Philosophy and Phenomenological Research
Vol. LXXII, No. 3, May 2006

The Problem of Induction[*]

GILBERT HARMAN
Princeton University

SANJEEV R. KULKARNI
Princeton University

The Problem

The problem of induction is sometimes motivated via a comparison between rules of induction and rules of deduction. Valid deductive rules are necessarily truth preserving, while inductive rules are not.

So, for example, one valid deductive rule might be this:

(D) From premises of the form "All *F* are *G*" and "*a* is *F*," the corresponding conclusion of the form "*a* is *G*" follows.

The rule (D) is illustrated in the following depressing argument:

(DA) All people are mortal.
I am a person.
So, I am mortal.

The rule here is "valid" in the sense that there is no possible way in which premises satisfying the rule can be true without the corresponding conclusion also being true.

A possible inductive rule might be this:

(I) From premises of the form "Many many *F*s are known to be *G*," "There are no known cases of Fs that are not *G*," and "*a* is *F*," the corresponding conclusion can be inferred of the form "*a* is *G*."

The rule (I) might be illustrated in the following "inductive argument."

(IA) Many many people are known to have been mortal.
There are no known cases of people who are not mortal.
I am a person.

[*] Editor's note: This paper was delivered at the 2005 Rutgers Epistemology Conference.

So, I am mortal.

The rule (I) is not valid in the way that the deductive rule (D) is valid. The "premises" of the inductive inference (IA) could be true even though its "conclusion" is not true. Even if there are no known cases up until now of people who are not mortal, it is certainly conceivable (at least to someone giving this argument) that one could live forever.

That possibility does not impugn the validity of the deductive rule (D), because if one does live forever, the first premise of the deductive argument (DA) will turn out to be false. But one's living living forever would not make any of the premises of the inductive argument (IA) false.

We might say that deduction has a kind of perfect reliability in a way that induction does not. One problem of induction then is the problem of saying in what way inductive rules might be reliable.

This issue about the reliability of induction is not the same as the issue of whether it is possible to produce a noncircular justification of induction. That other issues arises when one considers how to justify one or another inductive rule. It may seem that the only possible justification would go something like this.

Induction has been pretty reliable in the past.
So, induction will be pretty reliable in the future.

Any such justification is circular because it uses an inductive principle to justify an inductive principle. Perhaps we can justify one inductive principle in terms of another, but ultimately there will be an inductive principle for which we can supply no non-circular justification.

In any event, the issue of noncircular justification is not the problem of inductive reliability with which we will be concerned. Our problem is this. A deductive rule like (D) is perfectly reliable in the sense that, necessarily, it never leads from true premises to a false conclusion. An inductive rule like (I) is not perfectly reliable in that sense. There are instances of (I) with true premises but false conclusions. Our problem then is to explain what sort of reliability an inductive rule might have and to specify inductive rules that have that sort of reliability.

We might want to measure the reliability of a rule like (I) by the percentage of instances with true premises that have true conclusions. But there is a difficulty due to the fact that the rule may have infinitely many instances with true premises, infinitely many of which have false conclusions and infinitely many of which have true conclusions, and given infinitely many cases of each sort it is not clearly defined what the percentage is of instances with true conclusions. We might consider only inductive arguments fitting the rule that people have actually made or will make, presumably a finite number, in which case reliability might be measured by the percentage of actual inferences of this sort with true premises that also have true conclusions. But, of course, that would not be a useful measure of the reliability of inductive rules that people have not and never will use. So, we might consider the percentage of inferences of the relevant form with true premises that *would* also have true conclusions if people *were* to make inferences of that form. However, it isn't clear how to evaluate such a counter-factual criterion. A better idea is to consider the *statistical probability* that inferences of that form with true premises would also have true conclusions.

We will consider something like this last idea later, but first we need to discuss an oversimplification in the somewhat standard way in which we have stated this problem of induction.

Inference and Implication

Following tradition, we have been writing as if there were two kinds of reasoning, deductive and inductive, with two kinds of arguments, deductive and inductive. That traditional idea is wrong and correcting it complicates the issue of inductive reliability.

In the traditional view, reasoning can be modeled by a formal argument. You start by accepting certain premises, you then accept intermediate conclusions that follow from the premises or earlier intermediate conclusions in accordance with certain rules of inference. You end by accepting new conclusions that you have inferred directly or indirectly from your original premises.

