Error and Inference

Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science

Edited by

DEBORAH G. MAYO

Virginia Tech

ARIS SPANOS

Virginia Tech

CAMBRIDGE
UNIVERSITY PRESS
2010

FIVE

Induction and Severe Testing

Mill's Sins or Mayo's Errors?

Peter Achinstein[1]

Although I have offered some criticisms of her views on evidence and testing (Achinstein, 2001, pp. 132–40), I very much admire Deborah Mayo's book (1996) and her other work on evidence. As she herself notes in the course of showing how misguided my criticism is, we actually agree on two important points. We agree that whether *e*, if true, is evidence that *h*, in the most important sense of "evidence," is an objective fact, not a subjective one of the sort many Bayesians have in mind. And we agree that it is an empirical fact, not an a priori one of the sort Carnap has in mind. Here I will take a broader and more historical approach than I have done previously and raise some general questions about her philosophy of evidence, while looking at a few simple examples in terms of which to raise those questions. It is my hope that, in addition to being of some historical interest, this chapter will help clarify differences between us.

1 Mill under Siege

One of Mayo's heroes is Charles Peirce. Chapter 12 of Mayo's major work, which we are honoring, is called "Error Statistics and Peircean Error Correction." She has some very convincing quotes from Peirce suggesting that he was a model error-statistical philosopher. Now I would not have the Popperian boldness to say that Mayo is mistaken about Peirce; that is not my aim here. Instead I want to look at the view of induction of a philosopher whom Peirce excoriates, Mayo doesn't much like either, and philosophers of science from Whewell to the present day have rejected. The philosopher

[1] I am very grateful to Linda S. Brown for helping me to philosophically reflect properly, and to Deborah Mayo for encouraging me to philosophically reflect properly on her views.

is John Stuart Mill. I say that Peirce excoriated Mill. Here is a quote from Peirce:

> John Stuart Mill endeavored to explain the reasonings of science by the nominalistic metaphysics of his father. The superficial perspecuity of that kind of metaphysics rendered his logic extremely popular with those who think, but do not think profoundly; who know something of science, but more from the outside than the inside, and who for one reason or another delight in the simplest theories even if they fail to cover the facts. (Peirce, 1931–1935, 1.70)

What are Mill's sins – besides being an outsider to science, delighting in oversimplifications, and having the father he did? Well, there are many. In what follows I will focus on the sins Peirce notes that are either mentioned by Mayo or that are or seem inimical to her error-statistical philosophy. I do so because I want to give Mill a better run for his money – a fairer chance than Peirce gave him. Also, doing so may help to bring out strengths and weaknesses in error-statistical ideas and in Mill.

Here are four sins:

1. First and foremost, Mill's characterization of induction completely omits the idea of severe testing, which is central for Peirce and Mayo. For example, such inferences completely ignore the conditions under which the observed sample was taken, which is very important in severe testing.
2. Second, Mill's inductions conform to what later became known as the "straight rule." They are inductions by simple enumeration that license inferences of the form "All observed A's are B's; therefore, all A's are B's." Moreover, they are "puerile" inferences from simple, familiar properties directly observed in a sample to the existence of those properties in the general population. Such inferences cannot get us very far in science. Peirce, following in the footsteps of Whewell earlier in the nineteenth century, has considerable disdain for Mill's inductive account of Kepler's inference to the elliptical orbit of Mars (Peirce, 1931–1935, 7.419).
3. Third, Mill supposes that an induction is an argument in which the conclusion is inferred to be true and assigned a high probability. But according to Peirce (and Mayo), an induction or statistical inference "does not assign any probability to the inductive or hypothetic conclusion" (Peirce, 1931–1935, 2.748).
4. Finally, Mill assumes that all inductions presuppose a principle of the uniformity of nature, a presupposition that, besides being vague, is not warranted or needed.

In an attempt to give at least a partial defense of Mill, I will also invoke Isaac Newton, who proposed a view of induction that is similar in important respects to Mill's view. Newton, at least, cannot be accused by Peirce of not knowing real science. And although Mill offers an account of induction that is philosophically superior to Newton's, Newton shows more clearly than Mill how induction works in scientific practice. He explicitly demonstrates how his four Rules for the Study of Natural Philosophy, including rules 3 and 4 involving induction, were used by him in defending his law of gravity.

2 A Mill–Newton View of Induction: Sins 1 and 2

Mill offers the following classic definition of induction:

> Induction, then, is that operation of the mind by which we infer that what we know to be true in a particular case or cases, will be true in all cases which resemble the former in certain assignable respects. In other words, Induction is the process by which we conclude that what is true of certain individuals of a class is true of the whole class, or that what is true at certain times will be true in similar circumstances at all times. (Mill, 1888, p. 188)

Mill declared inductive reasoning to be a necessary first stage in a process he called the "deductive method." He regarded the latter as the method for obtaining knowledge in the sciences dealing with complex phenomena. It establishes the laws necessary to analyze and explain such phenomena.[2]

Unlike Mill, Newton never explicitly defines "induction" in his works. But, like Mill, he considers it to be a necessary component of scientific reasoning to general propositions: "In . . . experimental philosophy, propositions are deduced from the phenomena and are made general by induction. The impenetrability, mobility, and impetus of bodies, and the laws of motion and the law of gravity have been found by this method."[3] The closest Newton comes to a definition of induction is his rule 3 at the beginning of Book III of the *Principia*:

> Rule 3: Those qualities of bodies that cannot be intended or remitted and that belong to all bodies on which experiments can be made should be taken as qualities of all bodies universally.

Some commentators (including a former colleague of mine, Mandelbaum [1964]) take this rule to be a rule of "transdiction" rather than "induction."

[2] The second step, "ratiocination," involves deductive calculation. The third step, "verification," involves the comparison of the results of ratiocination with the "results of direct observation" (Mill, 1888, p. 303).

[3] Newton (1999, Book 3, *General Scholium*, p. 943).

By this is meant a rule for inferring what is true about all unobservables from what is true about all observables, whether or not they have been observed. In his discussion, however, Newton himself gives some examples involving inductions (from what has been observed to what may or may not be observable) and others involving transdictions (from all observables). And in his rule 4, we are told to regard propositions inferred by induction from phenomena as true, or very nearly true, until other phenomena are discovered that require us to modify our conclusion.

Let us focus first on Mill. It is quite correct to note in Sin 1 that Mill's characterization of induction, given earlier, completely omits the idea of severe testing. But Mill's response to this charge of sin will surely be this: "We must distinguish my *definition* of induction, or an inductive generalization, from the question of when an induction is justified." Recall one of Mill's definitions of an inductive inference as one in which "we conclude that what is true of certain individuals of a class is true of the whole class." But he does not say, and indeed he explicitly denies, that any induction so defined is justified. Indeed, on the next page following this definition he writes that whether inductive instances are "sufficient evidence to prove a general law" depends on "the number and nature of the instances." He goes on to explicitly reject the straight rule of enumerative induction as being universally valid. Some inductions with this form are valid; some are not.

Later in the book, in a section on the "evidence of the law of universal causation," he writes:

> When a fact has been observed a certain number of times to be true, and is not in any instance known to be false; if we at once affirm that fact as an universal truth or law of nature, without either testing it by any of the four methods of induction [Mill's famous methods for establishing causal laws], or deducing it from other known laws, we shall in general err grossly. (Mill, 1888, p. 373)

In an even later section called "Fallacies of Generalization," Mill again emphasizes that inductive generalizations are not always valid but are subject to different sorts of "fallacies" or mistakes in reasoning, including failure to look for negative instances and making the generalization broader than the evidence allows. My point is simply that Mill defines inductive generalization in such a way that there are good and bad inductions.

Moreover, for Mill – and this point is quite relevant for, and similar to, Mayo's idea – whether a particular inductive inference from "all observed As are Bs" to "all As are Bs" is a good one is, by and large, an *empirical* issue, not an a priori one. Mill notes that the number of observed instances of

a generalization that are needed to infer its truth depends on the kinds of instances and their properties. He writes (p. 205) that we may need only one observed instance of a chemical fact about a substance to validly generalize to all instances of that substance, whereas many observed instances of black crows are required to generalize about all crows. Presumably this is due to the empirical fact that instances of chemical properties of substances tend to be uniform, whereas bird coloration, even in the same species, tends not to be. Finally, in making an inductive generalization to a law of nature, Mill requires that we vary the instances and circumstances in which they are obtained in a manner described by his "four methods of experimental inquiry." Whether the instances and circumstances have been varied, and to what extent, is an empirical question and is not determined a priori by the fact that all the observed members of the class have some property.

