How about Bust? Factoring Explanatory Power Back into Theory Evaluation

Author(s): Larry Laudan

Source: *Philosophy of Science*, Jun., 1997, Vol. 64, No. 2 (Jun., 1997), pp. 306-316

Published by: The University of Chicago Press on behalf of the Philosophy of Science Association

Stable URL: https://www.jstor.org/stable/188310

How About Bust? Factoring Explanatory Power Back into Theory Evaluation*

Larry Laudan†
Department of Philosophy, National Autonomous University of Mexico

1. Introduction. The papers by Hellman and Mayo offer up a rich menu of problems and proposed solutions, so there is much here for a friendly critic to fasten on. In order to bring a modicum of focus to my commentary, I shall limit my remarks to the Duhem problem and its radiations in epistemology and methodology. Both Mayo and Hellman claim to have solutions to that hoary old problem and they tout these solutions as key indicators of the explanatory power of their respective technical epistemologies, whether Bayesian or Neyman/Pearsonian. Like Mayo, I shall be arguing that the Bayesian treatment of Duhem's problem is no solution at all; that, indeed, it fails to grapple with the core challenges posed by the purported ambiguities of falsification. My response to Mayo's more detailed, and I think more right-headed, treatment of the Duhem problem will be more complex. While I believe that she is moving in the right direction in many respects, I think that she fails to see one key dimension of the Duhemian conundrum. Indeed, she risks solving not Duhem's problem but quite a different one. I shall gently try to encourage her to steer her way back towards the central task.

For the last two decades, I have been arguing that, in the appraisal of theories and hypotheses, what does (and what should) principally matter to scientists is not so much whether those hypotheses are true or probable. What matters, rather, is the ability of theories to solve empirical problems—a feature that others might call a theory's explanatory or predictive power. As I read the essays by Hellman and Mayo, as well as other authors in those two traditions, I find a bland indifference to the issue of a theory's ability to account for many of the

*Received January 1997.

†Send reprint requests to the author, Calle de Tenaza, 23, San Javier, Guanajuato, GTO 36000, Mexico.

Philosophy of Science, 64 (June 1997) pp. 306–316. 0031-8248/97/6402-0006$2.00

phenomena falling in its domain. They respectively worry about the probability of an hypothesis or about its well-testedness; but neither seems concerned with how much an hypothesis explains nor how many problems it solves. On the contrary, their respective measure of epistemic acceptability run quite counter to considerations of broad scope. Because I think Duhem shared my concerns and because, in my view, the Duhem problem is ultimately as much about a theory's scope as it is about its correctness, I think it might be of some use to view the Hellman-Mayo debate about Duhem's problem through these spectacles.

Duhem did not think that theories were true or probable; he did not argue that the aim of science is to develop an account of nature consisting of theories which became more true through time. His was a classic form of instrumentalism: the aim of theories is, in the idiom, "to save the phenomena"—an aim about which he wrote one of his most penetrating books (Duhem 1969). Now, both for the astronomical tradition which bears that name and for Duhem himself, what it means to say that a theory saves the phenomena (roughly put) is that said theory manages to explain all the known, principal features of the objects under investigation by the theory in the domain in which the phenomena fall. *The demand to save the phenomena is a requirement of partial completeness.* It holds a theory epistemically liable when that theory fails to exhibit its ability to explain or predict the known "facts" about a particular area of investigation.

I dwell on this seemingly obvious point because, in our own time, the phrase "saving the phenomena" has sometimes taken on a quite different complexion. With van Fraassen (1980, for instance), a theory "saves the phenomena" just in case everything it entails about observable phenomena is true. This is both narrower and broader then the traditional conception. It is narrower—and this feature will be of prime interest to us—in that it permits a theory to qualify as "saving the phenomena" even if that theory fails to solve many problems in its domain. It is broader in that it insists that a theory saves the phenomena if and only if it deals with both the known and the unknown observable consequences; on the traditional conception, the focus was principally on a theory's ability to account for the *known* phenomena.

