, Mente & Ação

Editorial

The third Special Issue of ABSTRACTA is dedicated to *Reliable Reasoning: Induction and Statistical Learning Theory* (MIT Press, 2007) by Gilbert Harman and Sanjeev Kulkarni, both from Princeton University. We are proud to publish this critical discussion of such an important topic for philosophy of science in general (and of psychology, in particular) as well as contemporary epistemology.

In *Reliable Reasoning* Harman and Kulkarni take seriously the vexed philosophical problem of induction. They shed new light on it by proposing that philosophers can benefit from the application of statistical learning theory (and its mathematical framework) to approach that problem. As Harman and Kulkarni tell us in this symposium, their "intention in writing *Reliable Reasoning* was to suggest that basic statistical learning theory provides one sort of response to the traditional philosophical problem of induction, which asks what can be shown *a priori* about induction."

The result could not have been more stimulating. As Michael Strevens points out in his paper, "*Reliable Reasoning* is a simple, accessible, beautifully explained introduction to (…) statistical learning theory." Similarly, Glenn Shafer highlights that the authors "have written an enjoyable and informative book that makes [statistical learning theory] accessible to a wide audience" and whose "undertaking is important, and the execution is laudable." Paul Thagard, in his turn, contends that Harman & Kulkarni "have presented a strong case that statistical learning theory is highly relevant to issues in philosophy and psychology concerning inductive inferences." And Stephen José Hanson reminds us that the book "is a wonderful redux to a time in the early half of the 20th century when statistical learning theory was just developing, and when new methods and concepts were being discovered."

We are thankful to all those who have taken part in this symposium for giving us the opportunity of publishing such a high-level discussion. We would like to thank, first of

all, Gilbert Harman and Sanjeev Kulkarni for their attention, time, and generous support throughout the process of editing the symposium as well as for making it possible in the first place. We also thank the discussants, Stephen José Hanson, Glenn Shafer, Michael Strevens, and Paul Thagard, who dedicated their time and efforts to write their contributions, and who have made possible such an open and qualified academic debate. Last but not least, we are grateful to Ann Twiselton and to the MIT Press for their kind support.

We are confident that our readers will enjoy this intellectually challenging issue of ABSTRACTA.

André Abath &
Leonardo de Mello Ribeiro,
EDITORS.

April, 2009.

Abstracta SPECIAL ISSUE III, pp. 5 – 9, 2009

PRECIS OF "RELIABLE REASONING: INDUCTION AND STATISTICAL LEARNING THEORY" (MIT Press, 2007)

Gilbert Harman & Sanjeev Kulkarni

Reliable Reasoning seeks to show how results in basic statistical learning theory bear on issues in philosophy and psychology.

One is the ancient philosophical problem of induction. The problem is sometimes (misleadingly) motivated through a comparison of induction with deduction. Deduction is said to be perfectly reliable in the sense that the truth of the premises in a deduction guarantees the truth of the conclusion. This is said typically not to be the case for induction. Induction can lead from truths to falsehoods.

The comparison with deduction is of course misleading because a deductive relation can hold between premises and a conclusion even though one cannot reasonably infer that conclusion from those premises, for various reasons. Sometimes good reasoning leads one to abandon a premise rather than accept a conclusion. More generally, logic—the theory of deduction—is not a theory of inference.

Can any inductive methods be justified? It is sometimes said that the justification of methods of inference can only consist in adjusting those methods and one's various beliefs with each other until one reaches a *reflective equilibrium*. Although there is evidence that people do adjust methods and opinions in this way, there is also considerable evidence that the results are fragile and unreliable.

A better idea is that the justification of inductive methods must involve finding a way to assess the reliability of inductive methods, where reliability has to do with the statistical likelihood that the conclusions these methods lead to are correct. (It is hard to be in reflective equilibrium if one cannot believe one's methods of reasoning are reliable in this sense.)

Statistical learning theory is precisely concerned with assessing the reliability of inductive methods. How can data be used so as to arrive at a reliable rule for classifying new cases on the basis of certain values of observable *features* of those new cases?

In order to provide a formal mathematical theory, statistical learning theory appeals to the notion of a *D*-dimensional *feature space* in which each point represents a possible set of values of observable features. This framework assumes that an unknown probability distribution characterizes encounters with objects and the correlations between feature values of objects and their correct classifications. The unknown probability distribution determines the unknown best rule of classification, namely the Bayes Rule that minimizes expected error. In the simplest case, the same unknown probability distribution applies to the data as well as the new cases to be classified. It is also assumed in the simplest case that the probability of getting a particular object with such and such features and such and such a classification is independent of what other features and objects have or will occur.

For the case of a YES/NO classification, a classification rule can be identified with a set of points in feature space, perhaps certain disjoint areas or hyper-volumes, indicating which points in feature space are to be labeled YES and which are to be labeled NO. For example, linear rules divide the space into two regions separated by a line or plane or hyperplane.

Given a set *C* of rules, *enumerative induction* endorses a rule from *C* that minimizes error on the data. Enumerative induction makes sense only if there are significant limits on the rules included in *C*. Without such *inductive bias*, enumerative induction allows all possible inferences about new cases. On the other hand, if there are significant limits on the rules included in *C*, then, given enough data, it is highly probable that enumerative induction will endorse a rule whose expected error is close to the minimum error for rules in *C* (Vapnik and Chervonenkis, 1968).

Vapnik and Chervonenkis (1968) show that (subject to some very mild conditions) *no matter what the background probability distribution*, with probability approaching 1, as more and more data are considered, the expected error of the rules that enumerative induction endorses will approach the minimum expected error of rules in *C*, *if and only if C* has a finite index that has acquired the name, *VC dimension*.

VC dimension is explained in terms of *shattering*. Rules in *C* shatter a set of *N* data points if and only if, for every possible labeling of the *N* points with YESes and NOs, there is a rule in *C* that perfectly fits that labeling. In other words, there is no way to label those *N* points in a way that would falsify the claim that the rules in *C* are perfectly adequate.

The VC dimension of *C* is the largest number *N* such that some set of *N* points in the feature space can be shattered by rules in *C*. If for any *N*, there is a set of points that is shattered by rules in *C*, the VC dimension of *C* is infinite.

Vapnik and Chervonenkis show (again under very mild conditions) that, if and only if the set of rules *C* has finite VC dimension, expected error from the use of enumerative induction *uniformly converges* to the minimum expected error of rules in *C*. This means that it is possible to calculate for any given ε and δ, how much data are needed (no matter what the probability distribution) so that, with probability 1 - ε, the difference between the expected error and the minimum expected error of rules in *C* is less than δ.

This does not mean that the expected error of enumerative induction converges to that of a best possible rule, a Bayes rule, because the least expected error of rules in *C* may be greater than the expected error of a Bayes rule, perhaps much greater.

Enumerative induction uses data to choose a rule from *C* entirely on the basis of the empirical adequacy of the rule with respect to that data. There are alternative inductive methods that in choosing a rule from *C* balance such empirical adequacy against something else—some sort of simplicity perhaps.

What Vapnik calls *structural risk minimization* is an example of this second sort of inductive method. In structural risk minimization, the class *C* of rules is an infinite union of subclasses of increasing VC dimension. (The VC dimension of *C* is therefore infinite, so that enumerative induction may perform poorly.) For each rule in *C* let *m* be the number of the smallest subclass of *C* to which the rule belongs. Then structural risk minimization chooses a rule by balancing that number *m* against the rule's fit to the data.

Structural risk minimization allows *C* to be chosen in such a way that it contains rules whose expected error is arbitrarily close to the expected error of a Bayes rule. Furthermore, it can be shown that the method of structural risk minimization is *universally consistent* in the sense that the expected error of rules endorsed by such methods does

approach in the limit the expected error of a Bayes rule. But since the class *C* used in structural risk minimization has infinite VC dimension, we get no guarantee of uniform convergence to the expected error of the Bayes rule.

VC dimension is defined in terms of falsifiability in a way that is reminiscent of Popper. The VC dimension of a class of rules is at least *n* if there is a set of *n* points with the property that no assignment of labels to those points could falsify the hypothesis that some rule in *C* fits the data. Popper similarly measures the simplicity of a class of rules in terms of the amount of data needed to falsify the hypothesis that the correct rule is in that class. But there is a difference. And VC dimension turns out to be the more important characteristic of a class of rules than what might be called its "Popper dimension."

There are other universally consistent inductive methods. Certain *nearest neighbor* methods provide one example. There are also inductive methods that take into account not just the empirical adequacy of a rule but also the place of the rule in a well-ordering of the rules in *C*, e.g. by the length of its minimal description in a specified format.

In a final chapter, we explain how the ideas we have taken from statistical learning theory apply to the analysis of perceptrons and feed-forward neural networks that philosophers and psychologists often discuss as providing approximate models of aspects of the brain. We then go on to discuss a different model involving *support vector machines* (SVMs).

In effect, SVMs map the original feature space into a typically much higher dimensional space (sometimes even an infinite dimensional space) in which images of the data are linearly separated. We consider the possibility that SVMs might provide a useful model of psychological categorization that would compete with currently standard models.

We also discuss what Vapnik calls "transduction," a learning method that uses additional information beyond the labeled data—information about hard cases and about what cases have come up to be classified. The theory of transduction suggests new models of how people sometimes reason. The hypothesis that people sometimes reason transductively provides a possible explanation of some psychological data that have been interpreted as showing that people are inconsistent or irrational. It also provides a possible account of a certain sort of "moral particularism."

Gilbert Harman & Sanjeev Kulkarni

Princeton University

harman@princeton.edu / kulkarni@princeton.edu

Abstracta SPECIAL ISSUE III, pp. 10 – 17, 2009

COMMENTS ON HARMAN AND KULKARNI'S "RELIABLE REASONING"

Glenn Shafer

Gil Harman and Sanjeev Kulkarni have written an enjoyable and informative book that makes Vladimir Vapnik's ideas accessible to a wide audience and explores their relevance to the philosophy of induction and reliable reasoning. The undertaking is important, and the execution is laudable.

Vapnik's work with Alexey Chervonenkis on statistical classification, carried out in the Soviet Union in the 1960s and 1970s, became popular in computer science in the 1990s, partly as the result of Vapnik's books in English. Vapnik's statistical learning theory and the statistical methods he calls support vector machines now dominate machine learning, the branch of computer science concerned with statistical prediction, and recently (largely after Harman and Kulkarni completed their book) these ideas have also become well known among mathematical statisticians.

A century ago, when the academic world was smaller and less specialized, philosophers, mathematicians, and scientists interested in probability, induction, and scientific methodology talked with each other more than they do now. Keynes studied Bortkiewicz, Kolmogorov studied von Mises, Le Dantec debated Borel, and Fisher debated Jeffreys. Today, debate about probability and induction is mostly conducted within more homogeneous circles, intellectual communities that sometimes cross the boundaries of academic disciplines but overlap less in their discourse than in their membership. Philosophy of science, cognitive science, and machine learning are three of these communities. The greatest virtue of this book is that it makes ideas from these three communities confront each other. In particular, it looks at how Vapnik's ideas in machine learning can answer or dissolve questions and puzzles that have been posed by philosophers.