So, in the traditional view, a deductive logic is a theory of reasoning. Deductive logic is concerned with deductive rules of inference like (D). We have a good deductive logic but it is said we need an inductive logic that would be concerned with inductive rules of inference like (I).

The trouble is that this traditional picture of the relation between induction and deduction conflates two quite different things, the theory of reasoning and the theory of what follows from what. Deductive logic is a theory of what follows from what, not a theory of inference or reasoning.

One problem with the traditional picture is its implication that reasoning is always a matter of inferring new things from what you start out believing. On the contrary, reasoning often involves abandoning things you start out believing. For example, you discover an inconsistency in your beliefs and you reason about which to give up. Or you start by accepting a particular datum that you later reject as an "outlier." More generally, you regularly modify previous opinions in the light of new information.

A related problem with the traditional picture is its treatment of deductive principles like (D) as rules of inference. In fact they are rules about what follows from what. Recall what (D) says:

(D) From premises of the form "all *F* are *G*" and "*a* is *F*" the corresponding conclusion of the form "*a* is *G*" follows.

(D) says that a certain conclusion follows from certain premises. It is not a rule of inference. It does not say, for example, that if you believe "All F are *G*" and also believe "*a* is *F*" you may or must infer "*a* is *G*." That putative rule of inference is not generally correct, whereas the rule about what follows from what holds necessarily and universally. The alleged rule of inference is not generally correct because, for example, you might already believe "a is not *G*" or have good reason to believe it. In that case, it is not generally true that you may or must also infer and come to believe "*a* is *G*"

Perhaps you should instead stop believing "All *F* are *G*" or "*a* is *F*." Perhaps you should put all your energy into trying to figure out the best response to this problem, which may involve getting more data. Or perhaps you should go have lunch and work out how to resolve this problem later!

From inconsistent beliefs, everything follows. But it is not the case that from inconsistent beliefs you can infer everything.

Deductive logic is a theory of what follows from what, not a theory of reasoning. It is a theory of deductive consequence. Deductive rules like (D) are absolutely universal rules, not default rules, they apply to any subject matter at all, and are not specifically principles about a certain process. Principles of reasoning are specifically principles about a particular process, namely reasoning. If there is a principle of reasoning that corresponds to (D), it holds only as a default principle, "other things being equal."

Deductive arguments have premises and conclusions. Reasoning does not in the same way have premises and conclusions. If you want to say that the "premises" of inductive reasoning are the beliefs from which you reason, it is important to note that some of those beliefs may be given up in the course of your reasoning. An argument is an abstract structure of propositions. Reasoning is a psychological process.

Sometimes in reasoning, you do construct an argument. But you do not normally construct the argument by first thinking the premises, then the intermediate steps, and finally the conclusion. You do not generally construct the argument from premises to conclusion. Often you work backwards from the desired conclusion. Or you start in the middle and work forward towards the conclusion of the argument and backward towards the premises.

Sometimes you reason to the best explanation of some data, where your explanation consists in an explanatory argument. In such a case, the *conclusion* of the explanatory argument represents the "premises" of your reasoning, the data to be explained, and the "conclusion" of your reasoning is an explanatory *premise* of your argument.

It is what philosophers call a "category mistake" to treat deduction and induction as two species of the same genus, because deduction and induction are of very different categories. Deductive arguments are abstract structures of propositions. Inductive reasoning is a process of change in view. There are deductive arguments, but it is a "category mistake" to speak of deductive reasoning except in the sense of reasoning *about* deductions. There is inductive reasoning, but it is a "category mistake" to speak of inductive arguments. There is deductive logic, but it is a "category mistake" to speak of inductive logic.

One might object that there is a perfectly standard terminology used by some logicians according to which certain deductive rules are called "rules of inference." How could we object to this terminology? Our answer is that this is like saying that there is a perfectly standard terminology used by some gamblers according to which the so-called "gambler's fallacy" is a legitimate principle about probability. "That's just how they use the term *probable*!" The gambler's fallacy is a real fallacy, not just a terminological difference. It can have terrible results. In the same way, to call deductive rules "rules of inference" is a real fallacy, not just a terminological matter. It lies behind attempts to develop relevance logics or inductive logics that are thought better at capturing ordinary reasoning than classical deductive logic does, as if deductive logic offers a partial theory of ordinary logic. It makes logic courses very diffcult for students who do not see how the deductive rules are rules of inference in any ordinary sense. It is just wrong for philosophers and logicians to continue carelessly to use this "terminology," given the disastrous effects it has had and continues to have on education and logical research.