In short, although Mill *defines* induction in a way that satisfies the formal idea of induction by simple enumeration, he explicitly denies that any induction that has this form is thereby valid. Whether it is valid depends on nonformal empirical facts regarding the sample, the sampling, and the properties being generalized. I would take Mill to be espousing at least some version of the idea of "severe testing."

Although Newton does not define "induction," the examples of the inductive generalizations he offers in the *Principia* conform to Mill's definition. And I think it is reasonable to say that Newton, like Mill, does not claim that any induction from "all observed As are Bs" to "all As are Bs" is valid. For one thing, he never makes this statement or anything like it. For another, his inductive rule 3 has restrictions, namely, to inductions about bodies and their qualities that "cannot be intended and remitted." In one interpretation offered by Newton commentators, this rule cannot be used to make inductive generalizations on qualities of things that are not bodies (e.g., on qualities of forces or waves) or on all qualities of things that are bodies (e.g., colors or temperatures of bodies). Moreover, Newton seems to defend his third rule by appeal to a principle of uniformity of nature, much like the one Mill invokes. In his discussion of rule 3, Newton writes, "We [should not] depart from the analogy of nature, since nature is always simple and ever consonant with itself." Later I will discuss Mill's more elaborated view of such a principle. Suffice it to say now that, if Newton takes such a principle to be empirical rather than a priori, then whether some particular inductive generalization is warranted can depend empirically on whether certain uniformities exist in nature regarding the bodies and sorts of properties in question that help to warrant the inference.

However, one of the most important reasons for citing Newton is to combat the idea – formulated as part of Sin 2 – that inductive generalizations of the sort advocated by Newton and Mill are puerile ones like those involving the color of crows; that they cite only familiar properties whose presence can be established simply by opening one's eyes and looking; and that they are ones that someone like Mill, knowing little science, can readily understand and appreciate.

From astronomical observations regarding the various planets and their satellites, Newton establishes (what he calls) six "Phenomena." For example, Phenomenon 1 states that the satellites of Jupiter by radii drawn to the center of Jupiter describe areas proportional to the times, and their periods are as 3/2 power of the distances from the center of Jupiter. From the six phenomena together with principles established in Book 1, Newton derives a series of theorems or "propositions" concerning the forces by which the satellites of the planets and the planets themselves are continually drawn away from rectilinear motion and maintained in their orbits. These forces are directed to a center (the planet or the sun), and they vary inversely as the square of the distance from that center. Explicitly invoking his rule 1, which allows him to infer that one force exists here, not many, and his inductive rule 3, Newton infers that this force governs all bodies in the universe.

If someone claims that such an induction is puerile, or that the property being generalized – namely, being subject to an inverse square force – is something familiar and directly observable in the manner of a black crow, I would say what the lawyers do: *res ipse loquitur*.

Nor does Mill restrict inductions to generalizations involving familiar properties whose presence is ascertainable simply by opening one's eyes and looking. His definition of induction as "the process by which we conclude that what is true of certain individuals of a class is true of the whole class" has no such requirement. Moreover, he cites examples of inductions in astronomy where the individual facts from which the inductions are made are not establishable in such a simple manner – for example, facts about the magnitude of particular planets in the solar system, their mutual distances, and the shape of the Earth and its rotation (Mill, 1888, p. 187).

3 Assigning Probability to an Inductive Conclusion: Sin 3

Let me turn to what Peirce, and I think Mayo, may regard as the worst sin of all – assigning a probability to the inductive conclusion or hypothesis. Peirce claims that "in the case of an analytic inference we know the probability of our conclusion (if the premises are true), but in the case of synthetic

inferences we only know the degree of trustworthiness of our proceeding" (Peirce, 1931–1935, 2.693) Mayo (1996, p. 417) cites this passage from Peirce very approvingly. An example of an analytic inference for Peirce is a mathematical demonstration. Inductions are among synthetic inferences. Another quote Mayo gives from Peirce involves his claim that we can assign probability to a conclusion regarding a certain "arrangement of nature" only if

> universes were as plenty as blackberries, if we could put a quantity of them in a bag, shake them well up, draw out a sample, and examine them to see what proportion of them had one arrangement and what proportion another. (Peirce, 1931–1935, 2.684)

Because this idea makes little sense, what Peirce claims we really want to know in making a synthetic inductive inference is this: "[g]iven a synthetic conclusion, . . . what proportion of all synthetic inferences relating to it will be true within a given degree of approximation" (Peirce, 1931–1935, 2.686).

Mayo cites this passage, and says that "in more modern terminology, what we want to know are the error probabilities associated with particular methods of reaching conclusions" (Mayo, 1996, p. 416). The sin, then, is assigning probability to an inductive conclusion and not recognizing that what we really want is not this but error probabilities. So, first, do Newton and/or Mill commit this sin? Second, is this really a sin?

Neither in their abstract formulations of inductive generalizations (Newton's rule 3; Mill's definition of "induction") nor in their examples of particular inductions to general conclusions of the form "all As are Bs" does the term "probability" occur. Both write that from certain specific facts we can conclude general ones – not that we can conclude general propositions with probability, or that general propositions have a probability, or that they have a probability conditional on the specific facts. From the inductive premises we simply conclude that the generalization is true, or as Newton allows in rule 4, "very nearly true," by which he appears to mean not "probably true" but "approximately true" (as he does when he takes the orbits of the satellites of Jupiter to be circles rather than ellipses). Nor, to make an induction to the truth or approximate truth of the generalization, does Newton or Mill explicitly require the assignment of a probability to that generalization.

However, before we conclude that no probabilistic sin has been committed, we ought to look a little more closely at Mill. Unlike Newton, Mill offers views on probability. And there is some reason to suppose that he would subscribe to the idea that one infers an inductive conclusion with probability. In Chapter XVIII of Book 2 of *A System of Logic*, Mill advocates

what today we would call an *epistemic* concept of probability – a concept to be understood in terms of beliefs or expectations rather than in terms of belief-independent events or states of the world. He writes:

> We must remember that the probability of an event is not a quality of the event itself, but a mere name for the degree of ground which we, or someone else, have for expecting it. The probability of an event to one person is a different thing from the probability of the same event to another, or to the same person after he has acquired additional evidence. . . . But its probability to us means the degree of expectation of its occurrence, which we are warranted in entertaining by our present evidence. (Mill, 1888, p. 351)

Mill offers examples from games of chance, but, unlike the classical Laplacian, he says that probabilities should be assigned on the basis of observed relative frequencies of outcomes. For example, he writes:

> In the cast of a die, the probability of ace is one-sixth; not simply because there are six possible throws, of which ace is one, and because we do not know any reason why one should turn up rather than another, . . . but because we do actually know, either by reasoning or by experience, that in a hundred or a million of throws, ace is thrown in about one-sixth of that number, or once in six times. (Mill, 1888, p. 355)

For Mill, this knowledge of relative frequencies of events can be inferred from specific experiments or from a "knowledge of the causes in operation which tend to produce" the event in question. Either way, he regards the determination of the probability of an event to be based on an induction. For example, from the fact that the tossing of a die – either this die or others – has yielded an ace approximately one-sixth of the time, we infer that this will continue to be the case. So, one-sixth represents the degree of expectation for ace that we are warranted in having by our present evidence.

From his discussion it sounds as if Mill is an objectivist rather than a subjectivist about epistemic probability (more in the mold of Carnap's *Logical Foundations of Probability* than that of subjective Bayesians). Moreover (like Carnap), he is an objectivist who relativizes probability to an epistemic situation of an actual or potential believer, so that the probability depends on the particular epistemic situation in question. (In *The Book of Evidence* I contrast this view with an objective epistemic view that does not relativize probability to particular epistemic situations; in what follows, however, I will focus on an interpretation of the former sort, since I take it to be Mill's.)