This leaves out, of course, many other communities that debate probability and reliable reasoning: mathematical probabilists, the many tribes of mathematical statisticians,

economists, psychologists, information theorists, and even other tribes of computer scientists, including those within machine learning who study prediction with expert advice [2]. The work of Vapnik and Chervonenkis is only a tiny part of the vast literature on probability and prediction that is relevant to philosophy's questions about induction and reliable reasoning, and the next step beyond Harman and Kulkarni's book is surely to try to fit what they have done into a larger picture.

From a historical viewpoint, and also from the viewpoint of modern mathematical probability, Vapnik's statistical learning theory makes a very special assumption: it assumes that repeated observations are drawn from the same probability distribution. As Harman and Kulkarni explain (p. 35), "we assume that the data represent a random sample arising from the background probability distribution, and we assume that new cases that are encountered are also randomly produced by that distribution." This is the famous assumption that observations are independently and identically distributed. It can be weakened slightly to the assumption that the observations are exchangeable – i.e., that their probability distribution does not change when the order is permuted. Are there really many applications where either assumption is reasonable?

Leibniz thought not. In 1703, Jacob Bernoulli wrote to Leibniz to explain how his law of large numbers would use past examples to find probabilities for future ones: "For example, if I perceive, having made the experiment in very many pairs of young and old, that it happens 1000 times that the young person outlives the old person and the reverse happens only 500 times, then I may safely enough conclude that it is twice as probable that a young person will outlive an old one as the reverse." Leibniz responded skeptically: "Who is to say that the following result will not diverge somewhat from the law of all the preceding ones – because of the mutability of things? New diseases attack mankind. Even if you have observed the results for any number of deaths, you have not therefore set limits on the nature of things so that they cannot vary in the future." (See [1], pp.38-39.)

The history of mathematical probability in the three centuries after Leibniz's exchange with Bernoulli can be framed as a continuation of their debate. Mathematicians continually refined Bernoulli's law of large numbers, but its success in applications was spotty. In the 19th century, Laplace's theory of errors of measurement reigned in astronomy,

while its applications in human affairs were rightly ridiculed. Frank Knight, founder of the Chicago school of economics, coined the distinction between "risk" and "uncertainty" to distinguish between the situation of an insurance company, which can count on the law of large numbers, and the situation of a businessman, who does not enjoy the luxury of many repeated chances under constant conditions.

The great accomplishment of mathematical probability during the twentieth century was to move beyond the picture of successive independent draws from a single probability distribution to the idea of a stochastic process, in which probabilities evolve. This change was already underway in the 1920s, with the explosion of work on Markov chains [3] and Weiner's application of functional analysis to model Brownian motion. It was consecrated in 1953 by Joe Doob's general framework for stochastic processes, which applied to continuous as well as discrete time [4]. By 1960, Jerzy Neyman could declare that science had become the study of stochastic processes [6].

Neyman saw four periods in the history of indeterminism in science:

1. *Marginal indeterminism*, the period in the 19th century when scientific research was indeterministic except in the domain of errors of measurement.

2. *Static indeterminism*, the period at the end of the 19th and beginning of the 20th century when populations were the main subject of scientific study, so that the idea of independent draws from populations was dominant.

3. *Static indeterministic experimentation*, the period from 1920 to 1940 when R. A. Fisher's ideas were dominant and the basic ideas of statistical testing and estimation were developed.

4. *Dynamic indeterminism*, already in full swing in 1960, when every serious study in science was a study of some evolutionary chance mechanism.

"In order that the applied statistician be in a position to cooperate effectively with the modern experimental scientist," Neyman declared, "the theoretical equipment of the statistician must include familiarity and capability of dealing with stochastic processes."

In the half century since Neyman wrote, the theory and applications of stochastic processes have developed as he envisioned. Natural science and economics are awash with dynamic stochastic modeling. How can it be, then, that Vapnik's work, based on the tired old idea of independent identically distributed observations, has suddenly emerged as so powerful, finding so many applications in biology and other data-rich domains?

Are these domains in which probabilities do not evolve? I doubt it. The data sets that people in machine learning use to test competing methods generally fails tests of exchangeability, so much so that it is standard practice to permute the order of the observations in these data sets before applying methods, such as support vector machines, that assume exchangeability.

In fact, the results of statistical learning theory that use exchangeability – the guarantees of accuracy based on finite Vapnik-Chervonenkis dimension, for example – are so asymptotic (require such cosmic sample sizes in order to give interesting bounds) that they have little to say about the success of support vector machines. (Concerning accurate confidence levels for successive predictions of a support vector machine or other prediction method when exchangeability does hold, see [8,9].)

The key to the success of support vector machines seems to lie elsewhere – in a feature of their implementation that Harman and Kulkarni mention on pp. 85-87: the mapping of data to higher dimensional spaces where classes can be more nearly linearly separated. This mapping is actually implemented implicitly with kernels, which assign to pairs of vectors in the original space the angles between the vectors to which they would be mapped if the mapping were spelled out. Such kernels were studied by probabilists and mathematical statisticians starting in the 1940s, but it was computer scientists implementing support vector machines who first took advantage of them on a large scale, to process the type of data that has now become so common in medicine and other branches of biology, where the number of individuals measured many be reasonable but an immense number of variables are measured on each individual.

Kernels are becoming increasingly important in computer science and mathematical statistics, not only in support vector machines but in other techniques as well. What is crucial in all cases is the choice of the kernel. Choosing the kernel means choosing what features of the observations we want to use for prediction. Choosing which measurements or which aspects of the measurement (the mapping the kernel represents is a mapping from the original measurements to their many aspects) has always been the central question for statistical prediction, and it becomes only more acute in the high-dimensional problems where support vector machines and other kernel techniques are so useful. If I were to fault Harman and Kulkarni on one point, it is that they do not dwell on the experimentation and reasoning that goes into choosing the kernel. This seems to be where applications of machine learning generate new knowledge, and we might learn something from a philosophical analysis. Is the choice of a kernel an example of induction? Is it inference to the best explanation?

One reason support vector machines can be successful in spite of the failure of the exchangeability that Vapnik assumes in all his theoretical work is that the machines rely not so much on stability of the probability distribution from which examples are drawn as on the stability of the relation between the features used for prediction and what is predicted. In order to make this point as clearly as possible, let us write x for the vector of measurements we use for prediction (the *object*) and y for what we predict (the object's *label*). An *example* is a pair (x,y). A kernel is a function K that assigns a real number to every pair of examples (x,y) and (x',y'). We observe n examples, say $(x_1,y_1),\ldots,(x_n,y_n)$ and a new object x_{n+1}, and we want to predict the label y_{n+1}. The support vector machine determined by a kernel K is a way of making this prediction. Exchangeability requires that the n+1 examples $(x_1,y_1),\ldots,(x_n,y_n),(x_{n+1},y_{n+1})$ all be drawn from the same probability distribution. In particular, the x_i should all be drawn from the same distribution. But many methods of prediction, including support vector machines, can do a good job even when the distribution from which the x_i are drawn varies, provided the dependence of y_i on x_i remains somewhat stable. It is enough, for example, if the conditional probabilities $P(y_i|x_i)$ do not change and y_i is independent, given x_i, of the earlier examples [10].

I would also like to add a thought to Harman and Kulkarni's discussion of the contrast between induction and transduction (pp. 90-94). As Vovk, Gammerman, and I argue in [9], the contrast may be clarified if we first discuss *on-line prediction*. When we talk about induction, we usually think about deriving a rule from a batch of examples, say $(x_1,y_1),\ldots,(x_n,y_n)$, and then using that rule for prediction in many future examples. But in an on-line setting, where we see example after example and predict y from x each time, it may be practical to update the prediction rule each time. We predict y_{n+1} from x_{n+1} using a rule we learn from analyzing $(x_1,y_1),\ldots,(x_n,y_n)$, but then we observe y_{n+1}, and so before predicting y_{n+2} from x_{n+2}, we get a new prediction rule by analyzing all n+1 examples $(x_1,y_1),\ldots,(x_n,y_n),(x_{n+1},y_{n+1})$. And so on. We use each rule only once. Is a rule that we use only once a rule? Is finding a rule that we immediately discard induction? The question is obviously relevant to cognitive science if some mental routines are slightly modified whenever they are used.

Finally, I would like to mention recent work on on-line prediction by Vovk, Takemura, and myself [7], which leads to some new insights into the question of whether examples need to be drawn from a probability distribution in order for good probabilistic prediction to be possible. In general, probability predictions are considered good if they pass statistical tests that compare them with what actually happens. Consider, for example, a forecaster who uses information x_i to give a probability p_i. for whether it will rain ($y_i = 1$) or not ($y_i = 0$). It is easy to construct statistical tests for whether the p_i agree with the y_i well enough, and these tests can be reframed as strategies for a gambler who tries to multiply the capital he risks by a large factor betting at the odds given by the p_i. It turns out that the gambler can combine these strategies into a single strategy, which involves a kernel that measures how much a new example (x_{n+1},y_{n+1}) is like old examples $(x_1,y_1),\ldots,(x_n,y_n)$. It also turns out that the forecaster can defeat such a strategy, regardless of how the weather turns out. We call this *defensive forecasting*.

The possibility of defensive forecasting means that good on-line prediction does not depend on examples being drawn from a background probability distribution. The most crucial question in prediction is not whether examples are being chosen from probabilities but whether the prediction is on-line. If the prediction is on-line, there are many ways it can

be done well. Some of these seem to involve the estimation of a background probability distribution, but this is illusory, for the estimate of the background probability distribution can change drastically as prediction proceeds. The important point is that no matter how reality actually chooses the y_i, you can give p_i that avoid extending any trends that might lead to statistical rejection of your forecasting.

In their discussion of reflective equilibrium (pp. 13-19), Harman and Kulkarni mention the situation of a juror, who is scarcely in an on-line setting. The opposing counsels will propose to the juror quite different sequences of examples in which the case at hand might be placed. How to choose? This is Reichenbach's problem of choosing a reference class. Philosophy has something to say here. Bayesians and non-Bayesian theories of subjective probability have something to say. Methods of machine learning, it seems, do not.

Glenn Shafer

Rutgers Business School

&

Department of Computer Science,
Royal Holloway, University of London

gshafer@rutgers.edu

References

[1] Bernoulli, J. (2006). *The Art of Conjecturing*. Translated with an Introduction and Notes by Edith Dudley Sylla. Johns Hopkins: Baltimore.

[2] Cesa-Bianchi, N. & Lugosi, G. (2006). *Prediction, Learning, and Games*. Cambridge University Press: Cambridge.

[3] Bru, B. (2006). Souvenirs de Bologne. *Journal de la Société française de Statistique,* 144:135-226.

[4] Doob, J.L. (1953). *Stochastic Processes*. Wiley: New York.

[5] Knight, F.H. (1921). *Risk, Uncertainty and Profit*. Houghton Mifflin: Boston.

