THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Daniel Shapiro; *Treasurer*, Barbara Gimbel; *Other Trustees*: Lee Bollinger, Leigh S. Cauman, Kent Greenawalt, Michael J. Mooney, Lynn Nesbit.

All communications to the Editors and Trustees and all manuscripts may be sent to John Smylie, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.

You may also visit our website at: www.journalofphilosophy.org

THE JOURNAL OF PHILOSOPHY

2004

SUBSCRIPTIONS (12 issues)

Individuals	$35.00
Libraries and Institutions	$75.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are required.

Published monthly as of January 1977; typeset and printed by Capital City Press, Montpelier, VT.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY 10027. (212) 866-1742

POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices.

POSTMASTER: Send address changes to the *Journal of Philosophy* at MC 4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027.

THE JOURNAL OF PHILOSOPHY

VOLUME CI, NO. 5, MAY 2004

LIKELIHOOD, MODEL SELECTION, AND THE DUHEM-QUINE PROBLEM*

The Duhem-Quine problem is usually formulated *deductively* with a choice described *dichotomously*. When the conjunction of a hypothesis (H) and an auxiliary assumption (A) entails an observational prediction (O) that fails to come true, should one reject H or reject A? A more general formulation would be to ask what one should say when the conjunction (H & A) confers some probability on O, and instead of considering the two choices just mentioned, the problem would be to evaluate judgments that are a matter of degree. For example, if the observational outcome disconfirms the conjunction (H & A), what determines whether and how much each conjunct is disconfirmed? Indeed, the negative cast of this question can be discarded by generalizing further: How does the disconfirmation or confirmation of the conjunction affect the disconfirmation or confirmation of the conjuncts?[1] The epistemological holism associated with Pierre Duhem[2] and W.V. Quine[3] denies that evidence bearing on (H & A) can have an impact on H that differs from the impact it has on A. This holism can take two forms. *Nondistributive holism* asserts that only whole conjunctions are confirmed and disconfirmed, never their constituent conjuncts; *distributive holism* concedes that evidence

* My thanks to Julian Barbour, Martin Barrett, Richard Creath, John Earman, Ellery Eells, Branden Fitelson, Malcolm Forster, Michael Friedman, Alan Hájek, Daniel Hausman, Margaret Moore, John Norton, Michael Stölzner, Peter Turney, and the editors of this JOURNAL for useful suggestions.

[1] If the original Duhem-Quine problem concerns a question about acceptance/rejection, how is the problem described here concerning confirmation/disconfirmation related to that original problem? If decisions about acceptance and rejection need to include an evaluation of the evidence at hand, the second problem is part of the first.

[2] *The Aim and Structure of Physical Theory* (Princeton: University Press, 1954).

[3] "Two Dogmas of Empiricism," in *From a Logical Point of View* (Cambridge: Harvard, 1953), pp. 20–46, and *Philosophy of Logic* (Englewood Cliffs, NJ: Prentice-Hall, 1970).

0022-362X/04/0105/221–41

bearing on the conjunction can have an impact on a conjunct, but insists that the effect on one conjunct must be the same as the effect on the other.[4] Holists grant that hypotheses and auxiliary assumptions are often treated differently when predictions fail, but claim that it is *nonevidential considerations*, such as simplicity or conservatism, that does the work.[5] To refute holism, the challenge is to show how *the evidence* can have an effect on hypotheses that differs from its effect on auxiliary assumptions.

Previous attempts to bring probabilistic tools to bear on the Duhem-Quine problem have mainly been Bayesian.[6] It is intrinsic to this approach that one must discuss $\Pr(H|A)$, $\Pr(A|H)$, and the probability of the observations conditional not just on $(H \,\&\, A)$ but on $(H \,\&\, \text{not}A)$ and on $(\text{not}H \,\&\, A)$.[7] It is not essential to assign point values

[4] Distributive holists may assert that the effects on H and A are *qualitatively* the same (that is, that both are confirmed or both are disconfirmed) or, more ambitiously, that the effects are *quantitatively* the same (that is, that the degree of confirmation of H is identical with the degree of confirmation of A). For more on this taxonomy of holisms, see my "Quine's Two Dogmas," *Proceedings of the Aristotelian Society*, LXXIV (2000): 237–80.

[5] I do not concede that simplicity *is* always an extra-evidential consideration; the point here is that this is what holists happen to believe. For discussion, see my *Reconstructing the Past: Parsimony, Evolution, and Inference* (Cambridge: MIT, 1988), and my "Instrumentalism, Parsimony, and the Akaike Framework," *Philosophy of Science*, LXIX (2002): S112–S123.

[6] See, for example, J. Dorling, "Bayesian Personalism, the Methodology of Scientific Research Programs, and Duhem's Problem," *Studies in the History and Philosophy of Science*, X (1979): 177–87, Colin Howson and Peter Urbach, *Scientific Reasoning: The Bayesian Approach*, (La Salle, IL: Open Court, 1989), John Earman, *Bayes or Bust?* (Cambridge: MIT, 1992), and Michael Strevens, "A Bayesian Treatment of Auxiliary Hypotheses," *British Journal for the Philosophy of Science*, LII (2001): 515–37. For a nonBayesian treatment, see Deborah Mayo's *Error and the Growth of Experimental Knowledge* (Chicago: University Press, 1996), which analyzes the Duhem-Quine problem within the context of frequentist statistics.