Is Mill restricting probability to events and, more specifically, to events such as those in games of chance, in which relative frequencies can be determined in principle? He certainly goes beyond games of chance. Indeed, he

seems to want to assign probabilities to propositions or to allow conclusions to be drawn with a probability. In Chapter XXIII (pp. 386ff), "Of Approximate Generalizations and Probable Evidence," Mill considers propositions of the form "Most As are B" (and "X% of As are B"). He writes:

> If the proposition, Most A are B, has been established by a sufficient induction, as an empirical law, we may conclude that any particular A is B *with a probability proportional to the preponderence of the number of affirmative instances over the number of exceptions.* (Mill, 1888, p. 390; emphasis added)

To be sure, we have a sort of relative frequency idea here, and the conclusion he speaks of is not general but about a particular A being B.[4] However, what I find more interesting is the idea of concluding a proposition "with a probability" and the idea that the proposition concluded can concern a specific A rather than a type – ideas congenial to epistemic views of probability.[5]

Now let us assume that someone, call her Deborah, has made an induction in Mill's sense (which will be understood broadly to include inferences from a subset of a class to a universal or statistical generalization about that class, or to a claim about an individual arbitrarily chosen within the class but not in the subset. And let us suppose that her induction is justified. My question is this: Can Mill claim that Deborah's induction is justified only if he assumes that, given her knowledge of the premises and given what else she knows, the degree to which someone in that situation is warranted in drawing the inductive conclusion is high? If, as I believe, this condition is necessary for Mill, and if, as I am suggesting, probability for Mill is to be understood as objective epistemic probability, and if such probability is applicable to inductive conclusions, given the premises, then objective epistemic probabilists are ready for action.[6] They need not assign a point probability, or an interval, or a threshold (as I personally do). It can be quite

[4] Mill (1888, p. 186) explicitly allows inferences to "individual facts" to count as inductions. (In the quotation above, I take it he is talking about the "odds," or, in terms of probability, the number of affirmative instances over the sum of affirmative instances and exceptions.) Mill goes on to say that we do so assuming that "we know nothing except that [the affirmative instances] fall within the class A" (p. 391).

[5] Mill gives various examples and arguments of this type, including ones in which a proposition about a specific individual is concluded with a numerical probability (1888, pp. 391–2). It is clear from his discussion that he is presupposing standard rules of probability.

[6] Here I include those (such as Mill and Carnap) who relativize probability to an epistemic situation, and anyone (such as myself) who claims that such relativizations can but need not be made.

vague – for example, "highly probable," or with Carnap, "probability > k" (whatever k is supposed to be, even if it varies from one case to another). But objective epistemic probabilists – again I include Mill – are committed to saying that the inference is justified only if the objective epistemic probability (the "posterior" probability) of the inductive conclusion, given the truth of the premises, is sufficiently high.

Is Mayo committed to rejecting this idea? First, she, like Peirce, is definitely committed to doing so if probability is construed only as relative frequency. Relative frequency theorists reject assigning probabilities to individual hypotheses (for the Peircean "blackberry" argument). Indeed, she considers Mill's inference from "most As are Bs" to a probabilistic conclusion about a particular A being B to be committing what she calls the "fallacy of probabilistic instantiation" (Mayo, 2005, p. 114). Second, and most important, she believes rather strongly that you can assess inductive generalizations and other inductive arguments without assigning any probability to the conclusion. The only probability needed here, and this is a relative frequency probability, is the probability of getting the sort of experimental result we did using the test we did, under the assumption that the hypothesis in the conclusion is false. What we want in our inductions, says Mayo, are hypotheses that are "highly probed," not ones that are "highly probable" – that is, that have a high posterior probability.

Is part of her reason for rejecting the requirement of high posterior probability for an inductive conclusion this: that the only usable notion of probability is the relative frequency one, so that assigning an objective epistemic probability to a hypothesis either makes little sense, or is not computable, or some such thing? That is, does she reject any concept of objective epistemic probability for hypotheses? Or is Mayo saying that such a concept is just not needed to determine the goodness of an inductive argument to an empirical hypothesis? I believe she holds both views. In what follows, however, I will consider only the second, and perhaps the more philosophically challenging, view that, even if hypotheses have posterior probabilities on the evidence, for a hypothesis to be justified by the experimental data this posterior probability – whether vague or precise – need not be high, and indeed may be very low. It is this claim that I want to question.

Using Mill's terminology for probability, but applying what he says to hypotheses (and not just events), the probability of a hypothesis is to be understood as the "degree of expectation [of its truth] which we are warranted in entertaining by our present evidence" (Mill, 1888, p. 351). Or, using terminology I have proposed elsewhere, the probability of a hypothesis

h, given evidence e, is the degree of reasonableness of believing h, on the assumption of e. Now I ask, given that such a notion of probability makes sense, how could an inductive argument to a hypothesis from data be justified unless (to use my terminology) the degree of reasonableness of believing the conclusion given the data is high, or unless (to use Mill's terminology) the data warrant a high degree of expectation in the truth of the conclusion – that is, unless the posterior probability of the conclusion is high? (This I take to be a necessary condition, not a sufficient one; see Achinstein, 2001, ch. 6.)

I can see Mayo offering several responses. One, already mentioned, is to completely reject the idea of objective epistemic probability. A second, again mentioned earlier, is to say that Mill is committing the "fallacy of probabilistic instantiation." I do not think this is a fallacy when we consider epistemic probability. If 99% of As are B, then, given that a particular A was randomly selected, the degree of reasonableness of believing that this A is B is very high.[7]

A third response (which she may offer in addition to the first) is to say that, when we inductively infer a hypothesis, we infer not its truth or probability but its "reliability" in certain types of experimental situations. Although sometimes she says that from the fact that a hypothesis has passed a severe test (in her sense) we may infer the hypothesis (Achinstein, 2005, p. 99), and sometimes she says that we may infer that it is "correct,"[8] she also claims that passing a severe test shows that the hypothesis is "reliable." What does that mean? She writes: "Learning that hypothesis H is reliable, I propose, means learning that what H says about certain experimental results will often be close to the results actually produced – that H will or would often succeed in specified experimental applications" (Mayo, 1996, p. 10). Mayo wants to move beyond Karl Popper's view that passing a severe test simply means that, despite our efforts to falsify the hypothesis, the hypothesis remains unfalsified. Her idea of passing a severe test is such that, if a hypothesis does so, this tells us something important about how the hypothesis will perform in the future using this test. She writes: "Reliability deals with future performance, and corroboration, according to Popper, is only a 'report of past performance'" (Mayo, 1996, p. 9).

Let me formulate what I take to be Mayo's position concerning "passing a severe test" and "reliability." We have a hypothesis h, a test T for that

[7] See Mill (1888, pp. 390–1).

[8] "The evidence indicates the correctness of hypothesis H, when H passes a severe test" (Mayo, 1996, p. 64).

hypothesis, and some data D produced as a result of employing the test. For the hypothesis to pass a severe test T yielding data D, she requires (1) that the data "fit" the hypothesis in some fairly generous sense of "fit"[9]; and she requires (2) that the test T be a severe one in the sense that the probability is very high that T would produce a result that fits hypothesis h less well than data D if h were false (alternatively, the probability is very low that T would yield results that fit h as well as data D if h were false). In her view, the concept of probability here is relative frequency. If so, and taking a standard limiting relative frequency view, the probability she has in mind for claim 2 regarding a severe test can be expressed as follows:

Given that test T continues to be repeated, the relative frequency of outcomes fitting hypothesis h as well as data D will at some point in the testing sequence be and remain very low, under the assumption that h is false.

This is what is entailed by passing a severe test.

Now my question is this. Epistemically speaking, what, if anything, can one conclude from the fact that h has passed a severe test yielding data D?

Mill would want to conclude that "the degree of expectation . . . [in the truth of the hypothesis h] which we are warranted in entertaining by our evidence [i.e., by data D]" is high. More generally, I suggest, Mill requires that evidence should provide a good reason for believing a hypothesis. On my own view of evidence, the latter requires that the objective epistemic probability of the hypothesis h, given that data D were obtained as a result of test T, is very high. If Mayo were to accept this condition as necessary, then we would be in agreement that a posterior probability, however vaguely characterized, can be attributed to the hypothesis, and we would both agree that this probability is not a relative frequency.

The only problem I would anticipate here is that a posterior probability for h would presuppose the existence of a prior probability for h – a probability for h based on information other than data D. Suppose that such a prior probability for h is extremely low and that, despite the fact that h has passed a severe test, in Mayo's sense, with respect to data D, the posterior probability of h given data D is also very low. Then even though h has passed such a severe test with data D, I, and I believe Mill, would conclude that this fact does not warrant an inference to, or belief in, hypothesis h. If not, then "passing a severe test" in Mayo's sense would not be sufficient for inductive evidence in Mill's sense.

[9] Minimally she wants $p(D;h) > p(D;not\text{-}h)$. (See Mayo, 2005, p. 124, fn. 3.)