We are not arguing that there is *no* relation between deductive logic and inductive reasoning. Our limited point here is that deductive rules are rules about what follows from what, not rules about what can be inferred from what. Maybe, as has often been suggested, it is an important principle of reasoning that, roughly speaking, one should avoid believing inconsistent things, where logic provides an account of one sort of consistency. Whether or not there is such a principle and how to make it more precise and accurate is an interesting question that is itself not settled within deductive logic, however. Similar remarks apply to the thought that principles of inductive reasoning have to do with rational or subjective degrees of belief, where consistency then includes not violating the axioms of the probability calculus. There is a mathematical theory of probability. How that theory is to be applied to reasoning is not part of the mathematics. The same point holds for decision theories that appeal to utilities as well as probabilities. These theories may offer extended accounts of consistency or "coherence" in one's belief but leave open in what way such consistency or coherence is relevant to reasoning.

Various theories of belief-revision are sometimes described as logics, not just because there is a use of the term "logic" to refer to methodology but because these theories of belief revision have certain formal aspects. As will become clear in what follows, we certainly have no objection to the attempt to provide formal or mathematical theories or models of reasoning of this sort. We very much want to develop models that are, on the one hand, psychologically plausible or implementable in a machine and are, on the other hand, such that it is possible to know something useful about their reliability.

Anyway, to repeat the point of this section: it is a mistake to describe the problem of inductive reliability by comparison with deductive reliability. Deductive rules are rules about what follows from what; they are not rules about what can be inferred from what.

Reflective Equilibrium

So, induction is a kind of reasoned change in view in which the relevant change can include subtraction as well as addition. Can anything specific be said about how people actually do inductive reasoning? And can anything specific be said about the reliability of their inductive reasoning?

One obvious point is that actual reasoning tends to be "conservative" in the sense that the number of new beliefs and methods added and old beliefs and methods given up in any given instance of reasoned change in view will be quite small in comparison with the number of beliefs and methods that stay the same. The default is not to change.

At least two things can lead us to make reasoned changes in our beliefs. First, we may want to answer a question on which we currently have no opinion; reasoning from our present beliefs can then lead us to add one or more new beliefs. Second, we may find that some of our beliefs are inconsistent with or in tension with others; reasoning from our presently conflicting beliefs can then lead us to abandon some of those beliefs.

In making changes of either sort, we try to pursue positive coherence and to avoid incoherence. That is, given an interest in adding beliefs that would answer a particular question, we favor additions that positively cohere with things we already accept because, for example, the additions are implied by things we already accept or because the addition helps to explain things we already accept. Furthermore, we try to avoid incoherence in our beliefs due to contradictions or other sorts of conflict.

Paul Thagard (1989, 2000) has developed a "constraint satisfaction" model of coherence based reasoning using artificial neural networks, a model which has proved fruitful in research in decision-making (Holyoak and Simon, 1999; Simon et al., 2001; Simon and Holyoak, 2002; Read, Snow, and Simon, 2003; Simon, 2004).

The coherence based conception of reasoning plays a role in what Nelson Goodman (1953) says about justification. He says we cannot provide a noncircular justification of induction but we can test particular conclusions by seeing how they fit with general principles we accept, and we test general principles by considering how they fit with particular conclusions we accept. If our general principles conflict with our particular judgments, we adjust principles and particular judgments until they cohere with each. The resulting principles and judgments are justified, at least for us and at least for the moment.

John Rawls (1971) refers approvingly to Goodman's discussion and says that justification consists in modifying general principles and particular judgments with the aim of arriving at what he calls a "reflective equilibrium" in which our general principles fit with our "considered judgments" about cases and our judgments about cases fit with our general principles.

The reflective equilibrium view of justification is conservative in the sense that it assumes that each of our present beliefs and methods has a kind of default justification; our continuing to accept a given belief or method is justified in the absence of some special challenge to it from our other beliefs and methods. In this view, all of our current beliefs and methods represent default "foundations" for justification, at least if the foundations are understood to be the starting points for justification.

In the reflective equilibrium view of justification, the foundations are quite *general*. In contrast, what we might call *special foundations* theories suppose that the default starting points for justification are more restricted. In the strictest special foundations theories (Descartes 1641) the foundations are limited to what is completely obvious and indubitable at the present time. Such strict foundations theories give rise to various traditional epistemological problems—the problem of justifying beliefs based on the testimony of others, the problem of justifying beliefs in other minds, the problem of justifying beliefs in the existence of objects in the external world, the problem of justifying beliefs about the future based on past evidence, and the problem of justifying reliance on memory.

