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.

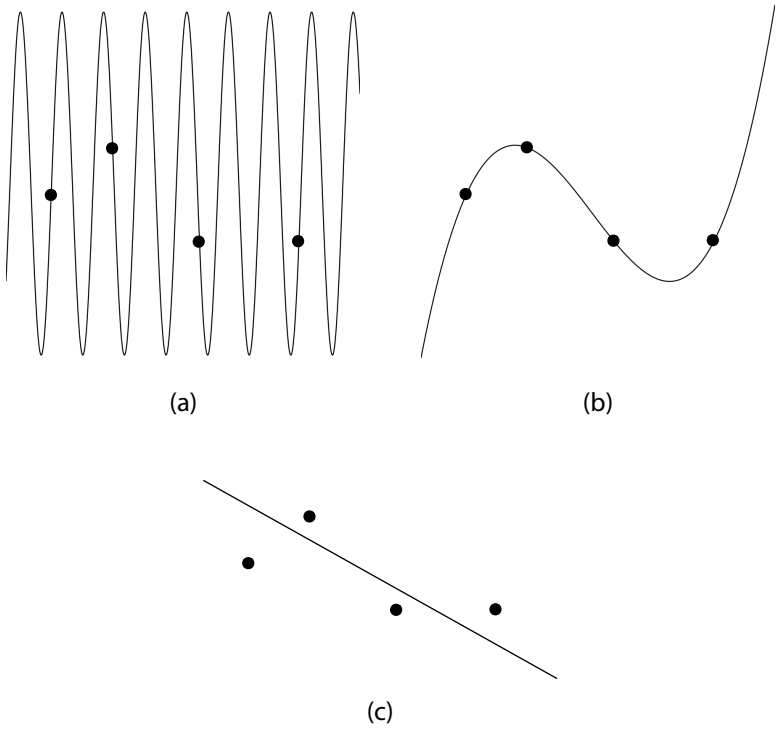

Figure 1: Curve-fitting: Any four data points can be fitted accurately by (a) some sine curve ($y = a \sin bx$) and (b) some cubic ($y = ax3 + bx2 + cx + d$), but not necessarily by (c) a line ($y = ax + b$).

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