To introduce a very simple example, suppose there is a disease *S* for which (while the patient is alive) there is one and only one test, *T*. This test can yield two different results: the test result turns red or it turns blue. Suppose that, of those who have disease *S*, 80% test red using *T*, so that P(red result with test *T*/disease *S*) = .8. And suppose that, of those who don't have disease *S*, 2 out of 100,000 test red, so that P(red result with test *T*/−*S*) = .00002. The probabilities here are to be construed as relative frequencies. Now let the hypothesis under consideration be that Irving has disease *S*. On the basis of the aforementioned probabilities, we can write the following:

1. P(Irving's getting a red result with test *T*/ Irving has disease *S*) = .8

and

2. P(Irving's getting a red result with test *T*/Irving does not have disease *S*) = .00002.

If so, then Irving's test result of red "fits" the hypothesis that Irving has *S*, in Mayo's sense, because probability 1 is much greater than probability 2. And because probability 2 is so low, Irving's getting the red result with test *T* should count as passing a severe test for the hypothesis that he has disease *S*.

Finally, suppose that only one person in ten million in the general population has disease *S*, so that P(*S*) = .0000001. Using Bayes' theorem, we compute that P(*S*/red result with test *T*) = .004; that is, four out of one thousand people who get the red result from test *T* (less than half of 1%) have disease *S*. I regard such a test for disease *S* as pretty lousy. But using these frequency probabilities as a basis for epistemic ones (represented with a small p), we obtain

p(Irving has disease *S*) = .0000001,

and

p(Irving has disease *S*/Irving gets a red result with test *T*) = .004.

If so, epistemically speaking, Irving's red test result gives very little reason to believe that he has disease *S*, despite the fact that Irving has passed a "severe test" for disease *S* in Mayo's sense. In such a case, if passing a severe test is supposed to give us a good reason to believe a hypothesis, it does not do the job.[10]

[10] Determining the posterior probability of having a certain disease (say), given a certain test result, without determining the prior probability of that disease (the "base rate") is an example of what is called the "base rate fallacy."

On the other hand, suppose that Mayo refuses to assign a posterior probability to *h*. Suppose she claims that we do not want, need, or have any such probability, whether or not this is epistemic probability. Then I, and I believe Mill, would have a problem understanding what passing a severe test has to do with something we regard as crucial in induction – namely, providing a good reason to believe *h*. We have data *D* obtained as a result of test *T*, and those data "fit" hypothesis *h* in Mayo's sense. Now we are told that if we were to continue indefinitely to test *h* by employing test *T*, and if *h* were false, the relative frequency of getting results that fit *h* as well as data *D* do would eventually become and stay low. Let us even suppose (as we did before) that *T* is the only test for disease *S* that can be made while the patient is living. We can say that, if the frequency probabilities are as previously reported, then giving the *T*-test to Irving and getting result "red" is as good as it gets for determining *S* – it is a better test for *S* than any other. If repeated and the hypothesis *h* is false, the frequency of outcomes "fitting" *h* as well as the present outcome does will remain low. In this sense, it is a "good test." But, given the additional information I have supplied, passing that test is not a good reason, or a good enough reason, to believe the hypothesis.

If we accept Mayo's definition of a hypothesis "passing a severe test," then, I think, we have to distinguish between

1. a hypothesis passing such a test with data results *D*, and
2. a hypothesis passing such a test with results *D* being a good reason to believe or infer that the hypothesis is true.

Establishing 1 is not sufficient for establishing 2.

The theory of evidence I have defended elsewhere distinguishes various concepts of evidence in use in the sciences. But the most basic of these (which I call "potential" evidence) in terms of which the others can be defined requires that some fact *e* is evidence that *h* only if *e* provides a good reason for believing *h*. Such a reason is provided, I claim, not simply when the (epistemic) posterior probability of *h*, given *e*, is high, but when the probability of an explanatory connection between *h* and *e*, given *e*, is high (which entails the former). I will not here defend the particulars of this view; I will say only that, where data from tests are concerned, the view of evidence I defend requires satisfying 2 above, not simply 1.

Finally, in this part of the discussion, let me note that Mayo's theory is being applied only to *experimental* evidence: data produced by experiments and observations as a result of what she calls a test. She is concerned with the question of whether the test results provide a severe test of the hypothesis

and, hence, evidence for it. I agree that this question about evidence and inductive reasoning is important. But what happens if we take it a step higher and ask whether the facts described by propositions highly probed by the experimental and observational data can themselves be regarded as evidence for higher level propositions or theories. This is precisely what Newton does in the third book of the *Principia.*

On the basis of different observations of distances of the four known moons of Jupiter from Jupiter – astronomical observations made by Borelli, Townly, and Cassini – Newton infers Phenomenon 1: "that the satellites of Jupiter, by radii drawn to the center of Jupiter, describe areas proportional to the times, and their periodic times . . . are as 3/2 powers of the distances from that center." Similar so-called phenomena are inferred about the moons of Saturn, about our moon, and about the primary planets. Now appealing to his principles and theorems of mechanics in Book 1, Newton regards the six phenomena as providing very strong evidence for his universal law of gravity. (He speaks of "the argument from phenomena . . . for universal gravity" [Newton, 1999, p. 796].) More generally, how will this be understood in the error-statistical view of evidence? Should only the observational data such as those reported by Borelli, Townly, and Cassini be allowed to count as evidence for Newton's universal law of gravity? Or can the error-statistical theory be applied to Newton's phenomena themselves? Can we assign a probability to Phenomenon 1, given the assumption of falsity of Newton's law of gravity? If so, how can this probability be understood as a relative frequency? More generally, how are we to understand claims about evidence when the evidence consists in more or less theoretical facts and the hypothesis is even more theoretical? I will simply assert without argument that an account of evidence that appeals to objective epistemic posterior probabilities can make pretty good sense of this. Can Mayo?

4 The Principle of the Uniformity of Nature: Sin 4

Now, more briefly, we have the fourth and final sin – the claim that Mill appeals to an unnecessary, unwarranted, and vague principle of the uniformity of nature.

Mill's discussion here (1888, pp. 200–6) is admittedly somewhat confusing. First, he says, "there is an assumption involved in every case of induction," namely that "what happens once will, under a sufficient degree of similarity of circumstances, happen again, and not only again, but as often as the same circumstances recur." He calls this a "universal fact, which

is our warrant for all inferences from experience." But second, he claims that this proposition, "that the course of nature is uniform," is itself "an instance of induction." It is not the first induction we make, but one of the last. It is founded on other inductions in which nature is concluded to be uniform in one respect or another. Third, he claims that, although the principle in question does not contribute to proving any more particular inductive conclusion, it is "a necessary condition of its being proved." Fourth, and perhaps most interesting, in an important footnote he makes this claim:

> But though it is a condition of the validity of every induction that there be uniformity in the course of nature, it is not a necessary condition that the uniformity should pervade all nature. It is enough that it pervades the particular class of phenomena to which the induction relates. . . . Neither would it be correct to say that every induction by which we infer any truth implies the general fact of uniformity as *foreknown*, even in reference to the kind of phenomena concerned. It implies *either* that this general fact is already known, *or* that we may now know it. (Mill, 1888, p. 203)

In this footnote Mill seems to abandon his earlier claim that the principle of uniformity of nature is "our warrant for all inferences from experience."

Possibly a more appealing way to view Mill's principle is as follows. Instead of asserting boldly but rather vaguely that nature is uniform, Mill is claiming somewhat less vaguely, and less boldly, that there are uniformities in nature – general laws governing various types of phenomena – which are inductively and validly inferred, perhaps using Mill's methods. That such laws exist is for Mill an empirical claim. Suppose that phenomena governed by such laws bear a similarity to others which by induction we infer are governed by some similar set of laws. Then, Mill may be saying, we can use the fact that one set of phenomena is so governed to strengthen the inference to laws regarding the second set.

So, for example, from the inductively inferred Phenomenon 1 regarding the motions of the moons of Jupiter, together with mechanical principles from Book 1, Newton infers Proposition 1: the force governing the motions of the moons of Jupiter is a central inverse-square force exerted by Jupiter on its moons. Similarly from the inductively inferred Phenomenon 2 regarding the motions of the moons of Saturn, Newton infers that a central inverse-square force is exerted by Saturn on its moons. The latter inference is strengthened, Mill may be saying, by the existence of the prior uniformity inferred in the case of Jupiter.