[6] Neyman, J. (1960). Indeterminism in Science and New Demands on Statisticians. *Journal of the American Statistical Association*, 55:625–639.

[7] Shafer, G. (2008). Defensive Forecasting: How to Use Similarity to Make Forecasts that Pass Statistical Tests. Pp. 215-147 of *Preferences and Similarities*, edited by Giacomo Della Riccia, Didier Dubois, Rudolf Kruse, and Hans-Joachim Lenz, CISM Series, Springer: New York. See also paper 22 at www.probabilityandfinance.com.

[8] Shafer, G. & Vovk, V. (2008). 'A Tutorial on Conformal Prediction'. *Journal of Machine Learning Research*, 9:371-421.

[9] Vovk, V., Gammerman, A. & Shafer, G. *Algorithmic Learning in a Random World*. Springer: New York.

[10] Vovk,V., Nouretdinov, I. & Gammerman, A. On-line Predictive Linear Regression. To appear in *Annals of Statistics*.

Abstracta SPECIAL ISSUE III, pp. 18 – 26, 2009

INFERENCE TO THE BEST INDUCTIVE PRACTICES

Paul Thagard

Harman and Kulkarni (2007) provide a rigorous and informative discussion of reliable reasoning, drawing philosophical conclusions from the elegant formal results of statistical learning theory. They have presented a strong case that statistical learning theory is highly relevant to issues in philosophy and psychology concerning inductive inferences. Although I agree with their general thrust, I want to take issue with some of the philosophical and psychological conclusions they reach. I will first discuss the general problem of assessing norms and propose a metanormative decision procedure that can then be applied to the assessment of inductive methods. In particular, I will apply it to the assessment of the method of inference that Harman and Kulkarni call transduction.

METANORMATIVITY

Philosophy is inherently normative in that it is concerned not just with how people *do* think and act, but also with how they *ought* to think and act. Normative issues arise in general epistemological deliberations about how people should make inferences and in general ethical deliberations about the morality of kinds of actions. But normative issues also abound in much more specific practices, for example concerning how scientists should use theories and experiments to investigate the world and how doctors should decide what medical treatments are best. By a *norm* I mean a practice that ought to be followed in a particular domain. In this sense, Harman and Kulkarni can be viewed as working to help establish norms for inductive inference. By *metanormativity* I mean the meta-level consideration of how norms ought to be established (Hardy-Valée and Thagard, 2008).

In contemporary philosophy, the most popular approach to metanormativity is a method that John Rawls (1971) called *reflective equilibrium,* which involves an ongoing process modifying general principles and particular judgments to better accord with each

other. Harman and Kulkarni have legitimate worries about this method, particularly that it is often fragile and unreliable. I have expressed similar worries about reflective equilibrium myself (Thagard 1988, 2000), and Harman and Kulkarni erroneously extrapolate from my enthusiasm for coherence-based approaches to many kinds of inference that I endorse an approach to metanormativity akin to reflective equilibrium.

To clarify my position on how to arrive at reasonable norms about how people ought to think and act, I will now offer a decision procedure for metanormativity generalized from less explicit previous accounts (Thagard 1988, Hardy-Valée and Thagard 2008). Abstractly, the procedure requires the following steps:

1. Identify a domain of practices.
2. Identify candidate norms for these practices.
3. Identify the appropriate goals of the practices in the given domain.
4. Evaluate the extent to which different practices accomplish the relevant goals.
5. Adopt as domain norms those practices that best accomplish the relevant goals.

Of course, this kind of inference is revisable as new information comes in about goals and the efficacy of practices. I will illustrate how this procedure works by describing its application to a narrow medical domain.

In specific medical domains, physicians are concerned with establishing norms of diagnosis and treatment that prescribe the best practices for recognizing and improving patients' conditions. For example, in accord with step 1, we could identify as a narrow domain of practice the need of cardiologists to determine how to deal with patients who have arteries sufficiently blocked that chest pain results. Step 2 requires identifying candidate best practices, such as bypass surgery, stent insertion, treatment with drugs for controlling blood pressure and cholesterol levels, or watchful waiting. A major way of accomplishing step 2 is to survey the broad range of current practices, but a more creative approach requires considering new kinds of treatments that need to be invented or refined. Step 3 requires figuring out what are the goals of treating people with blocked arteries, which might include improving the patients' life expectancy and quality of life, but could

also include matters of cost to the public health system in countries sufficiently civilized to have one. Step 4 requires the onerous process of determining as much as possible the extent to which each of the different processes accomplishes the various goals of treating arterial disease. Once this is done using the best available medical evidence, we should have a good idea of the best practices for treatment, which can then be adopted as norms in step 5.

This decision procedure can also be applied to the higher-order problem in step 4 of how to assess the efficacy of various treatments. A positive recent movement is *evidence-based medicine,* which advocates replacing unreflective judgments based on clinical experience with systematic evaluation of the best kinds of experimental evidence, particularly data gathered from randomized, controlled, clinical trials (e.g. Guyatt et al 1992). I won't attempt it here, but I would argue that the norms of evidence-based medicine should be adopted as best practices for establishing lower-level norms for particular kinds of treatment. Another interesting exercise would be to apply my decision procedure for metanormativity to itself, arguing that it is a better practice for adopting norms than such alternatives as reflective equilibrium and a priori reasoning. Instead, I want to apply my metanormativity decision procedure to the central problem of Harman and Kulkarni's book, establishing norms for inductive inference.

EVALUATING INDUCTIVE METHODS

The domain of practices now to be evaluated concerns inductive inferences, ones that introduce uncertainty (step 1). Despite the risk of making errors by inferring false conclusions, people have many ways of making inductive inferences, including: generalizing from samples to populations, using probability theory, statistical inference, the hypothetical-deductive method, inference to the best explanation, analogy, and wishful thinking. Step 2 of my decision procedure requires identifying these and many other possible ways of reasoning inductively in order to be able to assess which of them should be adopted as norms of inductive inference. Harman and Kulkarni are concerned with only a small range of kinds of inductive inference that can be analyzed using the tools of statistical learning theory.

Step 3 is much more problematic. Harman and Kulkarni interpret the philosophical problem of induction as the problem of the *reliability* of inductive inference. They do not define reliability, but I presume they mean something like the ratio of the number of true conclusions to the number of all conclusions reached (Goldman 1986). However, they seem to presuppose that reliability is the *only* goal of inductive inference, but I find this implausible.

First, in both science and everyday life, the goals of inductive inference include understanding as well as reliability. We want not only to achieve truths and to avoid error, but also to grasp why things happen. In everyday life, emphasis on reliability alone would restrict us to a kind of behaviorism, noting regularities in how people respond to their environments. But people cannot resist attributing mental states to each other, going beyond behavior to infer that people have various beliefs, desires, and emotions that cause their behavior. This ancient kind of reasoning was felicitously dubbed "inference to the best explanation" by Harman (1965). Inference to the best explanation can take us beyond observed regularity to non-observable states that tell us why regularities occur. It often leads to false conclusions, for we often err in our judgments about the mental states of other people, and even sometimes err about our own mental states. Still, inference to the best explanation about mental states should not be eschewed, because there is no better way of figuring out why people behave as they do.

More systematically, inference to the best explanation is an established part of the practice of inductive inference in science (Thagard 1988, 1992). In physics, chemistry, biology, and medicine, science has progressed by inferring the existence of entities that are not directly observable, such as electrons, molecules, genes, and viruses. This kind of inference is frequently *un*reliable, as we see from the pantheon of scientific mistakes that includes inferences to the existence of such discredited entities as phlogiston, caloric, and the luminiferous aether. Ultracautious positivist and empiricist philosophers of science want to stick closely to observation, but without the goal of understanding why things happen we would not have the best current theories. Hence we should include understanding beside reliability as a main goal of inductive inference.

Even more controversially, I would like to propose another goal to be used in evaluating competing methods of inductive inference: potential for practical importance. At any given moment, there is a huge range of inductive inferences that a person might make. I might devote the rest of the afternoon to collecting evidence that would support the generalization that all the items in my study weigh less than 100 kilograms. My inference would be reliable, but pointless. Inductive methods should be capable of producing conclusions that are useful, not just true. Once again, inference to the best explanation to non-observed entities wins out over more restrictive inductive methods that might be more reliable. Without theories about non-observable entities such as electromagnetic radiation and viruses, we would not have such marvels of modern technology as computers, television, and antiviral drugs. Hence I think that Harman and Kulkarni are unduly restrictive in considering only reliability as the concern of inductive reasoning.

Step 4 of my metanormativity decision procedure recommends assessing different methods with respect to all relevant goals. For inductive inference, the goals certainly do include reliability, and the assessment of inductive methods can legitimately be formal as well as empirical. Statistical learning theory strikes me as highly useful for assessing some inductive practices with respect to that particular goal, and I was impressed by the insights provided by Harman and Kulkarni concerning the VC dimension. My point is just that the goals and range of practices of inductive inferences are far broader than can be studied using these formal methods.

TRANSDUCTION

A particularly interesting part of the discussion by Harman and Kulkarni is their treatment of Vapnik's theory of transduction, which they mark as being novel in two main aspects. First, transduction does not involve inferring an inductive generalization that is then used for classification, but instead proceeds directly from information about previous cases to classification of new ones. Second, transduction uses information about what new cases have come up in its classification of them. This second aspect is indeed novel, but the first has a long history in philosophical and psychological discussions.

On the standard view, inductive inferences go from cases to rules, which can then be applied to new cases. The alternative view that inference can go directly from cases to cases was proposed by John Stuart Mill in the nineteenth century (Mill 1970, p. 364 – Ch. XX of Book III of the eighth edition):

> But we conclude (and that is all which the argument of analogy amounts to) that a fact *m,* known to be true of A, is more likely to be true of B if B agrees with A in some of its properties, (even though no connection is known to exist between *m* and those properties) than if no resemblances at all could be traced between B and any other thing known to possess the attribute *m.*

Similarly, Bertrand Russell (1967, p. 44) wrote early in the twentieth century:

> But the newness of the knowledge is much less certain if we take the stock instance of deduction that is always given in books on logic, namely, 'All men are mortal; Socrates is a man, therefore Socrates is mortal.' In this case, what we really know beyond reasonable doubt is that certain men, A, B, C, were mortal, since, in fact, they have died. If Socrates is one of these men, it is foolish to go the roundabout way through 'all men are mortal' to arrive at the conclusion that *probably* Socrates is mortal. If Socrates is not one of the men on whom our induction is based, we shall still do better to argue straight from our A, B, C, to Socrates, than to go round by the general proposition, 'all men are mortal'. For the probability that Socrates is mortal is greater, on our data, than the probability that all men are mortal. (This is obvious, because if all men are mortal, so is Socrates; but if Socrates is mortal, it does not follow that all men are mortal.) Hence we shall reach the conclusion that Socrates is mortal with a greater approach to certainty if we make our argument purely inductive than if we go by way of 'all men are mortal' and then use deduction.

