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.