Finally, Mill's uniformity principle can be understood not just as a claim that uniformities exist in nature, and that these can, in appropriate

circumstances, be used to strengthen claims about other uniformities, but also as a methodological injunction in science: look for such uniformities. So understood, what's all the fuss?

5 Mill: Saint or Sinner?

Shall we conclude from this discussion that Mill is a saint regarding induction? I do not know that I have proved that, but at the very least I hope I have cast some doubt on the view that he is an unworthy sinner. If it is a sin to assign a probability to an inductive conclusion, as Mayo, following Peirce, seems to believe, then at least we can understand the good motive behind it: When we want evidence sufficient to give us a good reason to believe a hypothesis, Mill would say, then, as good as Mayo's "severe-testing" is, we want more.

Appendix: The Stories of Isaac and "Stand and Deliver"

Mayo introduces an example in which there is a severe test, or battery of tests, T to determine whether high school students are ready for college (Achinstein, 2005, pp. 96–7, 115–17) The test is severe in the sense that passing it is very difficult to do if one is not ready for college. Mayo imagines a student Isaac who has taken the test and achieved a high score X, which is very rarely achieved by those who are not college-ready. She considers Isaac's test results X as strong evidence for the hypothesis

H: Isaac is college-ready.

Now suppose we take the probability that Isaac would get those test results, given that he is college-ready, to be extremely high, so that

$$p(X/H) \text{ is practically } 1,$$

whereas the probability that Isaac would get those test results, given that he is not college-ready, is very low, say

$$p(X/\sim H) = .05.$$

But, says Mayo's imagined critic, suppose that Isaac was randomly selected from a population in which college readiness is extremely rare, say one out of one thousand. The critic infers that

$$\text{(A)} \quad p(H) = .001.$$

If so, then the posterior probability that Isaac is college-ready, given his high test results, would be very low; that is,

$$\text{(B)} \quad p(H/X) \text{ is very low,}$$

even though the probability in B would have increased from that in A. The critic regards conclusion B as a counterexample to Mayo's claim that Isaac's test results X provide strong evidence that Isaac is college-ready.

Mayo's response is to say that to infer conclusions A and B is to commit the fallacy of probabilistic instantiation. Perhaps, even worse, it is to engage in discrimination, since the fact that we are assigning such a low prior probability to Isaac's being college-ready prevents the posterior probability of his being college-ready, given his test results, from being high; so we are holding poor disadvantaged Isaac to a higher standard than we do test-takers from an advantaged population.

My response to the probabilistic fallacy charge is to say that it would be true if the probabilities in question were construed as relative frequencies. However, as I stressed in the body of the chapter, I am concerned with epistemic probability. If all we know is that Isaac was chosen at random from a very disadvantaged population, very few of whose members are college ready, say one out of one thousand, then we would be justified in believing that it is very unlikely that Isaac is college-ready (i.e., conclusion A and, hence, B).

The reader may recall the 1988 movie *Stand and Deliver* (based on actual events) in which a high school math teacher in a poor Hispanic area of Los Angeles with lots of student dropouts teaches his students calculus with their goal being to pass the advanced placement calculus test. Miraculously all the students pass the test, many with flying colors. However, officials from the Educational Testing Service (ETS) are very suspicious and insist that the students retake the test under the watchful eyes of their own representatives. Of course, the students feel discriminated against and are very resentful. But the ending is a happy one: once again the group passes the test.

Now two questions may be raised here. First, given that all they knew were the test results and the background of the students, were the officials from the ETS justified in believing that, despite the high test results, it is unlikely that the students have a good knowledge of basic calculus? I think that this is a reasonable epistemic attitude. They assigned a very low prior epistemic probability to the hypothesis that these students know calculus and, hence, a low posterior probability to this hypothesis given the first test results. We may suppose that after the results of the second test, conducted under more stringent testing conditions, the posterior probability of the

hypothesis was significantly high. To be sure, we the viewers of the film had much more information than the ETS officials: we saw the actual training sessions of the students. So the posterior epistemic probability we assigned was determined not only by the prior probability but by the test results and this important additional information.

The second question, which involves a potential charge of discrimination, is one about what action to take given the epistemic probabilities. After the first test results, what, if anything, should the ETS officials have done? This depends on more than epistemic probabilities. Despite low probabilities, it might have been judged better on grounds mentioned by Mayo – discrimination – or on other moral, political, or practical grounds not to repeat the test but to allow the test results to count as official scores to be passed on to the colleges. Or, given the importance of providing the colleges with the best information ETS officials could provide, the most appropriate course might well have been to do just what they did.

References

Achinstein, P. (2001), *The Book of Evidence*, Oxford University Press, New York.

Achinstein, P., ed. (2005), *Scientific Evidence: Philosophical Theories and Applications*, Johns Hopkins University Press, Baltimore, MD.

Mandelbaum, M. (1964), *Philosophy, Science, and Sense Perception*, Johns Hopkins University Press, Baltimore, MD.

Mayo, D.G. (1996), *Error and the Growth of Experimental Knowledge*, University of Chicago Press, Chicago.

Mayo, D.G. (2005), "Evidence as Passing Severe Tests: Highly Probed vs. Highly Proved," pp. 95–127 in P. Achinstein (ed.), *Scientific Evidence: Philosophical Theories and Applications*, Johns Hopkins University Press, Baltimore, MD.

Mill, J.S. (1888), *A System of Logic*, 8th edition, Harper and Bros., New York.

Newton, I. (1999), *Principia* (trans. I.B. Cohen and A. Whitman), University of California Press.

Peirce, C.S. (1931–1935), *Collected Papers*, C. Hartshorne and P. Weiss (eds.), Harvard University Press, Cambridge, MA.

Sins of the Epistemic Probabilist
Exchanges with Peter Achinstein

Deborah G. Mayo

1 Achinstein's Sins

As Achinstein notes, he and I agree on several key requirements for an adequate account of evidence: it should be objective and not subjective, it should include considerations of flaws in data, and it should be empirical and not a priori. Where we differ, at least at the moment, concerns the role of probability in inductive inference. He takes the position that probability is necessary for assigning degrees of objective warrant or "rational" belief to hypotheses, whereas, for the error statistician, probability arises to characterize how well probed or tested hypotheses are (highly probable vs. highly probed). Questions to be considered are the following:

1. *Experimental Reasoning*: How should probability enter into inductive inference – by assigning degrees of belief or by characterizing the reliability of test procedures?
2. *Objectivity and Rationality*: How should degrees of objective warrant be assigned to scientific hypotheses?
3. *Metaphilosophy*: What influences do philosophy-laden assumptions have on interpretations of historical episodes: Mill?

I allude to our disagreeing "at the moment" because I hope that the product of this latest installment in our lengthy exchange may convince him to shift his stance, at least slightly.

In making his case, Achinstein calls upon Mill (also Newton, but I largely focus on Mill). Achinstein wishes "to give Mill a better run for his money" by portraying him as an epistemic probabilist of the sort he endorses. His question is whether Mill's inductive account contains some features "that are or seem inimical to [the] error-statistical philosophy." A question for us is whether Achinstein commits some fundamental errors, or "sins" in

his terminology, that are common to epistemic (Bayesian) probabilists in philosophy of science, and are or seem to be inimical to the spirit of an "objective" epistemic account. What are these sins?

1. *First*, there are *sins of "confirmation bias"* or "reading one's preferred view into the accounts of historical figures" or found in scientific episodes, even where this reconstruction is at odds with the apparent historical evidence.
2. *Second*, there are *sins of omission*, whereby we are neither told how to obtain objective epistemic probabilities nor given criteria to evaluate the success of proposed assignments.

These shortcomings have undermined the Bayesian philosophers' attempts to elucidate inductive inference; considering Achinstein's discussion from this perspective is thus of general relevance for the popular contemporary project of Bayesian epistemologists.

1.1 Mill's Innocence: Mill as Error Statistician

Before considering whether to absolve Mill from the sins Achinstein considers, we should note that Achinstein omits the one sin discussed in Mayo (1996) that arises most directly in association with Mill: Mill denies that novel predictions count more than nonnovel ones (pp. 252, 255). My reference to Mill comes from Musgrave's seminal paper on the issue of novel prediction:

> According to modern logical empiricist orthodoxy, in deciding whether hypothesis h is confirmed by evidence e... we must consider only the statements h and e, and the logical relations between them. It is quite irrelevant whether e was known first and h proposed to explain it, or whether e resulted from testing predictions drawn from h. (Musgrave, 1974, p. 2)

"We find" some variant of the logicist approach, Musgrave tells us, "in Mill, who was amazed at Whewell's view" that successfully predicting novel facts gives a hypothesis special weight. "In Mill's view, 'such predictions and their fulfillment are... well calculated to impress the uninformed' " (Mill, 1888, Book 3, p. 356). This logicism puts Mill's straight rule account at odds with the goal of severity.