In a foundations theory of justification, the extent to which our beliefs and methods are justified depends on how narrow the foundations are. Very narrow foundations imply that very little is justified and general skepticism results. Such an unwelcome result can be avoided by expanding the foundations, for example, to allow that perceptual beliefs about the environment are foundational. In such an expanded foundationalism, there is no longer the same sort of epistemological problem about the external world. A certain type of inductive reasoning might be treated as a foundational method, in which case there is no longer an epistemological problem of induction. Similar proposals have been made about our ordinary reliance on memory and testimony. For example, Burge (1993) and Foley (1994) treat reliance on testimony as a foundational method, which gets rid of the otherwise intractable epistemological problem of justifying reliance on testimony.

As foundations are widened, foundations theories tend more and more to resemble conservative general foundation theories which treat everything one accepts as foundational and thus avoid the traditional epistemological problems about justified belief.

Furthermore, the very process of widening foundation seems to be based on an implicit acceptance of the reflective equilibrium idea. The process occurs because the original idea of strict foundations conflicts with the particular nonskeptical judgments people find themselves committed to in ordinary life!

Worries about Reflective Equilibrium

Suppose that certain inductive methods survive as we adjust our views and methods in such a way as to attain reflective equilibrium. Why should we think that this makes those methods particularly reliable?

Goodman and Rawls say that the sort of adjustment of general principle to particular judgment is exactly how we in fact go about testing and justifying our views. But why should we assume that our ordinary methods of justification show anything about reliability? Stich and Nisbett (1980) observe in discussing this exact issue that there is considerable evidence that our ordinary reasoning practices are affected by "heuristics and biases" (Tversky and Kahneman, 1974), which can and often do produce clearly unreliable results.

To be sure, the fact that we can tell that these results are unreliable might indicate only that people are ordinarily not in reflective equilibrium, but (as Stich and Nisbett argue) various errors such as the "gambler's fallacy" might well survive ordinary reflective equilibrium. Stich and Nisbett argue that in determining what methods it is reasonable to use, we cannot rely on ordinary opinion even if it is in reflective equilibrium. They say we need instead to take expert opinion into account. But how do we determine who the experts are? And why should we trust them anyway?

Another and possibly more serious worry about ordinary reflective equilibrium is that it appears to exhibit an unwelcome fragility that undermines its claim to reliability.

We have already mentioned that Thagard (1989, 2000) develops models of the method of reflective equilibrium using a certain sort of connectionist system of constraint satisfaction and his models exhibit this worrisome fragility. These models involve networks of nodes representing particular propositions. A node receives positive excitation to the extent that it is believed and negative excitation to the extent that it is disbelieved. There are two sorts of links among nodes, excitatory and inhibitory. Excitatory links connect nodes with others that they explain or imply or stand in some sort of evidential relation to, so that as one of the nodes becomes more excited, the node's excitation

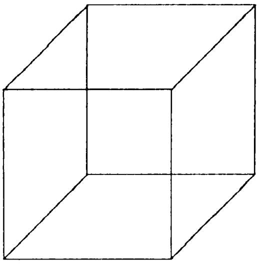

Figure 1: Necker Cube

increases the excitation of the other nodes. And as one such node becomes less excited or receives negative excitation, that decreases the excitation of the other nodes.

Inhibitory links connect nodes that conflict with each other so that as one such node receives more excitation, the others receive less and vice versa. Excitation, positive and negative, cycles round and round the network until it eventually settles into a relatively steady state. Nodes in the final state that have a positive excitation above a certain threshold represent beliefs and nodes in the final state that have a negative excitation beyond a certain threshold represent things that are disbelieved. Nodes in the final state with intermediate excitation values represent things that are neither believed nor disbelieved. The resulting state of the network represents a system of beliefs in reflective equilibrium.

It has often been noted that a connectionist network provides an appropriate model of certain sorts of Gestalt perception (Feldman, 1981). Consider a Necker cube (Figure 1).

A given vertex might be perceived as part of a near surface or as part of a farther back surface. There are excitatory links among the four vertices of each surface and inhibitory links between vertices of the different surfaces. The degree of excitation of a vertex represents how near it is. As excitation on a given vertex increases, that increases the excitation on the three other vertices of that face and drives down the excitation of the vertices on the other face. The result is that one tends to see the figure with one or the other face in front and the other in back. One tends not to see the figure as some sort of mixture or as indeterminate as to which face is in front.