[7] As a simple example of a Bayesian analysis, let us define the degree of confirmation that X receives from Y, $c[X,Y]$, as the ratio $\Pr(X|Y)/\Pr(X)$. So defined, $c[X,Y] > 1$ when Y positively confirms X and $c[X,Y] < 1$ when Y disconfirms X. By Bayes's theorem, this ratio equals $\Pr(Y|X)/\Pr(Y)$; thus, $c[H, \text{not}O] > c[A, \text{not}O]$ if and only if $\Pr(\text{not}O|H) > \Pr(\text{not}O|A)$, which expands to

$$\Pr(\text{not}O\,|H \,\&\, A)\Pr(A\,|H) + \Pr(\text{not}O|\ H \,\&\, \text{not}A)\Pr(\text{not}A\,|H) >$$
$$\Pr(\text{not}O\,|H \,\&\, A)\Pr(H\,|A) + \Pr(\text{not}O|\ \text{not}H \,\&\, A)\Pr(\text{not}H\,|A).$$

Note the occurrence of $\Pr(A|H)$ and of $\Pr(H|A)$ in this expression. If we assume that H and A are probabilistically independent, the inequality reduces to

$$\Pr(\text{not}O\,|H \,\&\, A)\Pr(A) + \Pr(\text{not}O|\ H \,\&\, \text{not}A)\Pr(\text{not}A) >$$
$$\Pr(\text{not}O\,|H \,\&\, A)\Pr(H) + \Pr(\text{not}O|\ \text{not}H \,\&\, A)\Pr(\text{not}H).$$

Notice the prior probabilities of A and of H. I do not mean to beg questions here about the proper definition of degree of confirmation, on which see Branden Fitelson, "The Plurality of Bayesian Measures of Confirmation and the Problem of Measure Sensitivity," *Philosophy of Science*, LXVI (1999): S362–S378. The point is just to identify the kinds of quantities that a Bayesian analysis must evaluate.

to these quantities; the formal treatments require only that value ranges or inequalities among these quantities be provided. The problem with this approach is that these quantities are often difficult to interpret *objectively*. What is the probability of Newton's theory if there are seven planets? What is the probability of there being seven planets, if Newton's theory is true? And what is the probability that the orbit of Uranus will have a certain shape, if Newton's theory is false and there are seven planets? It does no good to treat these probabilities as subjective degrees of belief. This is unsatisfactory because a subjective interpretation has the consequence that one's "solution" to the problem lacks normative force—one has offered no reason to think that disconfirmation *should* be assigned more to one conjunct than to the other.[8] In other words, Bayesianism in the context of the Duhem-Quine problem encounters the same limitations that Bayesianism often confronts in other settings.

In what follows I will discuss an example of the Duhem-Quine problem in which $\Pr(H|A)$, $\Pr(A|H)$, and $\Pr(O|\pm H \,\&\, \pm A)$ (where H is the hypothesis, A the auxiliary assumptions, and O the observational prediction) can be construed objectively; however, only some of those quantities are relevant to the analysis that I provide. The example involves medical diagnosis. The goal is to test the hypothesis that someone has tuberculosis; the auxiliary assumptions describe the error characteristics of the test procedure. Although it can make sense to talk about the objective probability that someone (randomly drawn from a given population) has tuberculosis and it also can make sense to talk about the objective probability that a test procedure has a certain set of error characteristics, neither of these quantities will enter into the analysis. The analysis proceeds entirely via *likelihoods*; what one needs to consider is just the probability of the observations conditional on four conjunctions of the form $(\pm H \,\&\, \pm A)$.[9] It is a special feature of the example that all four of these conjunctions are *simple statistical hypotheses* in the technical sense that each unambigu-

Another Bayesian approach would be to compare, not the *change* in probability that O induces in H with the change it induces in A, but the *absolute values* of $\Pr(H|O)$ and $\Pr(A|O)$. Bayes's Theorem entails that $\Pr(H|O) > \Pr(A|O)$ if and only if $\Pr(O|H)\Pr(H) > \Pr(O|A)\Pr(A)$. Prior probabilities occur in this expression, and $\Pr(H|A)$ and $\Pr(A|H)$ are involved as well, since $\Pr(O|H) = \Pr(O|\ H \,\&\, A)\Pr(A|H) + \Pr(O|\ H \,\&\, \text{not}A)\Pr(\text{not}A|\ H)$ and $\Pr(O|A) = \Pr(O|\ H \,\&\, A)\Pr(H|A) + \Pr(O|\ \text{not}H \,\&\, A)\Pr(\text{not}H|A)$.

[8] See Earman (*op. cit.*).

[9] In what follows, I use "likelihood" and "likely" in this technical sense—the likelihood of $(H \,\&\, A)$ with respect to the observation O, is $\Pr(O|\ H \,\&\, A)$, not $\Pr(H \,\&\, A|\ O)$.

In addition to the Madison company I have just described, suppose there is a company in Middleton that has been involved in the same project. They also want to develop a tuberculosis test kit, so they also try out their procedure on 1000 people whom they know have tuberculosis and 1000 people whom they know do not. The data they obtain on their kit are given in Table 3. They then use maximum likelihood estimation to estimate the error probabilities of their test:

(Middleton) $\Pr(+ \text{ test result} \mid S \text{ has tuberculosis}) = 990/1000$
$\Pr(+ \text{ test result} \mid S \text{ has no tuberculosis}) = 5/1000.$

Notice that the Middleton device is inferred to have slightly larger error probabilities, both positive and negative, than the Madison test. Still, it is a pretty good test. If someone has a positive result on the Middleton test, the likelihood ratio of the two hypotheses H_1 (S has tuberculosis) and H_2 (S does not have tuberculosis) is 990/5 favoring H_1, and a negative outcome engenders a likelihood ratio of 995/10 favoring H_2.

TABLE 3

MIDDLETON	1000 with Tuberculosis	1000 with no Tuberculosis
+ test result	990	5
− test result	10	995

We now can return to the original problem of finding out whether Newman has tuberculosis. I introduced the two tuberculosis tests to give this problem a Duhemian twist. Duhem emphasized that physical theories do not entail observational predictions all by themselves, but do so only when conjoined with auxiliary assumptions. Duhem's insight is preserved in the example at hand, even though the relationships are probabilistic, not deductive. Suppose we give Newman a tuberculosis test and obtain a positive result. The probability of obtaining that result depends both on whether Newman has tuberculosis and on whether we used a test kit from Madison or one from Middleton. The four probabilities are represented in Table 4.