Nevertheless, Achinstein gives evidence that Mill's conception is sufficiently in sync with error statistics so that he is prepared to say, "I would take Mill to be espousing at least some version of the idea of 'severe testing'" (p. 174). Here is why: Although Mill regards the form of induction to be "induction by simple enumeration," whether any particular instantiation is valid also "depends on nonformal empirical facts regarding the sample, the sampling, and the properties being generalized" (Achinstein, p. 174) – for example, Mill would eschew inadequately varied data, failure to look for negative instances, and generalizing beyond what the evidence allows. These features could well set the stage for expiating the sins of the straight ruler: The "fit" requirement, the first clause of severity, is met because observed cases of As that are Bs "fit" the hypothesis that all As are Bs; and the additional considerations Achinstein lists would help to satisfy severity's second condition – that the ways such a fit could occur despite the generalization being false have been checked and found absent.

Achinstein finds additional intriguing evidence of Mill's error-statistical leanings. Neither Mill's inductions (nor those of Newton), Achinstein points out, are "guilty" of assigning probabilities to an inductive conclusion:

> Neither in their abstract formulations of inductive generalizations... nor in their examples of particular inductions... does the term "probability" occur.... From the inductive premises we simply conclude that the generalization is true, or as Newton allows in rule 4, "very nearly true," by which he appears to mean not "probably true" but "approximately true." (p. 176)

Neither exemplar, apparently, requires the assignment of a probability to inductively inferred generalizations. One might have expected Achinstein to take this as at least casting doubt on his insistence that such posterior probabilities are necessary. He does not.

1.2 Mill's Guilt: Mill as Epistemic Probabilist

Instead, Achinstein declares that their (apparent) avoidance of any "probabilistic sin" is actually a transgression to be expiated, because it conflicts with his own account! "However, before we conclude that no probabilistic sin has been committed, we ought to look a little more closely at Mill." In particular, Achinstein thinks we ought to look at contexts where Mill *does* talk about probabilities – namely in speaking about events – and substitute what he says there to imagine he is talking about probabilities of hypotheses. By assigning to Mill what is a probabilistic sin to error statisticians, Mill is restored to good graces with Achinstein's Bayesian probabilist.

Mill offers examples from games of chance, Achinstein notes (p. 177), where probabilities of general outcomes are assigned on the basis of their observed relative frequencies of occurrence:

> In the cast of a die, the probability of ace is one-sixth... because we do actually know, either by reasoning or by experience, that in a hundred or a million of throws, ace is thrown about one-sixth of that number, or once in six times. (Mill, 1888, p. 355)

In more modern parlance, Mill's claim is that we may accept or infer ("by reasoning or by experience") a statistical hypothesis H that assigns probabilities

to outcomes such as "ace." For the error statistician, this inductive inference takes place by passing H severely or with appropriately low error probabilities. Although there is an inference to a probabilistic model of experiment, which in turn assigns probabilities to outcomes, there is no probabilistic assignment to H itself, nor does there seem to be in Mill. So such statements from Mill do not help Achinstein show that Mill intends to assign posterior probabilities to hypotheses.[1]

The frequentist statistician has no trouble agreeing with the conditional probabilities of events stipulated by Achinstein. For example, we can assert the probability of spades given the outcome "ace" calculated under hypothesis H that cards are randomly selected from a normal deck, which we may write as P(spade | ace; H) – noting the distinction between the use of "|" and ";". However, we would not say that an event severely passes, or that one event severely passes another event. A statistical hypothesis must assign probabilities to outcomes, whereas the event "being an ace" does not.

I claim no Mill expertise, yet from Achinstein's own presentation, Mill distinguishes the assignment of probabilities to events from assigning probabilities to hypotheses, much as the error statistician. What then is Achinstein's justification for forcing Mill into a position that Mill apparently rejects? Achinstein seems guilty of the sin of "reading our preferred view into the accounts of historical episodes and/or figures."

2 The Error-Statistical Critique

2.1 Some Sleight-of-Hand Sins

Having converted Mill so that he speaks like a Bayesian, Achinstein turns to the business of critically evaluating the error statistician: "[O]bjective epistemic probabilists – again I include Mill – are committed to saying that the inference is justified only if the objective epistemic probability (the 'posterior' probability) of the inductive conclusion" is sufficiently high (p. 179). At times during our exchanges, I thought Achinstein meant this assertion as a tautology, playing on the equivocal uses of "probability" in ordinary language. If high epistemic probability is just shorthand for high inductive

[1] On the other hand, they do raise the question of how Achinstein's epistemic probabilist can come to accept a statistical model (on which probabilistic assignments to events are based). Achinstein's formal examples start out by assuming that we have accepted a statistical model of experiment, H, usually random sampling from a binomial distribution. Is this acceptance of H itself to be a matter of assigning H a high posterior probability through a Bayesian calculation? The alternatives would have to include all the ways the model could fail. If not, then Achinstein is inconsistent in claiming that warranting H must take the form of a Bayesian probability computation.

warrant, then his assertion may be uncontroversial, if also uninformative. If hypothesis H passes a highly severe test, there is no problem in saying there is a high degree of warrant for H. But there is a problem if this is to be cashed out as a posterior probability, attained from Bayes's theorem and satisfying the probability axioms. Nevertheless, Achinstein assures me that this is how he intends to cash out epistemic probability. High epistemic probability is to be understood as high posterior probability (obtainable from Bayes's theorem). I construe it that way in what follows.

Achinstein presupposes that if we speak of data warranting an inference or belief in H, then we must be talking about an epistemic posterior probability in H, but this is false. Musgrave's critical rationalist would reject this as yet another variant on "justificationism" (see Chapter 3). Achinstein's own heroes (Mill and Newton) attest to its falsity, because as he has convincingly shown, they speak of warranting hypotheses without assigning them posterior probabilities.

In the realm of formal statistical accounts of induction, it is not only the error statistician who is prepared to warrant hypotheses while eschewing the Bayesian algorithm. The "likelihoodist," for instance, might hold that data $\mathbf{x}$ warrants H to the extent that H is more likely (given $\mathbf{x}$) than rivals. High likelihood for H means P($\mathbf{x}$;H) is high, but high likelihood for H does *not* mean high probability for H. To equate P($\mathbf{x}|H$) and P($H|\mathbf{x}$) immediately leads to contradictions – often called the prosecutor's fallacy. For example, the likelihood of H and not-H do not sum to 1 (see Chapters 7 and 9, this volume).

If Achinstein were to accept this, then we would be in agreement that probabilistic concepts (including frequentist ones) may be used to qualify the evidential warrant for hypothesis H while denying that this is given by a probability assignment to H. Achinstein may grant these other uses of probability for induction yet hold that they are inferior to his idea of objective epistemic posterior probabilities. His position might be that, unless the warrant is a posterior probability, then it does not provide an adequate objective inductive account. This is how I construe his position.

2.2 Sins of Omission: How Do We Apply the Method? What Is So Good About It?

For such a sweeping claim to have any substance, it must be backed up with (1) some guidance as to how we are to arrive at objective epistemic posteriors and (2) an indication of (desirable) inferential criteria that objective epistemic probability accounts satisfy.

So how do we get to Achinstein's objective epistemic posterior probabilities? By and large this is not one of Achinstein's concerns; he considers

informal examples where intuitively good evidence exists for a claim H, and then he captures this by assigning H a high degree of epistemic probability. If he remained at the informal level, the disagreements he finds between us would likely vanish, for then high probability could be seen as a shorthand for high inductive warrant, the latter attained by a severe test. When he does give a probabilist computation, he runs into trouble.