Thagard's constraint satisfaction connectionist network has been used to model the reasoning of jurors trying to assess the guilt of someone in a trial. The model makes certain predictions. For example, a juror might begin with a view about the reliability of a certain sort of eye-witness identification, a view about whether posting a message on a computer bulletin board is more like writing something in a newspaper or more like saying something in a telephone conversation, and so forth. Suppose the case being decided depends in part on an assessment of such matters. Then Thagard's model predicts that a juror's general confidence in this type of eye-witness identification should increase if the juror judges that in this case the testimony was correct and should decrease if the juror judges that in this case the testimony was not correct, the model predicts a similar effect on the juror's judgment about what posting on a computer network is more similar to, and so forth. The model also predicts that, because of these effect, the juror's resulting reflective equilibrium will lead to the juror's being quite confident in the verdict he or she reaches.

Experiments involving simulated trials have confirmed this prediction of Thagard's model (Simon 2004). In these experiments, subjects are first asked their opinions about certain principles of evidence about certain sorts of eye-witness identifications, resemblances, etc. Then they are given material about difficult cases involving such considerations to think about. Finally, the subjects' final verdicts and their confidence in their verdicts and in the various principles of evidence are recorded.

One result is that, as predicted, although subjects may divide in their judgment of guilt at the end, with some saying the defendant is guilty and others denying this, subjects claim to be quite confident in their judgments and in the considerations that support them. Furthermore, also as predicted, subjects' judgments about the value of that sort of eye-witness identification, about the resemblance of posting on a computer bulletin board to other things, and so forth, also change at least while they are still thinking about the particular case in question.

The model implies that judgements in hard cases are fragile and unreliable in the following sense. When there is conflicting evidence, there is considerable tension among relevant considerations, just as there is a certain sort of tension among the nodes representing vertices in the Necker cube problem. If some nodes acquire even slightly increased or decreased excitation, the relevant inhibitory and excitatory connections can lead to changes in the excitation of other nodes in a kind of chain reaction or snowballing of considerations leading to a clear verdict, one way or the other, depending on the initial slight push, just as happens in one's perception of a Necker cube.

After the Gestalt shift has occurred, however, the case seems quite clear to the juror because of ways the juror's confidence has shifted in response to the positive and negative connections between nodes.

One upshot of this is that the slight errors in a trial that look like "harmless errors" can have a profound effect that cannot be corrected later by telling the juror to ignore something. By then the ignored evidence may have affected the excitation of various other items in such a way that the damage cannot be undone. Similarly, the fact that the prosecution goes first may make a difference by affecting how later material is evaluated.

This fragility of reflective equilibrium casts doubt about using the method of reflective equilibrium to arrive at reliable opinions.

There is some recognition of this problem in the literature concerning Rawls' appeal to reflective equilibrium in his account of the justice of basic institutions of society. It has been said that the problem might be met by trying to find a "wide" rather than a "narrow" reflective equilibrium, where that involves not only seeing how one's current views fit together but also considering various other views and the arguments that might be given for them and one must be careful to try to avoid the sorts of effects that arise from the order in which one gets evidence or thinks about an issue (Daniels, 1979). One needs to consider how things would have appeared to one if one had gotten evidence and thought about issues in a different order, for example.

Experimenters have shown that if subjects are instructed to try for this sort of wide reflective equilibrium, they are less subject to the sorts of effects that occur when they are not (Simon, 2004).

Does this mean that inductive methods acceptable to wide reflective equilibrium are reliable? Maybe, but why should we think so? Once we come to doubt the reliability of methods acceptable to narrow reflective equilibrium, why should we believe in the reliability of inductive methods accepted in wide reflective equilibrium? At this point, it does not seem adequate to be content to say that this is just how we justify things and leave it at that.

Reliability

It seems we need to find another way to assess the reliability of inductive methods. Thagard (1988) discusses this issue at length and suggests various additions to the method of wide reflective equilibrium, including paying attention to methods that figure in what we take to be the best science and to the goals of reasoning. But we would now like to point out that the reliability of various inductive methods has been studied extensively by theorists interested in machine-learning, more specifically under the heading of "statistical learning theory."

To take a problem that has been studied extensively, suppose we want a method for reaching conclusions about the next F on the basis of observing prior *F*s. We want the results of the method to be correct, or correct most of the time. We are interested in finding a usable method that does as well as possible.