TABLE 4

		Possible Auxiliary Assumptions		
		A_1: Madison		A_2: Middleton
Hypotheses	H_1: Newman has tuberculosis	997/1000	>	990/1000
		∨		∨
	H_2: Newman does not have tuberculosis	2/1000	<	5/1000

Notice that there is a *qualitative asymmetry* between what the observational outcome says about the hypotheses H_1 and H_2 and what it says about the auxiliary assumptions A_1 and A_2. Newman's positive test result renders H_1 more likely than H_2, regardless of whether A_1 or A_2 is true. However, whether A_1 is more likely than A_2 depends on which of the hypotheses is true, and this, I am assuming, is something we do not already know. Of course, if we *do* already know whether Newman has tuberculosis, then the observed test result does provide information about whether the test kit came from Madison or from Middleton. However, the information provided is exceedingly modest. If Newman has tuberculosis, the positive test outcome slightly favors A_1 over A_2; the likelihood ratio here is only 997/990. Similarly, if Newman does not have tuberculosis, then the positive result favors A_2 over A_1, with a likelihood ratio of 5/2. On the other hand, if we not only do not know whether Newman has tuberculosis but cannot even assign a probability to this being the case, the test result tells us nothing about the provenance of the test kit.

If the data and the ensuing maximum likelihood estimates of error probabilities from either Madison or Middleton had been different, it could easily have turned out that there is no *qualitative* asymmetry between the observation's impact on the hypotheses and its impact on the auxiliary assumptions. That is, it is possible for the observation to provide information about both. However, this does not mean that the *amount* of information provided must be the same; there still can be a *quantitative* asymmetry, even if there is no qualitative asymmetry. By changing the probability in the lower-right cell in Table 4, we obtain Table 5. Now Newman's positive test result favors H_1 over H_2, regardless of which auxiliary assumption is true, and it also favors A_1 over A_2, regardless of which hypothesis is true. However, the observation provides much more information about whether Newman has tuberculosis than it does about whether the test kit came from Madison or Middleton. With respect to the hypotheses, the ratio is either 997/2 or 990/1, depending on which test procedure was used. With respect to the auxiliary assumptions, the ratio is either 997/990 or 2/1, depending on whether Newman has tuberculosis. Note that 997/2 and 990/1 are both much bigger than 997/990 and 2/1. In this example, the observation provides more information about the hypotheses than it does about the auxiliary assumptions. Of course, the reverse situation can also obtain and it also is possible for the situation to be perfectly symmetrical; two examples of symmetry will be discussed at the end of the paper.

TABLE 5

		Possible Auxiliary Assumptions		
		A_1: Madison		A_2: Middleton
Hypotheses	H_1: Newman has tuberculosis	997/1000	>	990/1000
		v		v
	H_2: Newman does not have tuberculosis	2/1000	>	1/1000

When a conjunction (H & A) makes a prediction that neither conjunct makes on its own, epistemological holism says that it is *never possible* for the outcome to have an evidential significance for H that differs from the significance it has for A. The generality of this thesis means that just one counterexample is enough to refute it. I claim that the example just described performs that function. Let H be the hypothesis that Newman does not have tuberculosis and let A be the hypothesis that one is using the Madison test procedure. The conjunction (H & A) predicts that the test result will be negative in the sense that it confers on that outcome a probability of 998/1000. But suppose the test comes out positive. To see what this outcome means for H and what it means for A, we need to know what the alternatives are to each. With the alternatives as described, the test result can have a bearing on the hypotheses that differs fundamentally from the bearing it has on the auxiliary assumptions. Both qualitative and quantitative asymmetries are possible. Epistemological holism is false.

II. TWEAKING THE EXAMPLE

The example just described, in which you do not know beforehand whether Newman has tuberculosis and also do not know which tuberculosis test kit you are using, is somewhat artificial. Scientists typically know the provenance of the test kits they use as well as their estimated error probabilities. But even in this more realistic setting, there still is room for Duhemian puzzlement. Suppose we *know* we are using the Madison test kit. However, we recognize that the error probabilities associated with this test kit are merely estimates—we have no certainty that the estimated values are exactly right. Again we give the test to Newman and again obtain a positive result. Does that outcome provide information about whether Newman has tuberculosis and does it also provide information about the test procedure's error characteristics? If so, does the outcome provide more information about one of these than it does about the other?

It might seem intuitive to say that Newman's test outcome provides zero information about the error characteristics of the test procedure. After all, Newman is quite unlike the 2000 subjects who were used to calibrate the test; we have no independent knowledge as to whether he has the disease. Of course, if we knew that he *probably* has the disease, or that he *probably* does not, that would tell us whether his positive test result is *probably* a true positive or is *probably* a false positive, and that would lead us to modify slightly our estimates of the test's error characteristics. But suppose we do not know even that. How, then, can the test result provide any information at all about the test's error characteristics?

Newman's test outcome could be a false positive or it could be a true positive. Let us consider these possibilities in turn. If Newman's test result is a true positive, we should add this result to the 2000 individuals already studied and change our estimate of the test's error characteristics to

$$\Pr(+ \text{ test result} \mid S \text{ has tuberculosis}) = 998/1001$$
$$\Pr(+ \text{ test result} \mid S \text{ has no tuberculosis}) = 2/1000.$$

On the other hand, if Newman's test result is a false positive, we should revise our estimate of the test's error characteristics as follows:

$$\Pr(+ \text{ test result} \mid S \text{ has tuberculosis}) = 997/1000$$
$$\Pr(+ \text{ test result} \mid S \text{ has no tuberculosis}) = 3/1001.$$

Of course, we do not know whether Newman's result is a false positive or a true positive, so we do not know which pair of estimates we should use to characterize the procedure's error characteristics. However, this uncertainty does not prevent us from formulating a pair of *conditional estimates*:

(New Madison) If Newman has tuberculosis, then $\Pr(+ \text{ test result} \mid S \text{ has tuberculosis}) = 998/1001$ and $\Pr(+ \text{ test result} \mid S \text{ has no tuberculosis}) = 2/1000$.