Unlike the standard Bayesian, Achinstein does not claim that high posterior probability is sufficient for warrant or evidence, but does "take [it] to be a necessary condition" (p. 180). It is not sufficient for him because he requires, in addition, what he describes as a non-Bayesian "explanatory connection" between data and hypotheses (p. 183). Ignoring the problem of how to determine the explanatory connection, and whether it demands its own account of inference (rendering his effort circular), let us ask: Are we bound to accept the necessity? My question, echoing Achinstein, is this: *Epistemically speaking, what, if anything, can one conclude about hypothesis H itself from the fact that Achinstein's method accords H a high objective epistemic probability?*

That is, we require some indication that H's earning high (low) marks on Achinstein's objective probability scale corresponds to actually having strong (weak) evidence of the truth of H. If Achinstein could show this, his account would be superior to existing accounts of epistemic probability! It will not do to consider examples where strong evidence for H is intuitively sound and then to say we have high Achinstein posterior epistemic probability in H; he would have to demonstrate it with posteriors arrived at through Achinsteinian means.

3 Achinstein's Straight Rule for Attaining Epistemic Probabilities

To allow maximum latitude for Achinstein to make his case, we agree to consider the Achinsteinian example. The most clear-cut examples instantiate a version of a "straight rule," where we are to consider a "hypothesis" that consists of asserting that a sample possesses a characteristic such as "having a disease" or "being college-ready." He is led to this peculiar notion of a hypothesis because he needs to use techniques for probability assignments appropriate only for events. But let us grant all of the premises for Achinstein's examples.[2] We have (see p. 187):

[2] In an earlier exchange with Colin Howson, I discuss converting these inadmissible hypotheses into legitimate statistical ones to aid the criticism as much as possible, and I assume the same here (Mayo, 1997).

Achinstein's Straight Rule for Objective Epistemic Probabilities: If (we know only that) a_i is randomly selected from a population where p% have property C, then the objective epistemic probability that a_i has C equals p.

Next, Achinstein will follow a variant on a Bayesian gambit designed to show that data **x** from test T can pass hypothesis H severely, even though H is accorded a low posterior probability – namely by assuming the prior probability of H is sufficiently low. (Such examples were posed post-EGEK, first by Colin Howson (1997a).) I have argued that in any such example it is the Bayesian posterior, and not the severity assignment, that is indicted (Mayo 2005, 2006). This criticism and several like it have been discussed elsewhere, most relevantly in a collection discussing Achinstein's account of evidence (Mayo, 2005)! I focus mainly on the update to our earlier exchange, but first a quick review.

3.1 College Readiness

We may use Achinstein's summary of one of the canonical examples (pp. 186–7). We are to imagine a student Isaac who has taken the test and achieved a high score X, which is very rarely achieved by those who are not college-ready. Let

H(I): Isaac is college-ready.

Let H' be the denial of H:

H'(I) Isaac is not college-ready, i.e., he is deficient.

I use H(I) to emphasize that the claim is about Isaac}.

Let S abbreviate: Isaac gets a high score (as high as X).

In the error statistical account, S is evidence against Isaac's deficiency, and for H(I). (Although we would consider degrees of readiness, I allow his dichotomy for the sake of the example.) Now we are to suppose that the probability that Isaac would get those test results, given that he is college-ready, is extremely high, so that

$$P(S|H(I)) \text{ is practically 1,}$$

Whereas the probability that Isaac would get those test results, given that he is not college-ready, is very low, say,

$$P(S|H'(I)) = .05.$$

Suppose that Isaac was randomly selected from a population – call it Fewready Town – in which college readiness is extremely rare, say one

out of one thousand. The critic infers that

$$(*)\ \mathrm{P}(H(\mathrm{I})) = .001.$$

If so, then the posterior probability that Isaac is college-ready, given his high test results, would be very low; that is,

$$\mathrm{P}(H(\mathrm{I})|S) \text{ is very low,}$$

even though in this case the posterior probability has increased from the prior in (*).

The critic – for example, Achinstein – regards the conclusion as problematic for the severity account because, or so the critic assumes, the frequentist would also accept (*) $\mathrm{P}(H) = .001$. Here is the critic's flaw. Although the probability of college readiness in a randomly selected student from high schoolers from Fewready Town is .001, it does not follow that Isaac, the one we happened to select, has a probability of .001 of being college-ready (Mayo, 1997a, 2005, p. 117). To suppose it does is to commit what may be called a fallacy of probabilistic instantiation.

3.2 Fallacy of Probabilistic Instantiation

We may abbreviate by $\mathrm{P}(H(x))$ the probability that a randomly selected member of Fewready has the property C. Then the probabilistic instantiation argues from the first two premises,

$$\mathrm{P}(H(x)) = .001.$$

The randomly selected student is I.

To the inference:

$$(*)\ \mathrm{P}(H(\mathrm{I})) = .001.$$

This is fallacious. We need not preclude that $H(\mathrm{I})$ has a legitimate frequentist prior; the frequentist probability that Isaac is college-ready might refer to generic and environmental factors that determine the chance of his deficiency – although I do not have a clue how one might compute it. But this is not what the probability in (*) gives us. Now consider Achinstein's new update to our exchanges on Isaac.

3.3 Achinstein's Update

Achinstein now accepts that this assignment in (*) is a sin for a frequentist:

> [T]he probabilistic fallacy charge would be true if the probabilities in question were construed as relative frequencies. However . . . I am concerned with epistemic probability. If all we know is that Isaac was chosen at random from a very disadvantaged population, very few of whose members are college ready, say one out of one thousand, then we would be justified in believing that it is very [improbable] that Isaac is college-ready. (Achinstein, p. 187)

Hence, (*) gives a legitimate objective epistemic frequentist prior.

Therefore, even confronted with Isaac's high test scores, Achinstein's probabilist is justified in denying that the scores are good evidence for $H(\mathrm{I})$. His high scores are instead grounds for believing $H'(\mathrm{I})$, that Isaac is not college-ready. It is given that the posterior for $H'(\mathrm{I})$ is high, and certainly an explanatory connection between the test score and readiness exists. Although the posterior probability of readiness has increased, thanks to his passing scores, for Achinstein this does not suffice to provide epistemic warrant for H (he rejects the common Bayesian distinction between increased support and confirmation). Unless the posterior reaches a threshold of a fairly high number, he claims, the evidence is "lousy." The example considers only two outcomes: reaching the high scores or not, i.e., S or $\sim S$. Clearly a lower grade gives even less evidence of readiness; that is, $\mathrm{P}(H'(\mathrm{I})|\sim S) > \mathrm{P}(H'(I)|S)$. Therefore, whether Isaac scored a high score or not, Achinstein's epistemic probabilist reports justified high belief that Isaac is not ready. The probability of Achinstein finding evidence of Isaac's readiness even if in fact he is ready (H is true) is low if not zero. Therefore, Achinstein's account violates what we have been calling the most minimal principle for evidence!

The *weak severity principle*: Data **x** fail to provide good evidence for the truth of H if the inferential procedure had very little chance of providing evidence against H, even if H is false.

Here the relevant H would be $H'(\mathrm{I})$ – Isaac is not ready.

If Achinstein allows that there is high objective epistemic probability for H even though the procedure used was practically guaranteed to arrive at such a high posterior despite H being false, then (to echo him again) the error statistician, and I presume most of us, would have a problem understanding what high epistemic probability has to do with something we regard as crucial in induction, namely ensuring that an inference to H is based on genuine evidence – on data that actually discriminate between the truth and falsity of H.

This fallacious argument also highlights the flaw in trying to glean reasons for epistemic belief by means of just any conception of "low frequency of error." If we declared "unready" for any member of Fewready, we would rarely be wrong, but in each case the "test" has failed to discriminate the particular student's readiness from his unreadiness. We can imagine a context where we are made to bet on the generic event – the next student randomly

selected from the population has property C. But this is very different from having probed whether this student, Isaac, is ready or not – the job our test needs to perform for scientific inference.

I cannot resist turning Achinstein's needling of me back on him: Is Achinstein really prepared to claim there is high epistemic warrant for $H'(I)$ even though the procedure had little or no probability of producing evidence against $H'(I)$ and for $H(I)$ even if $H(I)$ is true? If he is, then the error statistician, and perhaps Mill, would have a hard time understanding what his concept has to do with giving an objective warrant for belief.

Let us take this example a bit further to explain my ironic allegation regarding "reverse discrimination." Suppose, after arriving at high belief in Issac's unreadiness, Achinstein receives a report of an error: in fact Isaac was selected randomly, not from Fewready Town, but from a population where college readiness is common, Fewdeficient Town. The same score now warrants Achinstein's assignment of a strong objective epistemic belief in Isaac's readiness (i.e., $H(I)$). A high school student from Fewready Town would need to have scored quite a bit higher on these same tests than one selected from Fewdeficient Town for his scores to be considered evidence of his readiness. So I find it surprising that Achinstein is content to allow this kind of instantiation to give an objective epistemic warrant.