Suppose that a usable method uses data to select a rule from a certain set S of rules for classifying new cases on the basis of their observed characteristics. Ideally, we want the method to select the best rule from S, the rule that makes the least error on new cases, the rule that minimizes expected error on new cases.

In other words, suppose that all the rules in S have a certain "expected error" on new cases. We want a method for finding the rule with the least expected error, given enough data.

But what does it mean to talk about the "expected error" of a rule from S. We might identify the expected error with the (unknown) frequency of actual errors we will make using the rule. But as we mentioned earlier, we will want to consider the expected error for rules we don't use, where there is no frequency of actual errors. So perhaps we need to consider the frequency of errors we would make if we used the rule, which is perhaps to say that the expected error of a rule is the (unknown) probability of error using that rule.

But where does that probability come from? We are concerned with the actual reliability of one or another rule, which presumably cannot be identified with our degree of belief in the rule or even with any sort of epistemic probability. It has to be some sort of more or less objective statistical probability. Let us explain.

Here is an example illustrating a distinction between the sort of objective statistical probability relevant to actual reliability and subjective or evidential probability. Suppose we show you a pair of dice. Each of a die's six sides has a different number of dots, from 1 to 6. You are to throw the dice and record the total number of spots on their uppermost sides. What is the probability of getting a total of seven?

Given your evidence, let us suppose, each side of a die is equally likely to come up, so the probability of getting a six (for example) is 1/6 and the probability of both coming up six for a total of 12 is 1/36 . The evidential probability of the total of the sides being exactly 7 is the probability that the first die comes up 1 and the second comes up 6 plus the probability that the first comes up 2 and the second 5 plus . . . , or 6/36 = 1/6.

Now suppose that, unknown to either of us, the first die is weighted so that the statistical probability of getting a 4 is 2/3 and the other sides are equally likely. Also, the second die is weighted so that the probability of a 3 is 2/3 with the other sides equally likely. Then the (unknown to us) statistical probability of getting a total of 7 would be the probability of getting a 4 on the first die and a 3 on the second, namely $2/3 \times 2/3 = 4/9$, plus the probability of getting 1 on the first die and 6 on the second, namely $1/15 \times 1/15$, plus the similar probabilities of the other four combinations, or $4/9 + (5 \times 1/225) \approx 7/15$.

We suggest that actual reliability is determined by the unknown objective statistical probability rather than any sort of evidential probability. To think about reliability in this way we have to suppose that there is a certain sort of background statistical probability distribution.[1]

Earlier we said we were interested in finding an inductive method for using data to select a rule from a certain set S of rules for classifying new cases on the basis of their observed characteristics. The rules in S will be rules for estimating the classification of an item given observed characteristics. We want to find a rule from S whose expected error as measured by that background probability distribution is as low as possible.

Any conclusion about inductive reliability of the sort with which we are concerned presupposes such a background probability distribution. To seek a method that is reliable in this way is to seek a method that is reliable in relation to that probability distribution. Without the assumption of such an unknown background statistical probability distribution, it does not make sense to talk about this sort of reliability.[2]

The next question is this. How can we use data to choose a good rule from S? One obvious idea is to select a rule from S with the least error on the data. Then we use that rule in order to classify new data. This is basically the method of enumerative induction. Our question then is, "How good is this version of enumerative induction for choosing a rule from S?"

Clearly, it depends on what rules are in the set S from which a rule is to be chosen. If all possible rules are in that set, then there will be many rules that have the least error on the data but which give different advice about new cases. So, we won't be able to choose a good rule for classifying the new cases.

More generally, any inductive method must have some sort of inductive bias. It must prefer some rules over others. It must be biased in favor of some rules and against others. If the method is the sort of enumerative induction that selects a rule from S with the least error on the data, there has to be a restriction on what rules are in S. Otherwise, we will never be able to use data in that particular way to select rules for classifying new cases.

Notice furthermore that restricting the rules in S will sometimes allow enumerative induction to select a rule that is not completely in accord with

[1] It makes sense to speak of statistical probability only in relation to a level of analysis of a system as a certain sort of "chance set-up," to use Hacking's (1965) useful terminology. It may be that a process involving a roulette wheel can be described as a chance set-up at one level of analysis, as a deterministic process at a deeper level, and as a chance set-up again, at an even deeper level. Our present point is that the relevant sort of reliability has application only with reference to a level of analysis of a situation as a chance set-up in which the relevant statistical probabilities make sense. There are important issues about the interpretation of this sort of probability that we cannot discuss here, except to say that this notion plays an important role in various contemporary subjects studied in engineering and computer science, including statistical learning theory.