If Newman does not have tuberculosis, then $\Pr(+ \text{ test result} \mid S \text{ has tuberculosis}) = 997/1000$ and $\Pr(+ \text{ test result} \mid S \text{ has no tuberculosis}) = 3/1001$.

We now can ask whether there is a difference in likelihood between the old estimates (Madison) or the new, conditional, estimates (New Madison) that were obtained by taking account of Newman's positive test result.

Table 6 summarizes the situation; cell entries represent the probability of Newman's positive test result, conditional on different combinations of hypotheses and auxiliary assumptions. Notice first that there is a *qualitative symmetry* between what the observation says about the hypotheses and what it says about the auxiliary assumptions. Newman's positive test result renders H_1 more likely than H_2, regardless of which auxiliary assumption is true, and the result also favors (New

Madison) over (Madison) regardless of which hypothesis is true.[16] However, there is a *quantitative asymmetry*. The observation is *very* informative about whether Newman has tuberculosis; 997/2 and 998/3 are both large. In contrast, the observation is only modestly informative about the choice between the auxiliary assumptions; the ratios are approximately 998/997 and 3/2, and these are both rather small. Why does Newman's test result have such a negligible impact on the estimates of the test's error characteristics? The reason is that Newman is just one person out of 2001. Had we initially estimated the error probabilities by using just 200 subjects, or 20, or 2, Newman would have mattered more.

TABLE 6

		Possible Auxiliary Assumptions		
		Madison		New Madison
Hypotheses	H_1: Newman has tuberculosis	997/1000	<	998/1001
		∨		∨
	H_2: Newman does not have tuberculosis	2/1000	<	3/1001

It may seem odd that I even consider (New Madison). If estimates of the error characteristics of a test procedure must be based solely on frequency data, then speculations about what our maximum likelihood estimates would be if we knew whether Newman has tuberculosis are irrelevant. This sensible attitude flies in the face of epistemological holism—it entails that Newman's test outcome provides considerable evidence about whether he has tuberculosis and zero information about the error characteristics of the test procedure. If this were correct, there would be a qualitative as well as a quantitative asymmetry. The analysis in which (Madison) and (New Madison) are compared comes close to this result, but does not coincide with it exactly. I argued that the evidence *slightly* favors (New Madison) over (Madison), not that the observation is literally *informationless*. In terms of the larger picture of seeing what is wrong with epistemological holism, this difference does not matter. But in terms of the specifics of likelihood reasoning, it does.

III. SIGNIFICANCE OF THE TWEAKED EXAMPLE

This last example illustrates a very general fact about the calibration of measurement instruments and the validation of test procedures in

[16] (Madison) has a lower likelihood than (New Madison) in each row, since $a/b < (a+1)/(b+1)$, if $0<a<b$.

science. The typical situation is that the error characteristics of a test procedure are first ascertained and then the procedure is applied to new individuals. One usually does not already know whether these new individuals have the condition being tested (otherwise, why apply the test?); indeed, one often does not even know whether the new individuals *probably* have the condition. Many of us have opinions about the approximate frequency of tuberculosis in this or that population; if we were prepared to assume that Newman was drawn at random from such a population, we would be entitled to talk about the prior probability that he has tuberculosis. Many scientific tests are not like this. Galileo gauged the reliability of his telescope by training it on various *terrestrial* objects. He used it to identify the flags on ships coming over the horizon and the inscriptions on distant buildings; in all these cases it was possible to determine *independently* whether the reports were correct.[17] Galileo then looked through his telescope at *Jupiter*; his observations provided strong evidence that Jupiter has moons, but little or no information about the telescope's error characteristics. Understanding this asymmetry does not require that one assign a prior probability to Jupiter's having moons. This is fortunate, since Galileo was in no position to assign an objective prior probability to that proposition.[18]

There is a sense in which the likelihood analysis of the tuberculosis example is nonBayesian, but this is not because likelihood is an idea that Bayesianism abhors. On the contrary—likelihood is a fundamental quantity in Bayes's theorem. What I mean is that the analysis does not use the full-blown resources that Bayesianism assumes are available. First, prior and posterior probabilities play no role. Second, a likelihoodist will be happy to compare the likelihoods of two simple statistical hypotheses (S_1 and S_2), but often is loath to compare the likelihoods of a simple hypothesis (S_1) and its negation (notS_1) when

[17] See Philip Kitcher's *The Advancement of Science: Science without Legend, Objectivity without Illusions* (New York: Oxford, 1993), pp. 228–33, and his "Real Realism: The Galilean Strategy," *Philosophical Review*, CX (2001): 151–98.

[18] Galileo estimated the error characteristics of his telescope by using it in problems that involved relatively small terrestrial distances; he then applied this detection device to an astronomical object that was much farther away. There certainly was room to wonder, at the time, how trustworthy this bold extrapolation was. My point here is not to comment on the legitimacy of Galileo's inference, but to note how often scientists use the protocol I described in connection with the tuberculosis test. The question of how the behavior of the measuring device should be *parameterized* (for example, a single set of error characteristics for sightings of all objects, or two such sets—one for objects that are near and another for objects that are far away), as opposed to the question of how values for parameters should be *estimated*, will be discussed later.

that negation is composite. Suppose notS_1 is equivalent to a disjunction of simple hypotheses (S_2 or S_3 or...or S_n). If so, the likelihood of notS_1 will be a *weighted average* of the likelihoods of S_2, S_3,...S_n, where the weighting term has the form $\Pr(S_i \mid \text{not}S_1)$. This weighting term often lacks an objective interpretation. If Newton's theory is false, what is the probability of each of the theory's specific alternatives? Thus the problem with priors often recurs as a problem for likelihoods.[19] It is a very special property of the tuberculosis example that the two hypotheses considered are *exhaustive* (assuming that Newman exists) and the four conjunctions are simple.[20]