Thus something like transduction has long been recognized by philosophers.

Similarly, contrary to what Harman and Kulkarni suggest, the importance of inference from cases to cases has also been noted by many psychologists, from at least three different perspectives. First, there is a large body of psychological research on analogical inference, which is obviously inference from cases to cases without intervening generalizations, although its cognitive structure is more complex than Mill recognized (e.g. Holyoak and Thagard, 1995). Second, one prominent theory of concepts is called the *exemplar* view, which proposes that concepts are not stored as general representations but merely as collections of particular cases that are then used to classify new cases (e.g.

Murphy 2002, ch. 4). Third, the standard interpretation of neural networks that learn from examples using backpropagation is not that the connection weights encode rules, but that they encode statistical patterns rather than generalizations (e.g. Rumelhart and McClelland 1986). Thus the aspect of transductive inference that it goes from cases to cases without intervening rules has a strong place in psychology as well as philosophy.

Should people use transduction? It would seem to support the inductive goal of reliability, because going from cases to cases avoids the danger of overgeneralization that inferring rules can easily introduce, as the quote from Russell suggests. However, inductive generalization may have some benefits with respect to explanation, if it leads to the adoption of causal rules that can explain why things happen. Another currently prominent theory of concepts emphasizes their role not just in classification but in explanation (Murphy 2002, ch. 6). I may, for example, have seen many cases of drunks that I can use to classify a new staggering person as a drunk, but an inductive generalization may enable me to explain why he is drunk. For example, the generalization that people who drink a lot of alcohol lose motor control provides a causal explanation of why someone is staggering.

My third goal of inductive inference was practical importance, and it is here that the second aspect of transduction seems most relevant. Because transduction takes into account information about new information to be classified, it can potentially come up with more useful new classifications. I cannot think of any current psychological theory of concept learning that has this property, nor of any experimental evidence that people have the capability of using such anticipations in their concept learning. However, a notable aspect of the multiconstraint theory of analogical inference proposes that the retrieval, mapping, and application of cases to be used as analogies all involve the constraint of purpose, which concerns the practical use of the analogy (Holyoak and Thagard 1995). So perhaps analogical inference counts as transductive both in the sense of going from cases to cases and in the sense of taking into account the characteristics of cases to be inferred about. Purpose may not increase reliability in the abstract sense of improving the truth to error ratio, but should make case-to-case inferences more useful for the various problem-solving goals of analogical inferences.

In conclusion, Harman and Kulkarni's *Reliable Reasoning* is a highly informative and stimulating exploration of important topics in inductive inference, but many important descriptive and normative questions about induction inference require further investigation.

Paul Thagard

University of Waterloo

pthagard@uwaterloo.ca

References

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

Guyatt, G. , et al. (1992). 'Evidence-based medicine: A new approach to teaching the practice of medicine'. *Journal of the American Medical Association, 268*, 2420-2425.

Hardy-Vallée, B., & Thagard, P. (2008). 'How to play the ultimatum game: An engineering approach to metanormativity'. *Philosophical Psychology, 21*, 173-192.

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

Harman, G., & Kulkarni, S. (2007). *Reliable reasoning: Induction and statistical learning theory*. Cambridge, MA: MIT Press.

Holyoak, K. J., & Thagard, P. (1995). *Mental leaps: Analogy in creative thought*. Cambridge, MA: MIT Press/Bradford Books.

Mill, J. S. (1970). *A system of logic* (8 ed.). London: Longman.

Murphy, G. L. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.

Rawls, J. (1971). *A theory of justice*. Cambridge, MA: Harvard University Press.

Rumelhart, D. E., & McClelland, J. L. (Eds.). (1986). *Parallel distributed processing: Explorations in the microstructure of cognition*. Cambridge MA: MIT Press/Bradford Books.

Russell, B. (1967). *The problems of philosophy*. Oxford: Oxford University Press.

Thagard, P. (1988). *Computational philosophy of science*. Cambridge, MA: MIT Press/Bradford Books.

_____. (1992). *Conceptual revolutions*. Princeton: Princeton University Press.

_____. (2000). *Coherence in thought and action*. Cambridge, MA: MIT Press.

Abstracta SPECIAL ISSUE III, pp. 27 – 41, 2009

REMARKS ON HARMAN AND KULKARNI'S "RELIABLE REASONING"

Michael Strevens

Reliable Reasoning is a simple, accessible, beautifully explained introduction to Vapnik and Chervonenkis's statistical learning theory. It includes a modest discussion of the application of the theory to the philosophy of induction; the purpose of these remarks is to say something more.

1. A Patient Pessimist's Guide to Induction

Philosophical Learning Theory

Vapnik and Chervonenkis's statistical learning theory may be compared to formal learning theory, familiar to philosophers from the work of Putnam (1963) and Kelly (1996). There are significant technical differences between the two theories, but considered as philosophical frameworks for thinking about inductive reasoning, they have much in common. I will say that they are both—in their epistemological incarnations—species of *philosophical learning theory*.

The programmatic goal of formal learning theory is to investigate methods for learning from experience that are guaranteed to converge on the truth. (or at least guaranteed to come as close as possible) under some given set of circumstances. If you have a method that is sure to converge, the thought goes, then provided that the particular circumstances within which the guarantee is offered actually hold, you are sure to find the truth—eventually. The problem of induction is in that case solved. Or at least, *a* problem of induction is solved, since different methods may be recommended in different circumstances.

Statistical learning theory takes a similar approach; the most important difference for philosophical purposes is that, assuming as it does that we live in an inherently stochastic world, it does not pursue convergence per se but a kind (or several kinds) of probabilistic convergence. Rather than providing a guarantee of convergence on the truth, it

will provide, where it can, a probability of finding the truth or something close to the truth that converges to one, so that in the limit the probability of failing to find the truth is zero. For brevity's sake, I will use the term *convergence* in what follows to mean either true convergence or probabilistic convergence (in all its varieties); none of what I want to say turns on the distinction.

It is this approach to justifying induction by finding guarantees of long run convergence that I refer to as philosophical learning theory, or PLT.[1] Philosophical learning theory's approach to inductive reasoning manifests an interesting mix of daring and pessimism, which I will discuss in the remainder of this section. As you will see, Harman and Kulkarni do not advocate these tenets of PLT explicitly; you may think of what follows as an interpretation and an extrapolation of their philosophical hopes for statistical learning theory, intended to draw them out.

Pessimism

Philosophical learning theory's pessimism lies in the insistence that the best inductive methods are those that minimize, and if possible eliminate, the possibility of failure, no matter how unlikely the failure might be. Given a method that converges quickly on the truth in almost every world but misses it altogether in some, and a method that always learns the truth but only very slowly, the PLT theorist by temperament prefers the latter, in order to deal with the kind of Humean skeptical worries that must be overcome to vindicate induction.[2] This predilection for safety is very much in evidence in Harman and Kulkarni's book: they are interested in methods that are guaranteed to converge (probabilistically) on the truth *in any kind of world whatsoever*, provided only that the world contains something

[1] Even the convergentist strand of Bayesian confirmation theory is by this criterion a kind of PLT—though Bayesians also have many non-learning theoretic tricks up their sleeves.

[2] That said, the tools provided by statistical and formal learning theory may be of considerable help to a learner with the former preference. Likewise, they may be useful given goals other than finding the truth, including goals that care about a mix of truth and other properties; as formal methods, they are not constrained by philosophical ideology.

appropriate to converge on, in the form of a (possibly stochastic) regularity linking the phenomena under examination.[3]

To give you a sense of where this caution might lead, consider a method discussed by Harman and Kulkarni that has some particularly desirable properties from their learning-theoretic point of view, the nearest-neighbor rule. To make a prediction about a new data point, the nearest-neighbor rule polls a certain number of existing data points that are most similar to the new point in known respects; it then predicts that the as-yet-unknown features of the new data point will take on values similar to their values in these nearest neighbors. For example, to predict the fitness (in a given environment) of a newly discovered variant of some bacterium, you might look at the known variants of the organism most similar to the new variant and use the mean of their fitnesses in the environment as an estimate.

As a quick-and-dirty heuristic to be employed when there is no deeper theoretical understanding available, this seems quite unobjectionable. But if PLT is to be taken seriously as a theory of empirical enquiry in science, then it should be understood as delivering not just heuristics but final theories. So we have to consider the possibility of a science that has, as its core theoretical posit, a nearest-neighbor rule. What would such a science look like? Rather than being centered on a few simple, far-reaching laws of nature, or a small set of general schemas for building (say) causal models of a wide range of phenomena, or even a large number of phenomenological generalizations, this science would have at its heart nothing more than an enormous and ever-growing data bank. Predictions would be made solely by consulting the information in this data bank.

That such a method might be recommended to science is a sign of the powerful conservativism—perhaps a better word would be paranoia—that orients one axis of philosophical learning theory: above all, says PLT, do not allow Nature to take you by surprise.

You might wonder—I certainly do—whether Harman, a long-time defender of inference to the best explanation, can reasonably be regarded as advocating this sort of

[3] To suppose the existence of such a regularity is to suppose a certain kind of uniformity in nature—though even given such an assumption a version of the problem of induction can be posed, as Harman and Kulkarni show in §3.8.

extreme empirical caution. But *Reliable Reasoning* is quite coy when it comes to such questions. It tenders statistical learning theory as a topic of interest to philosophers, but it does not develop to any degree the philosophical application of the theory as I have here. I am eager to learn more.

Daring

Philosophical learning theorists are not simply, in the face of Nature's awesome variety, pessimistic; they are at the same time, in another respect, rather daring.

Their daring lies in their devil-may-care attitude to the hypotheses recommended by their methods in the short term (where the short term is in fact any finite term). Whereas a traditional confirmation theorist is at pains to say that, after a certain amount of data has been collected, we are—in many circumstances at least—justified in believing the hypotheses recommended by our inductive methods, the PLT theorist will allow no such thing. Justification, in their view, applies to methods—in virtue of their convergence properties—but not to the beliefs endorsed by those methods. At best, the beliefs have an incidental significance: like wood shavings on the floor after a particularly fulfilling carpentry session, we may regard them with satisfaction as byproducts of aptly applied technique.

Where is the daring? Consider the ancient example of the traveler about to board the aircraft. To this individual, the PLT theorist cannot say: you can fly with confidence, because we have very good reason to think that the theory of aerodynamics that supplies the basis for the aircraft's construction is true (or at least, empirically adequate). They must say instead: the best we can say about our current theory of aerodynamics is that it was arrived at by the application of a method that will *eventually* lead to the truth. Who knows—perhaps we are there already. Good luck!

How does the PLT theorist react to the confirmation theorist's reassuring words? They are merely words, signifying nothing of true epistemic value. For all we know, we could discover tomorrow that our best aerodynamic theory is false. Such are the consequences of taking the problem of induction seriously. Indeed, in their attitude to the scientific method, PLT theorists have much in common with Popperians, although where

Popper claims to have avoided induction altogether, PLT theorists claim to have solved (or at least to have ameliorated) Hume's problem.