[2] The term "reliable" might also be used of methods that give desired results "in the limit." Such results need not be reliable in the short term in the sense that concerns us.

the data. Accepting such a rule is not to accept that the data are completely correct. So, enumerative induction can involve giving up something previously accepted.

Of course, restricting the rules in S runs the risk of not including the best of all possible rules, the rule with the least expected error on new cases. That is a problem with this sort of enumerative induction, because there is no way to use such enumerative induction without restricting the rules in S.

There are other possible inductive methods for choosing rules—methods that do not just choose the rule with the least error on the data. One such method balances data-coverage against something else, such as the simplicity of a given rule. In that case, the idea is to choose a rule that has the best combination of data-coverage and simplicity as measured in one or another way. We will say a little about that idea in a moment, but now let us concentrate on what is needed for the sort of enumerative induction that simply chooses the rule in S with the least error on the data. The present point is that such simple enumerative induction cannot include all possible rules in S.

So now consider the question of how the rules in S might be restricted if enumerative induction in this sense is to be guaranteed to work, given enough evidence, no matter what the background statistical probability distribution.

The answer to this question is one of the great discoveries of statistical learning theory—the discovery of the importance of the *Vapnik-Chervonenkis* dimension, or VC-dimension, of a set of rules. The VC-dimension is a measure of the "richness" of the set of rules and it is inversely related to the degree of falsifiability of the set.[3] Roughly speaking, Vapnik and Chervonenkis' (1968) fundamental result is that enumerative induction in the relevant sense can be shown to work, given enough data, no matter what the background statistical probability distribution, iff the set S has finite VC-dimension.

As we mentioned, enumerative induction in this sense is not the only possible inductive method. But it is a method that applies to many examples of machine learning, including perceptron learning, feed-forward neural net learning, and support vector machines.

The other method we mentioned, in which data-coverage is balanced against something else, allows for choosing among a set of rules with infinite VC-dimension. Here it can be shown that the right thing to measure against data-coverage is VC-dimension rather than simplicity conceived in some more usual way. We will not try to explain that result here.

[3] More precisely, the VC-dimension of a set of rules S is the maximum number of data points that can be arranged so that S "shatters" those points. S shatters N data points iff for every one of the 2^N ways of assigning values to each of those points there is a rule in S that is in accord with that assignment. Vapnik connects the role of VC-dimension with Popper's (1934) discussion of the importance of falsifiability in science.

Vapnik (1979) describes a method of inference, which he has more recently (1998, 2000, p. 293) called "transduction," a method that infers directly from data to the classification of new cases as they come up. Under certain conditions, transduction gives considerably better results than those obtained from methods that use data to infer a rule that is then used to classify new cases (Joachims 1999, Vapnik 2000, Weston et al. 2003, Goutte et al. 2004).

More generally, the problem of induction as we have described it—the problem of finding reliable inductive methods—can be fruitfully investigated, and is being fruitfully investigated in statistical learning theory (Vapnik, 1998; Kulkarni et al., 1998, Hastie et al., 2001).[4]

Conclusion

Let us sum up. The problem of induction as we have been understanding it is the problem of assessing the *reliability* of inductive inference. The problem is sometimes motivated by comparing induction with deduction, a comparison that rests on a confusion about the relation between inference and logic. Some suggest that the only real problem is to try to specify how we do inductive reasoning. In this view issues about reliability are to be answered by adjusting one's methods and beliefs so that they fit together in a reflective equilibrium. While there is evidence that people do reason by adjusting their opinions in the way suggested, there is also considerable evidence that the results are fragile and unreliable, and it is hard to be in reflective equilibrium if you cannot believe your methods of reasoning are reliable. Given that reasoning often involves giving up things previously believed, it may seem unclear how even to specify the desired type of reliability. However, it does turn out to be possible to specify methods for doing one sort of enumerative induction and to address questions about their reliability that have answers in statistical learning theory, a theory that has results about other possible inductive methods as well.[5]

[4] Our recognition of this connection between one form of the philosophical problem of induction and the subject of statistical learning theory led us to plan and teach an introductory level course at Princeton in "Learning Theory and Epistemology," Electrical Engineering 218/Philosophy 218.