The relations of qualitative and quantitative asymmetry that I have described are purely formal, and therefore do not depend on one's interpretation of probability. Still, the question may be asked of what interpretation of probability I am using when I say that the likelihoods evaluated in the example concerning Newman's tuberculosis are "objective." Clearly, I cannot think of probabilities as subjective degrees of belief. But neither do I wish to endorse the objective interpretations—actual relative frequencies, hypothetical relative frequencies, propensities—now on offer. My preference is the *no-theory theory of probability*, which rejects the need for a reductive analysis of what probability statements mean. Probability is a theoretical quantity. It obeys the axioms of probability and it bears nondeductive inferential relations to observed relative frequencies. Probability, like other theoretical magnitudes, cannot be reduced to observations, nor does it need to be.[21]

This brings us to the question of how general the treatment provided here of Newman's tuberculosis test can be said to be. Does the structure of this quotidian example apply to all situations in which the Duhem-Quine problem arises? There are two types of situation in which it does not; I will describe one of them now and postpone the other until the next section. Scientists sometimes react to the predictive failure of a conjunction by formulating an alternative to

[19] See Richard W. Miller, *Fact and Method* (Princeton: University Press, 1999), and my "Bayesianism: Its Scope and Limits," in Richard Swinburne, ed., *Bayes's Theorem* (New York: Oxford, 2002), pp. 21–38.

[20] I have emphasized how the likelihood approach allows one to avoid considering prior probabilities. However, there is a quantity that the likelihood approach requires one to consider, that Bayesian treatments (of the kind described in footnote 7) do not. When all four conjunctions of the form ($\pm H$ & $\pm A$) are simple statistical hypotheses, the likelihood approach will consider the quantity $\Pr(O \mid \text{not}H \,\&\, \text{not}A)$.

[21] See Isaac Levi and Sidney Morgenbesser, "Belief and Disposition," *American Philosophical Quarterly*, I (1964): 221–32, and Levi, *Gambling with Truth* (New York: Knopf, 1967).

one conjunct without bothering to formulate an alternative to the other. For example, when John Crouch Adams and Jean Joseph Le Verrier tried to account for the fact that the conjunction of Newton's theory and the assumption that there are seven planets generates an inaccurate prediction of Uranus's orbit, they did not try to invent an alternative to Newton's theory.[22] Rather, they set their minds to constructing a specific alternative to the auxiliary assumption, and the result was the prediction and confirmation of a new planet, which we now call Neptune.[23] In this case, the comparison was between two conjunctions—(Newton & seven planets) and (Newton & eight planets)—first using the old observations of Uranus's orbit, and then assembling new observations drawn from pointing telescopes in the right direction.[24] This one-sided response to observational anomaly also occurs when the auxiliary assumptions include propositions of pure mathematics. It is entirely customary for alternative empirical hypotheses to draw on the same body of pure mathematics.[25] In such cases, the observations are used to test (H_1 & A) against (H_2 & A), but are not used to test the auxiliary assumption A against an alternative, since none was formulated in the first place. Cases of this sort involve epistemological asymmetry "by default," so to speak; they therefore differ from the case of Newman and his tuberculosis test in which alternative hypotheses and alternative auxiliary assumptions are both on the table for consideration.

Even so, the likelihood concept throws light on cases in which scientists decline to construct a new theory (or decline to construct a new set of auxiliary assumptions). They often do so because they

[22] Although Newtonian theory does not use the concept of probability, it still is a mistake to think of the theory plus auxiliary assumptions as *deductively entailing* a prediction about what one should observe concerning Uranus's orbit. The reason is that the observation procedures used are subject to error, and these error characteristics need to be modeled probabilistically, just as was true for Newman's tuberculosis test.

[23] W.M. Smart, "John Crouch Adams and the Discovery of Neptune," *Occasional Notes of the Royal Astronomical Society*, XI (1947): 33–88.

[24] Adams and Le Verrier were able to *accommodate* the observed orbit of Uranus within the Newtonian framework by postulating an eighth planet, but did Uranus's orbit *confirm* the hypothesis that there is an eighth planet? It is tempting to answer this question in the negative and to insist that it was only the observation of *Neptune* that provided evidence. The broad epistemological question at issue here is whether using an observation to construct a hypothesis means that the observation fails to provide evidence for the hypothesis. See Christopher Hitchcock and Elliott Sober, "Prediction, Accommodation, and the Problem of Overfitting," *British Journal for the Philosophy of Science*, LV (2004): 1–34, for discussion.

[25] See my "Indispensability and Mathematics," *Philosophical Review*, CII (1993): 35–57.

expect that the new construction would have low likelihood, relative to *all* the evidence. Newtonian theory exhibited excellent fit to lots of other data; it would have been a very tall order to construct an alternative theory that does a better job of handling the data on Uranus while still having high likelihood relative to these other observations. The assumption that there are just seven planets was much less enmeshed with other data sets, so it made sense for Adams and Le Verrier to have held on to Newtonian theory while attempting to revise the auxiliary assumption about the number of planets.[26] As is well known, the same strategy met with failure when applied to the problem of explaining Mercury's orbit. Einstein's general theory of relativity was able to solve the problem precisely because it fit the data that Newton's theory also fit, while fitting the data on Mercury better. The old auxiliary assumption, that there is no planet between Mercury and the Sun, turned out to be right all along, notwithstanding the fact that Le Verrier and others considered the possibility that an as-yet unobserved planet (Vulcan) is perturbing Mercury's orbit.[27]

Must a solution to the Duhem-Quine problem give scientists advice on whether they should formulate an alternative to the hypothesis H or an alternative to the auxiliary assumptions A when the conjunction $(H \& A)$ generates a failed prediction? I do not think so. Epistemology does not have the burden of predicting that Uranus's orbit should be handled in one way while Mercury's should be handled in another. It took an Einstein (namely, *the* Einstein) to discover this; there was nothing in the anomalous data and their relation to Newtonian theory that indicated what the facts would turn out to be. It is perhaps more reasonable for philosophy in this instance to remain on one side of the divide between the *context of discovery* and the *context of justification*.[28] The likelihood analysis describes how alternatives should be compared once they are formulated, not whether they are worth constructing in the first place.[29]

[26] This point concerning "enmeshment with other data sets" is an appeal to observational evidence, not to the extra-evidential considerations that holists think are essential.