In this discussion of the epistemology of PLT, I have been for the most part channeling Kelly (1996) and Glymour and Kelly (2004). What do Harman and Kulkarni think? It is, again, difficult to tell: they do not provide any story as to how to regard the hypotheses recommended by their learning methods. One thing seems clear, though: on Harman and Kulkarni's variant of PLT it would be utterly unreasonable to expect these hypotheses to capture the truth. Let me explain.

Harman and Kulkarni, following Vapnik and Chervonenkis, set things up as follows. The basic inductive task is to predict, given the known properties of a specimen, certain of its as-yet unknown properties. In the simplest case, there are a number of properties that are known, and one qualitative property that is unknown. The task is to predict whether or not this latter property is present. One class of cases that fits this description are categorization tasks: you are presented with, say, pictures of a wide range of animals, and your job is to say whether or not each animal is a dog.

The "truth" is assumed to be some probability distribution relating the various properties, for example, a distribution giving the probability that an animal is dog conditional on its having such and such appearances. But the goal of the formal learner in Vapnik and Chervonenkis's theory is not to learn this stochastic truth. It is rather to arrive at a *deterministic* prediction rule that minimizes predictive error. (Various other goals are also possible, but it is invariably a deterministic rule that is sought.) Harman and Kulkarni call this rule (following a tradition in statistics) the "Bayes Rule" for the particular predictive problem and probability distribution. To avoid confusion, let me call it instead the *Best Rule*. The goal of Vapnik and Chervonenkis's theory, then, is to find an inductive learning procedure that, in the limit, learns the Best Rule, or comes as close as possible, in any circumstances.

Clearly, unless the truth happens to be deterministic, the Best Rule is not the True Rule. The philosophical learning theorist's guarantee of convergence on the Best Rule is therefore not a guarantee of convergence on the truth, but rather of convergence on a particularly useful deterministic heuristic. How, then, to regard whatever hypothesis is

recommended by the learning theorist at any time? It seems that even in the limit, your attitude to this hypothesis must be pragmatic—you may regard the hypothesis as useful, but not as true. Indeed, you will be pretty sure that it is false. And in the short term, of course, you cannot know, or even (following Kelly and Glymour) have any evidence for, the proposition that the hypothesis has predictive value.

What, then, should philosophers of science take away from all of this? That the idea of evidential support in science is a fallacy? Is this supposed to be a revisionary position, a proposal for the epistemic reform of the scientific method? Or is it supposed to be consistent with scientists' most deeply held epistemic beliefs? I wish that Harman and Kulkarni had given us answers to these Popperian questions.

2. Simplicity and VC Dimension

VC Dimension

In this second part of my remarks I want to focus on the idea that Vapnik and Chervonenkis's theory supplies, in its notion of the VC dimension of a set of inductive rules, an interesting surrogate for simplicity in scientific reasoning. Let me begin with a brief overview of the principal role played by the VC dimension in statistical learning theory.

Vapnik and Chervonenkis are concerned not only with the problem of converging on the predictively Best Rule, but also on the problem of converging on the best rule in any given set of rules, if the set in question does not contain the Best Rule itself.

Suppose that you start out with an inductive bias: your learning method, rather than taking into account every possible inductive rule configured to the question at hand, will consider only those inductive rules falling into a certain class. Call the rules in this class the *workable rules*. (The workable rules are not the logically possible rules, then, but the logically possible rules that you are willing to countenance.)

You would like to find the best rule in your set of workable rules. Ideally, of course, you would like to find the Best Rule itself, which is possible only if the Best Rule is in the workable set. There is a tension between these two desiderata. On the one hand, the larger

the set of workable rules, the more likely it is to contain the Best Rule. On the other hand, the larger and more complicated the set of workable rules, the harder it is to find the best workable rule. Vapnik and Chervonenkis's theory gives some mathematical substance to this latter claim, from the characteristically pessimistic learning-theoretic point of view. It defines a simple and intuitive learning method called enumerative induction—simply the method of choosing from the workable set the rule that best fits the data observed so far—and it states a condition on the set of workable rules that is necessary and sufficient for enumerative induction to converge, in the limit, on the best workable rule.

That condition is as follows: the set of workable rules must have a finite VC dimension. I refer you to Harman and Kulkarni's excellent discussion for a definition of VC dimension that is better than anything I can squeeze in here. But the idea is roughly as follows. Familiar from the philosophical literature on "curve-fitting" is the idea that some families of curves have more leeway to fit the data than others. Linear functions are quite constrained in their ability to intersect with, or even to come close to, a set of two-dimensional data points; high-order polynomials and certain trigonometric functions have more "wiggle room" (see *Figure 1* below).

The "wiggling" in question is the adjustment of parameters; while most cubic polynomials will come nowhere near fitting a set of four data points, there is sure to be *some* cubic that fits them exactly, that is, some choice of parameters that delivers a cubic that gets everything exactly right.

Think of the VC dimension of a family of rules as being a kind of measure of wiggle room—very loosely, the maximum number of data points that can be accommodated by choosing the right member of the family.

How is the VC dimension of a family of workable rules related to the problem of finding the best rule in that family? Given a workable family with VC dimension n, you need at least n data points before you get to the point where in every case (that is, for every possible set of n data points), the rules that are best at predicting the data so far agree to some extent on what will happen in the future. To take a simple example, suppose that your workable family is the linear rules—you are restricting yourself to linear hypotheses—and that you have only a single data point so far. Say that the set of rules that intersects, or

comes close to intersecting, this data point are the "empirically adequate" rules in the workable set. At this stage, the empirically adequate rules give you no guidance about the future: for any future point that might come along, there is some rule that matches the data and also intersects this new point. Clearly, you cannot converge on the best rule in the workable set in any useful sense without first getting to the stage where the empirically adequate rules do agree on the future, since convergence requires that they say roughly the same thing about the future as the best rule. It follows that the workable set's having a finite VC dimension is a necessary condition for this method of enumerative induction to converge on the best rule in that set—otherwise, there is always some arbitrarily large amount of data that, with respect to the workable set, badly empirically underdetermines the future. Vapnik and Chervonenkis show that having finite VC dimension is also a sufficient condition for convergence.

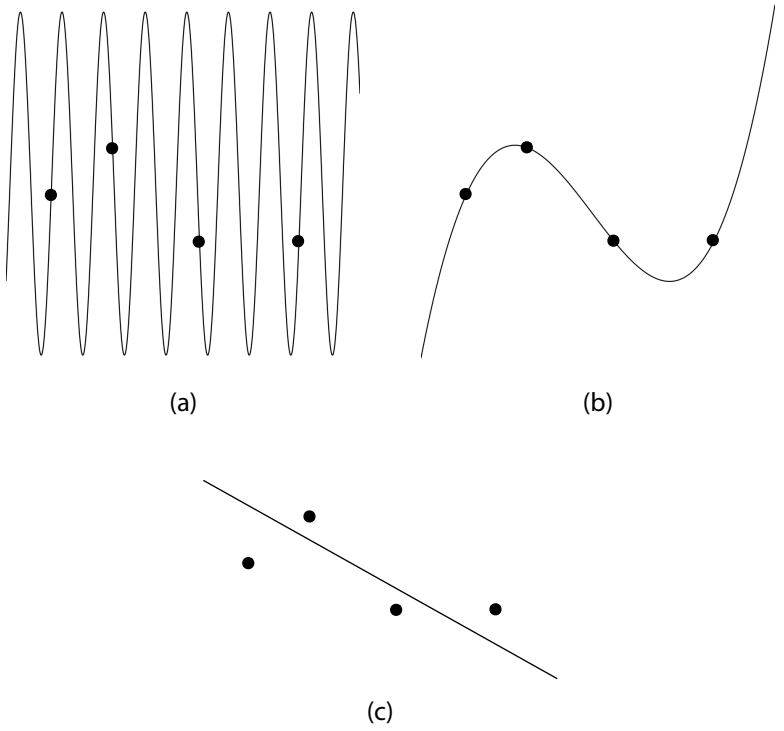

Figure 1: Curve-fitting: Any four data points can be fitted accurately by (a) some sine curve ($y = a \sin bx$) and (b) some cubic ($y = ax3 + bx2 + cx + d$), but not necessarily by (c) a line ($y = ax + b$).

The Workability Hierarchy

Now, you might think that the setup of the Vapnik and Chervonenkis theory is limiting in the following way: though scientists may start out by limiting themselves to the rules in a certain set—as some mid-century social scientists limited themselves to linear causal models—they do not thereby foreclose all possibility of moving outside the set if they run into trouble. They might start out with a fairly simple set, then, but if it becomes clear that the best rule in the set is not very good, they may move on to a set with more "wiggle room"—with a higher VC dimension.

Vapnik and Chervonenkis represent this more open-minded learning procedure as follows. You designate a "workability hierarchy": a sequence of sets of rules with ever greater VC dimension. The first set in the sequence might have a VC dimension of 2, the second set a VC dimension of 3, and so on. You then begin to collect data. Rather than explicitly restricting yourself to the first set of rules in the sequence, you rather choose the rule at any given time that minimizes the combination of empirical error on the existing data and VC dimension (thus departing from the method of enumerative induction, which minimizes empirical error alone). That is, you find a rule that fits the data well while appearing as early as possible in the workability hierarchy, or in other words, a rule that retrodicts what has been observed without exploiting a greater than necessary degree of "wiggle room". Vapnik and Chervonenkis call this method "structural risk minimization".

Harman and Kulkarni argue that the virtues of such a method underlie science's preference for simple over complex theories, in those cases where the conditions required for statistical learning hold. In doing so, they imply that VC dimension provides a good approximation for, perhaps even a good account of, simplicity in at least some parts of science. It is this claim that I wish to examine.

Simplicity in Science

Here are some things that are said of simple hypotheses in science, listed in a non-exhaustive spirit:

1. Simpler hypotheses are less hospitable to ad hockery—they offer less "wiggle room", and so are harder to fit to a given set of data.
2. Simpler hypotheses are easier to falsify.
3. We should prefer simpler hypotheses.
4. Simpler hypotheses are more likely to be true.
5. Simpler hypotheses make better explanations.

An implicit "all other things being equal" rider should be understood as attached to each of these maxims; maxim (5), for example, does not imply that simplicity is the sole factor that affects a hypothesis's explanatory potential, but rather that it is one such quality.

Note that hypotheses here should be understood as encompassing *families* of possible laws of nature—a hypothesis about two variables might state that they stand in a linear relationship, then, but it will not specify a value for the constant of proportionality. Throughout this discussion, I will assume that all questions about simplicity are asked with respect to well-defined families. In reality, of course, things are not so straightforward: you may be asked about the simplicity of a particular putative law of nature, constants and all, and it may not be clear what family of laws it should, for the purpose of answering the question, be considered to belong to.

The five simplicity maxims can be subdivided into three groups: (1) and (2) concern *accommodation*, (3) and (4) concern *acceptance*, and (5) concerns, of course, *explanation*.