3.4 The Case of General Hypotheses

When we move from hypotheses like "Isaac is college-ready" (which are really events) to generalizations – which Achinstein makes clear he regards as mandatory if an inductive account is not to be "puerile" – the difficulty for the epistemic probabilist becomes far worse if, like Achinstein, we are to obtain epistemic probabilities via his frequentist straight rule.

To take an example discussed earlier, we may infer with severity that the relativistic light deflection is within $\pm\ \varepsilon$ units from the GTR prediction, by finding that we fail to reject a null hypothesis with a powerful test. But how can a frequentist prior be assigned to such a null hypothesis to obtain the epistemic posterior?

The epistemic probabilist would seem right at home with a suggestion some Bayesians put forward – that we can apply a version of Achinstein's straight rule to hypotheses. We can imagine that the null hypothesis is

H_0: There are no increased risks (or benefits) associated with hormone replacement therapy (HRT) in women who have taken HRT for 10 years.

Suppose we are testing for discrepancies from zero in both positive and negative directions (Mayo, 2003). In particular, to construe the truth of a general hypothesis as a kind of "event," it is imagined that we sample

randomly from a population of hypotheses, some proportion of which are assumed true. The proportion of these hypotheses that have been found to be true in the past serves as the prior epistemic probability for H_0.

Therefore, if H_0 has been randomly selected from a pool of null hypotheses, 50% of which are true, we have

$$(*)\ P(H_0) = .5.$$

Although (*) is fallacious for a frequentist, once again Achinstein condones it as licensing an objective epistemic probability. But which pool of hypotheses should we use? The percentages "initially true" will vary considerably, and each would license a distinct "objective epistemic" prior. Moreover, it is hard to see that we would ever know the proportion of true nulls rather than merely the proportion that have thus far not been rejected by other statistical tests!

The result is a kind of "innocence by association," wherein a given H_0, asserting no change in risk, gets the benefit of having been drawn from a pool of true or not-yet-rejected nulls, much as the member from Fewready Town is "deficient by association." Perhaps the tests have been insufficiently sensitive to detect risks of interest. Why should that be grounds for denying evidence of a genuine risk with respect to a treatment (e.g., HRT) that *does* show statistically significant risks?

To conclude, here is our answer to Achinstein's question:

Does Mayo allow that H may pass with high severity when the posterior probability of H is not high?

In no case would H pass severely when the grounds for warranting H are weak. But high posteriors need not correspond to high evidential warrant. Whether the priors come from frequencies or from "objective" Bayesian priors, there are claims that we would want to say had passed severely that do not get a high posterior (see Chapter 7, this volume). In fact, statistically significant results that we would regard as passing the nonnull hypothesis severely can show a decrease in probability from the prior (.5) to the posterior (see Mayo, 2003, 2005, 2006).

4 Some Futuristic Suggestions for Epistemic Probabilists

In the introductory chapter of this volume, we mentioned Achinstein's concession that "standard philosophical theories about evidence are (and ought to be) ignored by scientists" (2001, p. 3) because they view the question of whether data **x** provide evidence for H as a matter of purely logical computation, whereas whether data provide evidence for hypotheses is not an a priori but rather an empirical matter. He appears to take those

failures to show that philosophers can best see their job as delineating the concepts of evidence that scientists seem to use, perhaps based on rational reconstructions of figures from the historical record.

Some may deny it is sinful that *epistemic probabilists* omit the task of how to obtain, interpret, and justify their objective epistemic probabilities, and claim I confuse the job of logic with that of methodology (Buldt 2000). Colin Howson, who had already denied that it was part of the job of Bayesian logic to supply the elements for his (subjective) Bayesian computation, declared in 1997 that he was moving away from philosophy of statistics to focus on an even purer brand of "inductive logic"; and he has clearly galvanized the large contemporary movement under the banner of "Bayesian epistemology" (see Glymour, Chapter 9, this volume). Two points: First, it is far from clear that Bayesian logics provide normative guidance about "rational" inference[3]; after all, error statistical inference embodies its own logic, and it would be good to explore which provides a better tool for understanding scientific reasoning and inductive evidence. Second, there is a host of new foundational problems (of logic and method) that have arisen in Bayesian statistical practice and in Bayesian-frequentist "unifications" in the past decade that are omitted in the Bayesian epistemological literature (see Chapter 7.2). It is to be hoped that days of atonement will soon be upon us.

References

Achinstein, P. (2001), *The Book of Evidence*, Oxford University Press, Oxford.

Buldt, B. (2000), "Inductive Logic and the Growth of Methodological Knowledge, Comment on Mayo," pp. 345–54 in M. Carrier, G. Massey, and L. Ruetsche (eds.), *Science at Century's End, Philosophical Questions on the Progress and Limits of Science*, University of Pittsburgh Press, Pittsburgh.

Howson, C. (1997a), "A Logic of Induction," *Philosophy of Science*, 64: 268–90.

Mayo, D.G. (1996), *Error and the Growth of Experimental Knowledge*, The University of Chicago Press, Chicago.

Mayo, D.G. (1997a), "Response to Howson and Laudan," *Philosophy of Science*, 64: 323–33.

Mayo, D.G. (2003), "Could Fisher, Jeffreys and Neyman Have Agreed on Testing? Commentary on J. Berger's Fisher Address," *Statistical Science*, 18: 19–24.

Mayo, D.G. (2005), "Evidence as Passing Severe Tests: Highly Probed vs. Highly Proved," pp. 95–127 in P. Achinstein (ed.), *Scientific Evidence*, Johns Hopkins University Press, Baltimore, MD.

[3] Even for the task of analytic epistemology, I suggest that philosophers investigate whether appealing to error-statistical logic offers a better tool for characterizing probabilistic knowledge than the Bayesian model. Given Achinstein's threshold view of evidential warrant, it is hard to see why he would object to using a severity assessment to provide the degree of epistemic warrant he seeks.

Mayo, D.G. (2006), "Philosophy of Statistics," pp. 802–15 in S. Sarkar and J. Pfeifer (eds.), *Philosophy of Science: An Encyclopedia*, Routledge, London.

Mill, J.S. (1888), *A System of Logic*, 8th edition, Harper and Bros., New York.

Musgrave, A. (1974), "Logical versus Historical Theories of Confirmation," *British Journal for the Philosophy of Science*, 25: 1–23.

Related Exchanges

Achinstein, P. (2000), "Why Philosophical Theories of Evidence Are (and Ought to Be) Ignored By Scientists, pp. S180–S192 in D. Howard (ed.), " *Philosophy of Science*, 67 (Symposia Proceedings).

Giere, R.N. (1997b), "Scientific Inference: Two Points of View," pp. S180–S184 in L. Darden (ed.), *Philosophy of Science*, 64 (PSA 1996: Symposia Proceedings).

Howson, C. (1997b), "Error Probabilities in Error," pp. S185–S194 in L. Darden (ed.), *Philosophy of Science*, 64 (PSA 1996: Symposia Proceedings).

Kelly, K., Schulte, O., Juhl, C., (1997), "Learning Theory and the Philosophy of Science," *Philosophy of Science*, 64: 245–67.

Laudan, L. (1997) "How about Bust? Factoring Explanatory Power Back into Theory Evaluation," *Philosophy of Science*, 64(2): 306–16.

Mayo, D.G. (1997a), "Duhem's Problem, the Bayesian Way, and Error Statistics, or "What's Belief Got to Do with It?" 64: 222–44.

Mayo, D.G. (1997b), "Error Statistics and Learning From Error: Making a Virtue of Necessity," pp. S195–S212 in L. Darden (ed.), *Philosophy of Science*, 64 (PSA 1996: Symposia Proceedings).

Mayo, D.G. (2000a), "Experimental Practice and an Error Statistical Account of Evidence," pp. S193–S207 in D. Howard (ed.), *Philosophy of Science*, 67 (Symposia Proceedings).

Mayo, D.G. (2000b), "Models of Error and the Limits of Experimental Testing," pp. 317–44 in M. Carrier, G. Massey and L. Ruetsche (eds.), *Science at Century's End, Philosophical Questions on the Progress and Limits of Science*, University of Pittsburgh Press, Pittsburgh.

Woodward, J. (1997), "Data, Phenomena, and Reliability," pp. S163–S179 in L. Darden (ed.), *Philosophy of Science*. 64 (PSA 1996: Symposia Proceedings).