[27] Whereas the Adams/Le Verrier approach to the orbit of Uranus involved "asymmetry by default," this was not the case with respect to later discussion of the anomalous perihelion of Mercury, in that modifications of the auxiliary assumptions and of Newtonian theory were both developed. See N. Roseveare, *Mercury's Perihelion from Le Verrier to Einstein* (New York: Oxford, 1982), and John Earman and Michel Janssen, "Einstein's Explanation of the Motion of Mercury's Perihelion," in Earman, Janssen, and John D. Norton, eds., *The Attraction of Gravitation: New Studies in the History of General Relativity* (Boston: Birkhäuser, 1993), pp. 129–72.

[28] See Hans Reichenbach, *Experience and Prediction* (Chicago: University Press, 1938).

[29] The distinction between *context of discovery* and *context of justification* is distinct from the distinction between *rules for accepting and rejecting hypotheses* and *rules for saying which hypotheses are better supported by the evidence.* I draw the latter distinction *within* the context of justification.

In discussing Newman's test result, the hypotheses considered were exhaustive (assuming that Newman exists), but the range of alternative auxiliary assumptions was narrowly circumscribed. In both versions of the problem, I assumed that applications of a tuberculosis test are independent and identically distributed; the same pair of error probabilities applies each time someone takes a test and the error probabilities that apply when one person takes the test are independent of what the outcomes happened to be when others did so. But surely this too is a background assumption that is up for grabs; it is not immune from revision in the context of the Duhem-Quine problem. The different auxiliary assumptions we considered disagree about the *values* of two parameters, but they agree on how the problem should be *parameterized.* It is perfectly legitimate to consider alternative *models* of how the test procedure works, where different models parameterize the problem in different ways. This takes us to our next topic.

IV. MODEL SELECTION

Although the likelihood approach is enough to show that epistemological holism is false, I do not claim that it is able to handle all instances of the Duhem-Quine problem. The main limitation concerns the treatment of composite (nonsimple) statistical hypotheses whose likelihoods cannot be interpreted objectively.[30] For example, consider the hypothesis that two physical quantities—say, the temperature and pressure in a closed chamber of gas—are related linearly. Although a specific straight-line hypothesis (for example, $P = 4 + 3T + U$, where U is an error distribution) confers a probability on a given value for pressure, given a value for temperature, it is more puzzling how one should think about the likelihood of the weaker claim that the relationship is linear (that is, that there exist values of a and b such that $P = a + bT + U$). This is because the likelihood of the hypothesis of linearity (LIN) is a weighted average of the likelihoods of all possible straight lines $(L_1, L_2, \ldots)$:[31]

$$\Pr(\text{Data} \mid \text{LIN}) = \sum_i \Pr(\text{Data} \mid L_i) \Pr(L_i \mid \text{LIN}).$$

If one has no objective basis for saying how probable this or that straight line is, conditional on (LIN), one will not be able to treat the likelihood of (LIN) as an objective quantity.

[30] See M. Forster, "Bayes and Bust: Simplicity as a Problem for a Probabilist's Approach to Confirmation," *British Journal for the Philosophy of Science,* XLVI (1995): 399–429, and my "Bayesianism: Its Scope and Limits" (*op. cit.*).

[31] As an expository convenience, I represent the average likelihood of (LIN) as a discrete summation, not as a continuous integration.

The term "model" is used in the statistics literature on model selection to refer to hypotheses that contain at least one adjustable parameter. The hypothesis of linearity is a model in this sense, but specific straight-line hypotheses are not. Perhaps the most widely used model selection criterion is the Akaike information criterion (AIC), proposed by Akaike (*op. cit.*).[32] AIC is nonBayesian; the goal is not to compute the probability of a model or its likelihood. Rather, AIC aims to provide an estimate of the model's *predictive accuracy*.[33] But how can (LIN), as opposed to a specific straight-line hypothesis, provide a prediction (either accurate or inaccurate) about the gas's temperature when the gas is raised to a particular temperature? The answer is that the model must first be fitted to a set of old data; the parameters a and b are estimated from that data, using the method of maximum likelihood estimation. By substituting these estimates for the adjustable parameters, (LIN) is replaced with the specific straight-line hypothesis L(LIN); this is the specific straight line that renders the old data maximally probable. One then draws a new data set from the chamber of gas and determines how well L(LIN) predicts this new data. The average performance of (LIN) in this prediction task—first being fitted to old data, then seeing how well the fitted model predicts new data—defines the model's predictive accuracy. Estimating the predictive accuracy of models is the goal; the next question is how we should go about attaining that goal. If we have just one data set at hand, how are we to use this evidence to judge how predictively accurate a model is?

An important lesson that scientists absorb from working with models is that making a model too complex will reduce its predictive accuracy. It is easy to get a model to fit the available data by making it sufficiently complex, but the price is often that the fitted model does a poor job predicting new data. This is not a brute fact in the life experience of scientists; rather, Akaike provided a mathematical framework that helps explain why overfitting tends to reduce predictive accuracy. Using some very general assumptions, Akaike proved a result concerning how the predictive accuracy of a model can be estimated. He showed that an unbiased estimate of a model's predictive accuracy can be obtained by looking at two of its properties—how well it fits

[32] See also Y. Sakamoto, M. Ishiguro, and G. Kitagawa, *Akaike Information Criterion Statistics* (New York: Springer, 1986); and Kenneth P. Burnham and David R. Anderson, *Model Selection and Inference: A Practical Information-Theoretic Approach* (New York: Springer, 1998).

[33] See Forster and Sober, "How to Tell When Simpler, More Unified, or Less *Ad Hoc* Theories Will Provide More Accurate Predictions," *British Journal for the Philosophy of Science*, XLV (1994): 1–36.

the data at hand, and how complex it is (where complexity is measured by the number of adjustable parameters the model contains). Akaike's theorem is the basis for AIC, which assigns a score to a model that reflects both its fit-to-data and its simplicity.[34] By comparing the AIC scores of different models, one can estimate which models will make more accurate predictions.