Accommodation

Insofar as the central concept of the theory of accommodation is room, or more precisely, wiggle room—the ability to find space for whatever data come along—the notion of VC dimension is obviously well equipped to play the role of simplicity in the accommodation-related maxims.

Harman and Kulkarni champion the VC-dimension notion over other notions of simplicity for this reason. In particular, they criticize Popper's suggestion that the simplicity of a hypothesis is proportional to its number of adjustable parameters. These two characterizations of simplicity sometimes come apart, Harman and Kulkarni argue: the

family of sine curves $y = a \sin bx$ has only two adjustable parameters yet has a great deal of wiggle room, as you can see from figure 1, and so a high—indeed, an infinite—VC dimension (Harman and Kulkarni, 72). Consequently, the hypothesis that some phenomenon is characterized by a sine curve counts as simple on Popper's account and as complex on Harman and Kulkarni's account; because the hypothesis can accommodate almost any set of data points and so is difficult or impossible to falsify, it is clearly Harman and Kulkarni's rather than Popper's definition of simplicity that vindicates the maxims in this particular case.

Acceptance

Next, the role of simplicity in deciding whether or not to accept a hypothesis. A preference for simple hypotheses may be motivated in various ways. One way is articulated by maxim (4), according to which, all other things being equal, a simpler hypothesis is more likely to be true. If two hypotheses fit the data equally well, then, we will be on safer ground if we choose the simpler of the two. Such a motivation does not suit the PLT theorist, however, who has no truck with the probabilities of particular hypotheses at particular times (see "Daring" above).

A more pragmatic approach to justification is germane to PLT. In its most straightforward form, it might run as follows: it is more expensive to engineer a complex hypothesis than a simple hypothesis. Thus we will save money by sticking to the simplest hypothesis that fits the data reasonably well.

Does the VC-dimension notion of simplicity fit this line of reasoning? I am not sure. Insofar as a complex hypothesis (in the VC-dimension sense) offers more "wiggle room", it might be less expensive to start out a scientific investigation with an all-purpose complex hypothesis and optimize the wiggling process (writing highly efficient computer programs to compute the best values for the parameters and so on) than to start with simpler hypotheses and then retool every time they fail to fit the data (or at least, every time they fail in the kind of ongoing, discouraging way that suggests that fresh ideas are needed).[4]

[4] For a pragmatic defense of a preference for simplicity that turns on this very issue of the costs of cognitive retooling when evidence forces a theory to be "retracted", see Kelly (2007).

Indeed, why not begin with the family of sine curves? It is simple to compute, and can be made to fit almost any data. The optimal choice of starting point from a cost-benefit point of view might well be, then, a family such as the sine curves that is structurally very simple (simple, perhaps, in Popper's sense) but that has a very high VC dimension.

Is this an argument that the VC-dimension notion of simplicity is unsuitable for the purposes of making a pragmatic case for preferring simple hypotheses, or is it an argument against the pragmatic case itself? A bit of both, I think: on the one hand, what makes the sine family attractive to the practically-minded curve-fitter is its simplicity in some sense not captured by its VC dimension; on the other hand, it is unclear to me, given the merits of the "start out complex and optimize the wiggling" strategy of the previous paragraph, that cost-benefit considerations could ever fully motivate our preference for theoretical simplicity. In this latter respect, I suppose that I am unable to escape the pull of Glymour's (1980) suggestion that the fundamental problem with hypotheses that are overly complex with respect to the available data is that they contain content that is not empirically confirmed by the data.

Explanation

My final topic is explanation. The explanatory maxim (5) is perhaps the most controversial of the group, in the sense that a substantial number of philosophers would deny that simplicity per se has any role to play in explanatory goodness at all. (Except, that is, as a sign of some deeper virtue: a causal theorist of explanation would concede, say, that explanations that omit causally irrelevant factors are both better and simpler explanations than those that include them, but here simplicity is a mere byproduct of the requirement of causal relevance.)

One account of explanation, however—the unification account—invokes simplicity explicitly as a desideratum (Friedman 1974; Kitcher 1989). Could the VC-dimension notion of simplicity be useful to a unification theorist?

Let me answer this question with the help of an example. I take it that a paradigm of explanatory simplicity for a unificationist is Newton's gravitational theory. With only the

geometry of space and time, the three laws of motion, and the gravitational force law, Newton is able to explain a vast range of phenomena.

That "vast range" should give you pause. How high, exactly, is the VC dimension of the Newtonian theory? It is not immediately clear. On the one hand, the theory articulates a tight constraint on the movements of any object, given the properties and movements of all the other objects. The tightness of this constraint suggests a lack of wiggle room. On the other hand, what matters for VC dimension is not the wiggle room given all the other objects, but the wiggle room given all the other *known* objects. In this respect Newtonian theory offers quite a bit of wiggle room, as several famous episodes from the history of science, each involving the positing of unseen matter, will remind you. The first is the postulation of the planet Neptune to explain irregularities in the orbit of Uranus. The second is the postulation of the (in fact non-existent) planet Vulcan to explain irregularities in the orbit of Mercury. The third (not an amendment to *Newtonian* theory, but you get my point) is the postulation of dark matter to explain irregularities in the internal movements of galaxies.

Of course, there are checks on these acts of accommodation, most of them coming from outside Newtonian gravitational theory itself. But the theory must, I think, be credited with a fairly large suite of empirical rooms, or in other words, an impressive power to accommodate any given set of data. I want to suggest that this ability to accommodate does not in any way undermine the simplicity of Newtonian explanation, in the sense that matters to a unificationist. If anything, quite the contrary: the unifying power of Newtonian theory comes in part from its uniform applicability to all matter, yet this same applicability is what enables the ad hoc postulation of additional, unseen bodies to account for empirical anomalies. Any deep unification, I suggest, will tend allow such strategies; thus, the sense of simplicity employed by the unificationist to capture this kind of depth will not coincide with the VC-dimension notion of simplicity.

Perhaps there is hope for a revised notion of simplicity, useful to a unificationist, based on the same mathematical ideas as the VC dimension. This revised notion would attend to a hypothesis's ability to fit the totality of relevant facts—not just the known bodies, but all the bodies. In this respect, as I remarked above, Newtonian theory does seem

to offer a tight constraint on what is allowed to happen, in the sense that adjustments of its one parameter—the gravitational constant—give you very little leeway when it comes to fitting the complete set of facts about the motion of massive bodies in space and time.

Then again, does the Newtonian theory have only one adjustable parameter? As Einstein so fruitfully remarked, the theory employs two notions of mass, inertial mass and gravitational mass, which it considers to be identical. But could this identity claim not be seen as a claim about the value of a parameter? Perhaps the family of rules to which Newtonian theory ought to be regarded as belonging for the purposes of simplicity determination includes rules that posit a wide range of relationships between rest mass and inertial mass, a range that would certainly increase Newtonianism's power to accommodate and so decrease its simplicity. Such are the difficulties of a family-relative notion of simplicity; I am not sure how they should be resolved.

Michael Strevens
New York University

strevens@nyu.edu

References

Friedman, M. (1974). 'Explanation and scientific understanding'. *Journal of Philosophy* 71:5–19.

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

Glymour, C. and K. T. Kelly. (2004). 'Why probability does not capture the logic of scientific justification'. In C. R. Hitchcock (ed.), *Contemporary Debates in Philosophy of Science*. Blackwell, Oxford.

Harman, G. and S. Kulkarni. (2007). *Reliable Reasoning: Induction and Statistical Learning Theory*. MIT Press, Cambridge, MA.

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

____. (2007). 'Simplicity, truth, and the unending game of science'. In S. Bold, B. Löwe, T. Räsch, and J. van Benthem (eds.), *Foundations of the Formal Sciences V: Infinite Games*. College Publications, London.

Kitcher, P. (1989). 'Explanatory unification and the causal structure of the world'. In P. Kitcher and W. C. Salmon (eds.), *Scientific Explanation*, volume 13 of *Minnesota Studies in the Philosophy of Science*, pp. 410–505. University of Minnesota Press, Minneapolis.

Putnam, H. (1963). 'Degree of confirmation and inductive logic'. In P. A. Schilpp (ed.), *The Philosophy of Rudolf Carnap*, volume 11 of *Library of the Living Philosophers*. Open Court, Chicago. Captions

Abstracta SPECIAL ISSUE III, pp. 42 – 46, 2009

COMMENTARY ON "RELIABLE REASONING"

Stephen José Hanson

This book is a wonderful redux to a time in the early half of the 20th century when statistical learning theory was just developing, and when new methods and concepts were being discovered. By the end of the 1970s learning theory had been sidetracked by the need for practical results that were consistent with the prevailing dogma of AI (artificial intelligence). Proponents of AI had posited that human thought was propositional in fact, specifically related to some unknown logical form. Consequently, much of the technology and computer science was devoted to developing the LISP or Prolog language that would somehow be more consistent the "programming language" of the mind. Early work in Machine Learning (1970s-1980s) focused on algorithms that had some functional relationship to some fundamental human task: categorization, perceptual object recognition, language understanding, reasoning, for example. But the algorithms were often too narrowly focused on these types of tasks and were termed as "brittle", in that small variations in the input conditions caused the algorithm to completely fail. One of the often repeated stories that was associated with the abrupt decline of AI during the late 1980s, was a DARPA project involving autonomous navigation of a tank in off-road environments. With the assembled VIPs of DARPA project Managers, Directors and military brass, the tank performed flawlessly through a set of standard obstacles and off-road variations, until, the sun went behind the clouds, causing the tank to immediately take a right turn into a tree continually ram the tree over and over again. After this event and others of a similar vein, much of the AI funding at major centers (MIT, CMU) was almost entirely cut. Insult was added to injury when a graduate student at CMU (Dean Pomerleau) developed a modest neural network (he dubbed ALVNN-autonomous learning vehicle Neural Network) that could be trained under various weather, road and obstacle conditions by driving the vehicle for a few hours in diverse conditions and thereafter would perform flawlessly on the same tasks as the AI programmed tank including varied lighting conditions which proved

disastrous for the more brittle AI system. ALVINN morphed over time to other autonomous vehicle systems based on combinations of Neural network technology and various systems integration methodology that recently and ironically won the 1M$ DARPA off-road autonomous vehicle competition (Thrun et al, in press).