[5] We are indebted to discussions with Vladimir Vapnik and to Daniel Osherson and James Pryor for comments on an earlier version.

Bibliography

Burge, T., (1993). "Content Preservation," *Philosophical Review*.

Daniels, N., (1979). "Wide Reflective Equilibrium and Theory Acceptance in Ethics." *Journal of Philosophy* 76: 256-82.

Descartes, R., (1641). *Meditationes de Prima Philosophia*. Paris.

Feldman, J. A., (1981). "A Connectionist Model of Visual Memory." In G. E. Hinton and J. A. Anderson (Eds.). *Parallel Models of Associative Memory*, (Hillsdale, NJ.: Erlbaum), 49-81.

Foley, R., (1994). "Egoism in Epistemology." In F. Schmitt, ed., *Socializing Epistemology*. Lanham: Rowman and Littlefield.

Goodman, N., (1953). *Fact, Fiction, and Forecast*, Cambridge, MA: Harvard University Press.

Goutte, C., Cancedda, N., Gaussier, E., Dèjean, H. (2004) "Generative vs Discriminative Approaches to Entity Extraction from Label Deficient Data." *JADT 2004, Les Journèes internationales d'Analyse statistique des Donnèes Textuelles*, Louvain-la-Neuve, Belgium, 10-12 mars.

Hacking, I., (1965). *The Logic of Statistical Inference*. Cambridge: Cambridge University Press.

Hastie, T., Tibshirani, R., and Friedman, J., (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.

Holyoak, K. J. and Simon, D., (1999). "Bidirectional Reasoning in Decision Making by Constraint Satisfaction," *Journal of Experimental Psychology: General*, 128: 3-31.

Joachims, T. (1999) "Transductive Inference for Text Classification Using Support Vector Machines." In I. Bratko and S. Dzeroski, editors, Proceedings of the 16th International Conference on Machine Learning: 200-9. San Francisco: Morgan Kaufmann.

Keynes, J. M., (1924) *A Tract on Monetary Reform*. New York: Harcourt, Brace.

Kulkarni, S. R., Lugosi, G., and Vendatesh, L. S., (1998). "Learning Pattern Classification: A Survey," *IEEE Transactions on Information Theory* 44: 2178- 2206.

Popper, K., (1934). *Logik der Forschung*. Vienna: Springer. Translated as *The Logic of Scientific Discovery* (London: Routledge, 2002).

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

Read, S. J., Snow, C. J., and Simon, D., (2003). "Constraint Satisfaction Processes in Social Reasoning." *Proceedings of the 25th Annual Conference of the Cognitive Science Society*: 964-969.

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

Simon, D., (2004). "A Third View of the Black Box," *University of Chicago Law Review*, 71, 511-586.

Simon, D. and Holyoak, K. J., (2002). "Structural Dynamics of Cognition: From Consistency Theories to Constraint Satisfaction," *Personality and Social Psychology Review*, 6: 283-294.

Simon, D., Pham, L. B., Le, Q A., and Holyoak, K. J., (2001). "The Emergence of Coherence over the Course of Decision Making," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27: 1250-1260.

Stich, S. and Nisbett, R., (1980). "Justification and the Psychology of Human Reasoning," *Philosophy of Science* 47: 188-202.

Thagard, P., (1988). *Computational Philosophy of Science*. Cambridge, MA: MIT Press. Thagard, P., (1989). "Explanatory Coherence." *Brain and Behavioral Sciences*, 12: 435-467.

Thagard, P., (2000). *Coherence in Thought and Action*. Cambridge, MA: MIT Press.

Tversky, A. and Kahneman, D., (1974). "Judgment under Uncertainty: Heuristics and Biases." *Science* 185: 1124-1131.

Vapnik, V., (1979). *Estimation of Dependencies Based on Empirical Data* (in Russian), Moskow: Nauka. English translation (1982) New York: Springer.

Vapnik, V., (1998). *Statistical Learning Theory*. New York: Wiley.

Vapnik, V., (2000) *The Nature of Statistical Learning Theory*, second edition. New York, Springer. Vapnik, V., and Chervonenkis, A. Ja., (1968). "On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities," *Doklady Akademii Nauk USSR* 181.

Weston, J., Pèrez-Cruz, F., Bousquet, O., Chapelle, O., Elisseeff, A., and Schölkopf, B. (2003) "KDD Cup 2001 Data Analysis: Prediction of Molecular Bioactivity for Drug Design-Binding to Thrombin." *Bioinformatics*.