As an example of a model selection problem in which models can be viewed as conjunctions (and so the Duhem-Quine problem can arise), let us consider the task of phylogenetic inference.[35] The goal is to evaluate the plausibility of different phylogenetic trees. Are human beings more closely related to chimps than they are to gorillas, or is the tree topology something different? There are three bifurcating trees that need to be considered—$(HC)\,G$, $H(CG)$, and $(HG)\,C$—but none of these confers a probability on the data until some model of the evolutionary process is provided. There are several process models to consider. In the context of molecular evolution, the simplest model is that of T. Jukes and C. Cantor,[36] which assumes that each of the four nucleotides has the same probability per unit time of changing into any of the others. More complex models, such as the one due to M. Kimura,[37] allow different changes to have different probabilities.[38] When a tree topology is conjoined with a process model, its adjustable parameters may be estimated from the data, and the AIC score of the conjunction may then be computed, as depicted in Table 7. Nothing prevents these cell entries from exhibiting the same asymmetries we saw in the likelihood analysis of Newman's tuberculosis test. The data might favor one tree topology over the other, regardless of which process model is used, but fail to provide any robust

[34] Does AIC play into the hands of holism by invoking simplicity? I would say not, in that the justification for AIC depends on empirical assumptions (though ones of great generality). In addition, I will be using AIC only as an example of a model selection criterion; *cross-validation* is another such criterion, and it involves no appeal to simplicity. As it happens, take-one-out cross-validation is asymptotically equivalent with AIC. See M. Stone's "Cross-validatory Choice and Assessment of Statistical Predictions (with Discussion)," *Journal of the Royal Statistical Society*, B XXXVI (1974): 111–47, and his "An Asymptotic Equivalence of Choice of Model by Cross-validation and Akaike's Criterion," *Journal of the Royal Statistical Society*, B XXXIX (1977): 44–47.

[35] See my "The Contest between Likelihood and Parsimony," *Systematic Biology* (2004, forthcoming).

[36] "Evolution of Protein Molecules," in H. Munro, ed., *Mammalian Protein Metabolism* (New York: Academic, 1969), pp. 21–132.

[37] "A Simple Method for Estimating Evolutionary Rates of Base Substitutions through Comparative Studies of Nucleotide Sequences," *Journal of Molecular Evolution*, XVI (1980): 111–20.

[38] For a survey of the different models now on offer, see Roderic Page and Edward Holmes, *Molecular Evolution: Phylogenetic Approach* (Malden, MA: Blackwell, 1998), pp. 148–62.

indication of which process model is better. It also is possible for the data to provide information about both tree topology and process model while providing more information about one than it does about the other. There is nothing special about AIC in this regard; all model selection criteria can generate both qualitative and quantitative asymmetries. This shows that the solution to the Duhem-Quine problem that I am proposing does not require a commitment to likelihood as the one true way to interpret evidence.

TABLE 7

		Process Models	
		Jukes-Cantor	Kimura
Tree topologies	$(HC)G$	AIC_1	AIC_2
	$H(CG)$	AIC_3	AIC_4

V. CONCLUDING COMMENTS

How is the likelihood analysis described here related to a deductivist formulation of the Duhem-Quine problem in which we recognize that there are three choices (and not just two)—we can reject H_1, reject A_1, or both? If the conjunction $(H_1 \,\&\, A_1)$ entails O, then $\Pr(\text{not}O \mid H_1 \,\&\, A_1) = 0$. The four likelihoods we need to consider, relative to the observation (not$O$), are shown in Table 8. Of course, the Law of Likelihood does not make recommendations about acceptance and rejection, but merely describes the differential support that the evidence provides. However, if acceptance and rejection require an evaluation of evidence, then the relationship of these four likelihoods is relevant to deciding what to accept and what to reject. If H_2 *dominates* H_1 (that is, if $p_1<p_3$ and $p_2<p_4$), then the evidence favors H_2 over H_1 regardless of which auxiliary assumption is true. If, in addition, A_2 dominates A_1 (that is, if $p_1<p_2$ and $p_3<p_4$), then $(H_2 \,\&\, A_2)$ is the conjunction with the highest likelihood, and so likelihood considerations point away from both H_1 and A_1. In this circumstance, the likelihood inequalities provide no qualitative asymmetry between the observation's impact on the hypotheses and its impact on the auxiliary assumptions, though a quantitative asymmetry may nonetheless obtain. If dominance holds in one direction but not the other, there is a qualitative asymmetry.

TABLE 8

		Possible Auxiliary Assumptions	
		A_1	A_2
Hypotheses	H_1	$p_1=0$	p_2
	H_2	p_3	p_4

In both the likelihood approach and the model selection approach, all four conjunctions of the form $(H_i \,\&\, A_j)$ $(i,j = 1,2)$ must be considered.[39] This brings out a further difference between these analyses and Quine's holistic epistemology. As noted earlier, the holist recognizes that people manage to decide which conjunct to blame when a conjunction generates a false prediction, but contends that this decision must rely on extra-evidential considerations, such as simplicity or conservatism. The problem is typically formulated in terms of what a person *believes*—if you believe the conjunction $(H_1 \,\&\, A_1)$ and then find that conjunction refuted by the evidence, which conjunct should you abandon? Quine recommends a policy of "minimum mutilation"—you should impose the smallest change in your web of belief that suffices to restore consistency with the observations, where changing a more "central" belief is regarded as a larger modification than changing a belief that is more "peripheral."[40] It follows that if you believe $(H_1 \,\&\, A_1)$, there can be no reason for you to abandon *both* conjuncts and embrace $(H_2 \,\&\, A_2)$, if a more conservative reformation to either $(H_1 \,\&\, A_2)$ or to $(H_2 \,\&\, A_1)$ would manage to restore consistency. It is interesting that the Law of Likelihood and model selection criteria place no premium on minimizing change in which propositions you believe. What you happen to believe plays no role at all; in fact, *you* do not enter into the analysis. The question is entirely about the relationship of propositions to data and has nothing to do with people and their affections or mobility. If $(H_2 \,\&\, A_2)$ has the highest likelihood or the best AIC score, so be it.[41]

The falsity of epistemological holism does not mean that it is *never* true; the point is that it is *often* untrue. Indeed, it is possible to describe a circumstance that can arise within a likelihood framework in which holistic intuitions are vindicated. Suppose that dominance fails, both with respect to the comparison of hypotheses and also with respect to the comparison of auxiliary assumptions. That is, let $p_1<p_2$, $p_3>p_4$,

[39] I hope it is clear that my focus on *four* conjunctions is an expository convenience. Both the Law of Likelihood and model selection criteria can address any number of conjunctions, and the asymmetries I have described can arise in that larger comparative context as well.