By the 1990s the Machine Learning field was bankrupt. Conference attendance declined and many of the leaders in the field began to attend the emerging and soon to be premier Neural Network conference , Neural Information Processing (NIPS), and retooling in Neural Network methods. Through the 1990s innovation between statistical learning theory and neural network architectures and various learning algorithms (esp. Reinforcement Learning; general kernel methods, and Bayesian search methods) that all began to flourish in the "wild west" of neural computation. I noticed this when I was Program Chair of NIPS in 1992, and wrote in the introduction of that years' volume, my sense that Neural Computation was becoming a mosaic or patchwork of statistical learning theory, neural network architectures and algorithms and goals from the now defunct AI/Machine Learning field. I argued that this type of diversity at the time was to be expected for a field that was absorbing so many directions ideas and tensions from other fields. NIPS was turning into a kind of hothouse, a kind of Cambrian period for speciation, in effect an engine for change and diversity. In the 1992 meeting, we carefully struck a chord between the emerging statistical learning science and the cyberpunk neural network slash and burn –"I can make a robot dance" applications. In the first session, which I chaired, I recall having 3 talks in a row, first Michael Kearns (*Estimating Average-Case Learning Curves Using Bayesian, Statistical Physics and VC Dimension Methods* - David Haussler, Michael Kearns, Manfred Opper, and Robert Schapire) who was developing constraints and bounds on learning algorithms more generally, John Moody, who had done a theoretical analysis of why Neural Networks would learn under conditions where they were over parameterized (which of course for the first 9 years was most of the time!) and showed remarkably that they automatically, adaptively modeled the underlying data complexity and developed the notion of "effective number of parameters" (*The Effective Number of Parameters:An Analysis of Generalization and Regularization in Nonlinear Learning Systems* - John E. Moody*)* finally Vladimir Vapnik (*Principles of Risk*

Minimization for Learning Theory - Vladimir Vapnik) who revealed a new concept in risk minimization and a new method that had yet to be tried on any data-something he later called a Support Vector Machine (SVM). You could sense the audience was stirred, that they knew they just heard earth-shattering ideas and directions, but had no idea where it would go, just that field was about to explode in size and directions in what would soon be called the "learning sciences".

Vapnik's paper is of particular interest in the context of the Harman and Kulkarni book. It wasn't till the next year that Vapnik's new algorithm was coded by ATT folks (L. Boutou, I. Guyon) and tested on some difficult (handwritten zip code data) data sets and shown to perform remarkably well. The method was particularly interesting since it appeared not be as ad hoc as other methods that also were engineered to work well with these data sets (Yann Le Cun's "LeNet"), at the same time it appeared to (purposely) lose information about the underlying probability distributions—not a great idea if one wants to understand the feature space. Nonetheless, the lack of underlying explanation of why a classifier works well, never seemed to deter the NIPS crowd and it was found to be remarkably robust in generalization on very hard problems and allowed simple variations leading to a new field now beginning to be called "Large Margin Classifiers". Vapnik had in fact, by theory and development of an accessible, easily generalizable algorithm, created a new subfield in Neural Computation (not by the way in Machine Learning, as often erroneously quoted), that began to draw twice the number of attendees to the NIPS workshop than to the entire conference held the week before (not so much these days).

SVM and its Kernel or Large Margin or whatever it is called now –classifiers have become de rigueur for solving hard classification problems. Introduction of nonlinear kernels and the "kernel trick", slack variables and Cost parameters as well as ways to visualize ALPHA per voxel, made the method truly useful and frankly what one turns to when other approaches fail. I have recently used it with Brain Imaging Data (Hanson & Halchenko, 2008) and found it both amazing and annoying. We were able to classify brain images from subjects viewing either pictures of Faces or pictures of Houses using full brain (40k voxels) with 92% out-of-sample cross validation for all subjects. Although remarkable, SVM does not really allow one to examine the diagnosticity of the underlying

features without some strong assumptions which unfortunately are probably not true. Nonetheless, we forged ahead using a method called Recursive feature elimination which allowed us to titrate down to the necessary(?) number of voxels required to maintain high generalization accuracy. We were able to find 100s of contiguous voxels that are most diagnostic for these two categories, which unfortunately for some theories of cognitive/perceptual modularity, were highly overlapped. But that's another story.

The reason for imparting all of this nitty-gritty detail is to try and bring SVM a bit down to earth as well as the notions of transduction that are framed so beautifully in the book. One point in passing is that it is highly unlikely that SVM has anything to do with human cognition or categorization. As Harman and Kulkarni point out, it has no way to express a “prototype”, again due to the lack of a probability distribution in the implementation of the method (also categorical perception is not really an example of SVM…but this is also a long story—see Harnad, Hanson & Lubin, 1991). Of course, this is the very tactic that allows it to attack such high-dimensional spaces. On the other hand this is not quite the same as solving the so called “curse of dimensionality” which is often confused with the power of SVM in classifying problems, which has more to do with its ability to pick good examples per class. Frankly, SVM, as my friend Yann Le Cun once put it, is nothing more than a dumb Perceptron, that can find interesting exemplar cases but can't remember how many it found. Now this is a bit harsh, and probably missing the larger picture of this remarkable idea. Nonetheless, there is a bit of over-romanticism here concerning SVM and I can't help to think that the Harman and Kulkarni treatment of SVM is not unlike Woody Allen's view of Manhattan: “'Chapter One. He adored New York City. He idolized it all out of proportion.' Uh, no, make that: 'He-he...romanticized it all out of proportion. Now...to him...no matter what the season was, this was still a town that existed in black and white and pulsated to the great tunes of George Gershwin.'” SVM is an interesting statistical learning theory tool ...but SVM ..Gershwin.. I don't think so.

Stephen José Hanson

Rutgers University

jose@psychology.rutgers.edu

References

Harnad, S., Hanson, S.J. & Lubin, J. (1991). 'Categorical Perception and the Evolution of Supervised Learning in Neural Nets'. IN D.W. Powers & L. Reeker (eds.), *Working Papers of the AAAI Spring Symposium on Machine Learning of Natural Language and Ontology*, pp. 65-74.

Hanson, S. & Halchenko, Y.O. (2008). 'Brain Reading Using Full Brain Support Vector Machines for Object Recognition: There is no "Face" Identification Area'. *Neural Computation*, 20, 2:486-503.

Hanson, S. J. Cowan, J.D., & Giles, C.L. (1993) 'Advances in Neural Information Processing Systems 5', *NIPS Conference*, Denver, Colorado, USA, 1992], Morgan Kaufmann 1993.

Thrun, S., Montemerlo, M., Dahlkamp, H., Stavens, D., Aron, A., Diebel, J., Fong, P., Gale, J., Halpenny, M., Hoffmann, G., Lau, K., Oakley, C., Palatucci, M., Pratt, V., Stang, P., Strohband, S., Dupont, C., Jendrossek, L.-E., Koelen, C., Markey, C., Rummel, C., van Niekerk, J., Jensen, E., Alessandrini, P., Bradski, G., Davies, B., Ettinger, S., Kaehler, A., Nefian, A. and Mahoney, P. 'Stanley, the robot that won the DARPA Grand Challenge'. *Journal of Field Robotics*, In Press.

Abstracta SPECIAL ISSUE III, pp. 47 – 56, 2009

RESPONSE TO SHAFER, THAGARD, STREVENS, AND HANSON

Gilbert Harman & Sanjeev Kulkarni

Like Glenn Shafer, we are nostalgic for the time when "philosophers, mathematicians, and scientists interested in probability, induction, and scientific methodology talked with each other more than they do now", [p.10].[1] Shafer goes on to mention other relevant contemporary communities. He himself has been at the interface of many of these communities while at the same time making major contributions to them (e.g., Shafer 1976, 1996; Shafer and Vovk 2001; Vovk, Ammerman, and Shafer 2005) and this very symposium represents something of that desired discussion.

We begin with a couple of general points about issues several commentators have raised and then discuss other more particular issues.

1. General Remarks

Scope

Shafer asks skeptically how statistical learning theory might provide advice to jurors trying to decide on the guilt or innocence of someone on trial. Thagard suggests that inference to the best explanation does not always fit the conditions of statistical learning theory. He worries that restricting scientific reasoning to those conditions would yield behaviorism in psychology. And he stresses that "the goals of inductive inference include understanding as well as reliability", [p.21]. Similarly, Strevens wonders how the acceptance of Newtonian physics could be explained using statistical learning theory. And Hanson points out that learning to make predictions does not necessarily allow deeper explanations of hidden principles.

These comments indicate a serious error in our exposition. We did not intend to suggest that statistical learning theory provides a complete theory of inductive reasoning or

[1] Page references in square brackets are to the papers of this symposium.

a general "theory of empirical inquiry in science." We think it is an interesting part of such a general theory, but we do not think it is the whole thing.

Our reason for concentrating on basic statistical learning theory is that it is a relatively a priori subject, allowing mathematical proofs of interesting results. In this sense, it speaks to interest in the traditional philosophical problem of induction: what can we show a priori about induction? It would be lovely to have similar results for other cases.

Transduction

Shafer and Thagard interpret "transduction" differently from the way we use the term (based on our understanding of Vapnik's use). Shafer counts all cases of "on-line prediction" as instances of transduction. To say that the prediction is on-line is to say that data are used to make a prediction about a new case, then, once it is known what the actual value of the new case is, that information is added to the data and the augmented data are used to make a prediction about a second new case, etc.[2] (An alternative is to use data to come up with a rule that is used to make predictions about various new cases as they come up without further changes in the rule.) But, as we use the term "transduction" (following Vapnik), it does not suppose we learn the actual value of the new cases. All we learn is that certain new cases have come up to be assessed (Harman and Kulkarni, 2007, pp. 89-95).

Thagard notes that "transduction uses information about what new cases have come up in its classification of them" [p.22] but argues that "the aspect of transductive inference that it goes from cases to cases without intervening rules has a strong place in psychology as well as philosophy", [p.24]. He mentions cases of "direct inference" as discussed by Mill, Russell, and other philosophers. He also includes under this heading the sort of analogical inference discussed in Holyoak and Thagard (1995), exemplar theories of concepts, and inferences modeled by neural networks. But it is unclear that his examples are cases without "intervening rules." For example, a fixed feed-forward neural network defines a function from input features to a classification. That function specifies a rule associated with that network. Similarly, an exemplar model of a concept in the form of a

[2] On-line prediction allows for the game-theoretic approach described in Shafer and Vovk (2001).

nearest neighbor model of classification defines a function or rule from input features to a classification associated with the concept.

2. Further Response to Glenn Shafer

Shafer notes that the theory we discuss derives from "The work of Vapnik and Chervonenkis" which "is only a tiny part of the vast literature on probability and prediction that is relevant to philosophy's questions about induction and reliable reasoning", [p.11]. We think it is a part worth knowing about that also is, for example, not well known in philosophy. We agree, of course, that there are many other developments in probability and prediction that are worth knowing about, for example Shafer and Vovk's (2001) surprising demonstration that game theory can provide a basis for statistical reasoning without the assumptions needed for basic statistical learning theory.

Shafer observes that the basic statistical learning theory we discuss assumes that "observations are independently and identically distributed", [p.11]. He notes that this assumption can often be replaced by assuming that they are exchangeable. But he asks "Are there really many applications where [this] assumption is reasonable?" [p.11]. Our answer is that there are at least some real life situations in which this assumption leads to some useful applications of the basic theory, e.g. the post office problem of learning to recognize handwriten zip codes, even though Vovk, Gammerman, and Shafer (2005) say this is now considered a "toy problem", (p. 3).