[40] *Philosophy of Logic*, p. 7; see also "Two Dogmas of Empiricism," p. 44.

[41] Since Quine's principle concerns change in what one believes, whereas the Law of Likelihood and model selection criteria do not provide rules of acceptance, there is no formal incompatibility here. However, when likelihood is placed in the context of a full Bayesian framework with prior probabilities, it can turn out that $(H_1 \,\&\, A_1)$ is the conjunction with the highest prior probability whereas $(H_2 \,\&\, A_2)$ is the one whose posterior probability is greatest. The imperative to find the conjunction that differs minimally from the one with the highest prior probability and that is logically consistent with the new evidence is not a principle of Bayesian epistemology.

$p_1 < p_3$, and $p_2 > p_4$ in Table 8. The cell entries in the table will then have two peaks (at H_1 & A_2 and at H_2 & A_1) and two valleys (at H_1 & A_1 and at H_2 & A_2). If the two peaks have equal likelihood, the observations will not discriminate between them, even though each peak is better supported than the conjunctions that occupy the two valleys. The history of epistemology is peppered with examples of this type. I will mention two; the first is Cartesian, while the second is due to Hans Reichenbach.[42] In both cases, (H_1 & A_1) predicts O, but not$O$ is what you observe:

not$O$: There seems to be a printed page in front of me.
H_1: There is a salami (and not a printed page) in front of me.
H_2: There is a printed page (and not a salami) in front of me.
A_1: My senses are functioning normally.
A_2: An evil demon is causing printed pages to look like salamis, and vice versa.

not$O$: My measurement device indicates that the triangle I have just measured has an angle sum that exceeds 180°.
H_1: Space has zero curvature.
H_2: Space has constant positive curvature.
A_1: There are no universal forces.
A_2: Universal forces are in operation.

Given the observational outcome not$O$, likelihood allows you to discriminate between conjunctions in the same row and conjunctions in the same column, but not between the anti-diagonal entries. The same situation can arise in the context of model selection, if the AIC scores in Table 7 have the pattern of peaks and valleys just described in connection with Table 8.

The likelihood approach views theory evaluation as inherently comparative. The Law of Likelihood describes what an observation says about *two* hypotheses, but is silent on the question of what, if anything, it says about a single hypothesis taken on its own. Likelihoodism elevates that omission to the level of principle; when a hypothesis is tested, it must be tested *against* alternatives.[43] In addition, there is no single alternative that counts as the uniquely correct alternative to consider. There are many specific alternatives, and the Law of Likelihood may be asked for its assessment of all these competitions. This marks an important difference between likelihoodism and Bayesianism. Bayesianism also has its comparative element, but there is just one alternative to a proposition that Bayesians need to consider; this is simply the proposition's negation. The comparative character of likelihood assessments (and of model selection criteria as well) is central to the analysis proposed here of the Duhem-Quine problem: *whether holism is right in what it says about the bearing of evidence on a conjunction crucially depends on what the alternative conjunctions are.* In many scientific contexts, the competing hypotheses and the competing auxiliary assumptions are such that the evidence has an impact on one conjunct that differs from the impact it has on the other. However, there are other discrimination problems in which the evidence cannot penetrate from an assessment of conjunctions to an assessment of conjuncts. Holism has its place, but in the context of scientific inquiry it is, by far, the exception and not the rule.

Quine famously opined that "any statement can be held true come what may, if we make drastic enough adjustments elsewhere in the system."[44] Quine's point was not just that it is *logically possible* to continue to believe a proposition in the face of apparently recalcitrant experience, but that *the evidence* does not say that this is a mistake. Decisions about retention or rejection are dictated by nonevidential considerations. I suspect that the allure of this form of holism derives from a hypothetico-deductive view of theory testing. If an observation O confirms a proposition P precisely when P entails O and O turns out to be true, and O disconfirms P precisely when P entails O and O turns out to be false, then epistemological holism is correct when the conjunction (H & A) entails O, but neither conjunct does. In this circumstance the observational outcome confirms or disconfirms the conjunction, but not the constituent conjuncts. However, it is abundantly clear that confirmation and disconfirmation can be mediated by *non*deductive relationships. In the examples discussed here concerning tuberculosis diagnosis and phylogenetic inference, the conjunctions considered were each consistent with the observational outcome without entailing it, but that does not mean that the evidence failed to discriminate among them. Evidence can do more than hypothetico-deductivism imagines. It is that extra power that undermines epistemological holism.

ELLIOTT SOBER

Stanford University
University of Wisconsin/Madison

[42] *The Philosophy of Space and Time* (New York: Dover, 1958).

[43] An obvious exception arises when a proposition entails an observational prediction that fails to come true. In this instance, there is no need to consider an alternative; *modus tollens* permits one to reject the proposition without further ado. However, in the very common case in which H, or the conjunction (H & A), confers a probability on O that is less than unity, the falsity of O cannot be interpreted until alternatives are considered. Likelihoodism therefore rejects "probabilistic *modus tollens*" (a.k.a. Fisherian significance testing), according to which a hypothesis should be rejected when an event occurs that the hypothesis says was very improbable. For criticisms of probabilistic *modus tollens*, see Hacking, Edwards, and Royall.

[44] "Two Dogmas of Empiricism," p. 43.