Shafer also says, "By 1960, Jerzy Neyman could declare that science had become the study of stochastic processes. ...[E]very serious study in science was a study of some evolutionary chance mechanism", [p.12]. We agree that conditions change over time. If the change is slow enough, the basic theory works well enough. Shafer observes that good results in practice are possible even when the assumptions of the basic case are not met. Much of the basic theory has extensions to various types of stochastic processes, notably those with suitable mixing conditions.

We think it is pedogically useful to start with the basic theory and later consider extensions to other cases. Furthermore, as Shafer observes, the assumption can be

weakened in various ways. We also believe that many cases in which philosophers appeal to the reliability of certain epistemic methods are cases that approximate the conditions of the basic theory.

Shafer says, "If I were to fault Harman and Kulkarni on one point, it is that they do not dwell on the experimentation and reasoning that goes into choosing the kernel" (for a support vector machine), [p.14]. Certainly this is an important practical issue. The choice of representation of features and decision rules is crucial for a learning task, but this is much more an art than a science.

Finally, Shafer wonders what illumination research in machine learning might provide about the situation of a juror in a trial. We mentioned research about jurors (based on Thagard's models of their reasoning) in order to indicate (a) that a juror typically reasons by making mutual adjustments in his or her beliefs in a way that aims at reaching a kind of reflective equilibrium and (b) that this method is quite fragile in a way that makes it relatively unreliable. Can statistical learning theory suggest ways of judging the reliability of jury verdicts or ways of improving the reliability of verdicts?

We see two possible difficulties here. First, in order to apply statistical learning theory, we would need to have data as to which verdicts are correct, we need labeled examples. Perhaps we could have experienced judges provide these labels?

Second, it is legally problematic to use statistical reasoning to decide on guilt or innocence at a trial (Tribe, 1971), even though such reasoning might be useful in a different context.

In any event, we repeat that we do not think that all induction is capturable in statistical learning theory or other machine learning approaches known to us.

3. Further Response to Paul Thagard

Thagard says, "Harman and Kulkarni erroneously extrapolate from my enthusiasm for coherence-based approaches to many kinds of inference that I endorse an approach to metanormativity akin to reflective equilibrium", [p.19]. We apologize for having given this

impression. We did not mean to suggest that Thagard endorsed this approach. And we like his account of his own "decision procedure for metanormativity".

Thagard worries that "emphasis on reliability alone would restrict us to a kind of behaviorism ...But people cannot resist attributing mental states to each other, going beyond behavior to infer that people have various beliefs, desires, and emotions that cause their behavior", [p.21]. Perhaps he is thinking that the data available to the statistical learner must consist entirely in relations among observable features, so that there can be no statistical learning of how to classify someone's psychological states on the basis of observable features. But data include feature vectors plus *labels*. The labels typically represent relatively unobservable properties. For example, they might be characterizations of psychological states. The labels on the data items might be provided by "experts" (e.g., people in those states).

Thagard argues that "the goals of inductive inference include understanding as well as reliability", [p.21]. How might that idea figure in statistical learning theory? Of course, statistical learning theory can allow for values beyond getting answers that are correct rather than incorrect. It can allow that some errors are worse than others, for example, so that the goal is to minimize expected cost rather than just to minimize expected error. Furthermore, the goal of achieving understanding might be in part reflected in an inductive bias that favors some classification rules rather than others. Empirical risk minimization chooses from a limited set C a rule that minimizes error cost on the empirical data; and the choice of the limited set C may reflect the goal of coming up with a rule that provides understanding. Similarly, structural risk minimization chooses a rule that balances empirical cost against something else, using an ordering of hypotheses that may reflect the potential understanding they might provide.

However, as we noted above, we do not mean to suggest that statistical learning theory provides a general account of inductive inference.

4. Further Response to Michael Strevens

Strevens compares the statistical learning theory that we discuss in *Reliable Reasoning* to the "formal learning theory" (FLT) developed by Putnam (1963) and Kelly (1996). (See also Jain et al., 1999; Reichenbach, 1949; Kulkarni and Tse, 1994, and Schulte, 2008.) However, statistical learning theory and FLT are concerned with quite different matters. FLT is a theory of long term learning in the limit; given a potentially infinite stream of data, the task is either to arrive at a hypothesis about the stream that is eventually correct, or to approach a correct hypothesis in the limit. Statistical learning theory is a theory whose goal might be learning to how to characterize certain items. Given data and a minimal assumption about the objective probability distribution of items with various features and labels, the task is to assign labels to next items that turn up hoping to minimize (costs of) errors.[3] FLT is concerned with coming up with a hypothesis about an infinite data stream, statistical learning theory is not. Statistical learning theory is concerned with coming up with a hypothesis about the next items. FLT is not. Statistical learning theory assumes there is an unknown background probability distribution of a certain sort; FLT makes no such assumption. Statistical learning theory is appropriate for machine learning, for example to recognize zip codes from handwriting on envelopes; FLT is not.

Strevens takes both FLT and statistical learning theory as examples of "philosophical learning theory" and says, "Philosophical learning theory's pessimism lies in the insistence that the best inductive methods are those that minimize, and if possible eliminate, the possibility of failure, no matter how unlikely the failure might be", [p.28]. This is just not true of statistical learning theory, a mathematical subject that refrains from making any general claim about what inductive methods are best.

To be sure, in our discussion of statistical learning theory in *Reliable Reasoning* we describe theorems that apply in the worst case, no matter what the underlying background objective probability distribution may be. But that is not to endorse "pessimism." In addition, there is no assumption that all inductive inference fits the paradigm for statistical

[3] Statistical learning theory can be applied to other issues too, such as function estimation, although in *Reliable Reasoning* we mainly discuss learning categorization.

learning theory. As we have emphasized already, statistical learning theory is not intended as a general "theory of empirical inquiry in science."

Furthermore, even if when the basic statistical learning theory paradigm holds, one may have reason to make additional assumptions about the background objective probability distribution that go beyond or otherwise modify the assumptions of basic statistical learning theory.

Strevens wonders, "How, then, to regard whatever hypothesis is recommended by the learning theorist at any time? ...What, then, should philosophers of science take away from all this? That the idea of evidential support in science is a fallacy?" [pp.31-2]. Our answer is that a given learning method offers an account of one kind of evidential support. Suppose that the relevant learning method is empirical risk minimization (enumerative induction). Given data, this method recommends a classification rule. Given features of a new case, the classification rule recommends a particular classification of that case. So, the learning data plus the features of the new case provide evidential support for that classification.

Strevens refers to "the idea that Vapnik and Chervonenkis's theory supplies, in its notion of the VC dimension of a set of inductive rules, an interesting surrogate for simplicity in scientific reasoning" [p.32] and later says, "Harman and Kulkarni imply that VC dimension provides a good approximation for, perhaps even a good account of, simplicity in at least some parts of science", [p.35]. But following Vapnik, we prefer to think of VC dimension as providing an *alternative* to simplicity in some scientific reasoning (for reasons we explain on pp. 69-73).

Strevens takes VC dimension to have something to do with "wiggle room" and, thinking of points in the *xy* plane, says, "while most cubic polynomials will come nowhere near fitting a set of four data points, there is sure to be *some* cubic that fits them exactly, that is, some choice of parameters that delivers a cubic that gets everything exactly right", [p.33]. This is too strong. There is a possibility that two of the four data points have the same x value but different y values, which would rule out any function $f(x)=y$ that captures those points.

5. Further Response to Stephen Hanson

Our intention in writing *Reliable Reasoning* was to suggest that basic statistical learning theory provides one sort of response to the traditional philosophical problem of induction, which asks what can be shown *a priori* about induction—especially that part of basic statistical learning theory concerned with worse case results. Our focus is on this philosophical use of this aspect of statistical learning theory rather than on any particular technique per se.

As Stephen Hanson points out, many people studying computational learning have been less interested in this philosophical use of the theory than in developing practical systems that actually learn useful things. As he indicates, starting in the 1950s, attempts were made to develop learning systems by trying to simulate learning by reasoning that follows explicit principles of propositional logic. Such learning was modeled by production systems and other "artificial intelligence" approaches. Alas, as Hanson explains, this approach resulted in fragile systems that could be applied only to small toy problems. By the mid 1980s attention shifted to learning in neural networks, involving systems that were less fragile and less limited in problem size. Since then various new theoretical ideas have influenced the practice of those interested in developing systems that actually learn, ideas such as support vector machines (which we say a little about in *Reliable Reasoning*) or "boosting," (which we do not discuss).

As Hanson observes, many of the new methods may do relatively well in learning to make predictions at the cost of not allowing deeper explanations of what is going on—explanations involving hidden principles (if there are any). There is an interesting methodological issue here concerning when it is useful to try to discover underlying principles and when it is better to go ahead with methods that give good predictions without uncovering such principles. But we cannot discuss that issue here.

Hanson takes it to be unlikely that support vector machines provide insights into animal or human cognition, something we briefly discuss in *Reliable Reasoning*. We think it is premature to rule this out. He also takes us to be overly "romantic" [p.45] about support vector machines. We are not clear why he says that. We discuss support vector machines as one approach of many without any implication that it is the only or best

learning method or the one we are endorsing. Our focus is on how this and other learning techniques (including many that we do not discuss) illustrate principles of basic statistical learning theory that have something of value to contribute to discussions of the philosophical problem of induction.

Gilbert Harman & Sanjeev Kulkarni

Princeton University

harman@princeton.edu / kulkarni@princeton.edu

References

Harman, G. (1965). 'Inference to the Best Explanation,' *Philosophical Review* 74: 88-95.

_____. (1967). 'Enumerative Induction as Inference to the Best Explanation,' *Journal of Philosophy* 64: 529-533.

Harman, G., and Kulkarni, S. (2007). *Reliable Reasoning: Induction and Statistical Learning Theory*. Cambridge, MA: MIT Press.

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

Kelley, K. T., (1996). *The Logic of Reliable Inquiry*. Oxford: Oxford University Press.

Kulkarni, S. and Tse, D. N. C. (1994). 'A Paradigm for Class Identification Problems,'; *IEEE Transactions on Information Theory*, 40, pp. 696-705.

Putnam, H., (1963). 'Degree of Confirmation and Inductive Logic.' In *The Philosophy of Rudolf Carnap*, ed. A. Schillp. LaSalle, Indiana: Open Court.

Reichenbach, H., (1949). *The Theory of Probability*. Berkeley: University of California Press.

Schulte, Oliver, (2008). 'Formal Learning Theory.' *The Stanford Encyclopedia of Philosophy* (Winter 2008 Edition), Edward N. Zalta (ed.), http://plato.stanford.edu/archives/win2008/entries/learning-formal/.

Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton: Princeton University Press.

_____. (1996). *The Art of Causal Conjecture*. Cambridge, MA: MIT Press.

Shafer, G. and Vovk, V. (2001). *Probability and Finance: It's Only a Game*. New York: Wiley.

Tribe, L. (1971). 'Trial by Mathematics: Precision and Ritual in the Legal Process,' *Harvard Law Review* 84: 1329-1393.

Vovk, V., Ammerman, A., Shafer, G. (2005). *Algorithmic Learning in a Random World*. New York: Springer.