The central point is that Duhem's concern, in judging a theory, was not principally with whether everything the theory says about the observable domain is true (this is van Fraassen's preoccupation) but rather with the question whether the theory manages correctly to address all those phenomena for which we hold it responsible. Duhem was emphatic about the fact—stoutly denied in this Tarski and post-Tarski age—that a theory is to be judged not only by what it requires and forbids but also against that yardstick provided by the well-established phenomena in the theory's domain. For Duhem, unlike both the verificationists and the falsificationists, a theory's empirical content provides only a part of the facts relevant to that theory's appraisal. For him, evidence of a theory's incompleteness (in the specific sense of its failure to address well-established phenomena in its domain) was at least as troubling as evidence of its falsity.

I think that Duhem was on the right track in this emphasis on what we should expect an adequate theory or hypothesis to do. Issues of explanatory power or completeness are at least as important as questions of truth, probability, or falsity. Unfortunately, however, this demand finds little support within either a Bayesian or an error-statistical framework. On the contrary, what those traditions share—for all their other differences—is an implicit commitment to the idea that issues of explanatory scope have minimal or no epistemic standing.

With this background in mind, let us turn to the Duhem problem itself, and perhaps cast it in a rather different light.

2. The Stages in the Accommodation of an Anomaly, or Substituting Honest Toil for Bayesian Sloth. Let me begin by very briefly restating the problem, using what has become the canonical symbology. We have an hypothesis, H. Supposing, with Duhem, that testing is holistic, we imagine that H can be brought to bear on experience only in conjunction with certain auxiliary assumptions. Call one set of those assumptions, A. Now, deploying our inferential sagacity, we discover that $(H\&A)\vdash O$, where O is some observable state of affairs. Looking into the matter, we discover the prediction is false: O' (which entails $\sim O$) obtains. Supposing O' to be veridical, we have refuted the complex (H&A). The Duhemian problem thus arises: Can we localize blame further or does experience leave us quite unable to ascertain where (as Lakatos used to say) the incriminating arrow of modus tollens falls?

For the last decade and more, a variety of latter-day Bayesians (Dorling, Howson, Redhead, Earman, and now Hellman) have claimed to have the solution to the problem of localizing blame—at least under certain circumstances. Specifically, they claim that so long as H is vastly more probable than A prior to the derivation of the mistaken result, then the rational response to the anomaly is to retain H and reject A. (Obviously, precisely the opposite conclusion would be indicated if the respective initial probabilities of H and A were reversed.) On their view, the degree of rational belief of H suffers modestly from the anomalous O', but provided that H remains more probable than not (which it will if its initial probability was sufficiently higher than A's), we know what to do in the face of the anomaly. Hey presto, falsification disambiguated!

Like so many of the other rabbits which subjectivist Bayesians pull out of their hat—syntax demands the plural here, but Bayesians have only one—this neither accords with much scientific practice nor does it begin to grapple with the subtleties of anomaly accommodation that a serious epistemology will have to embrace if it is to contend with Duhem's problem. The adjacent box lays out a series of stages that scientists typically go through when confronted by an apparently genuine anomaly for a complex involving plausible and important hypotheses. Although idealized in various respects, its diagrammatic character will enable me to suggest the distance between Bayesian precept and scientific practice.

If we are concerned, as Duhem was and I am, to find an hypothesis which can both explain the known phenomena (including O′) and do so in as robust a fashion as possible, then we need to takes stages II, III and IV seriously before deciding the epistemic impact of the failed prediction. By contrast, the Bayesians would short-circuit that arduous process by making a judgment about the rational acceptability of H immediately after Stage I.[1] For them, it does not matter whether we

Idealized Schema for Response to Duhemian Anomalies

Background: $(H\&A) \vdash O$; H is initially more likely than A

Stage I: Observe the anomaly O′ (inconsistent with O).

Stage II: Attempt to develop an A′, such that $(H\&A) \vdash O'$

Stage IIA: If II is successful, ask whether there is any independent evidence for A′

Stage IIB: If A′ exhibits independent evidence, ask whether A′ will substitute for A in all those cases—independent of the context of O and O′—where A may have functioned effectively

Stage III: Attempt to develop H′ such that $(H'\&A) \vdash O'$

Stage IIIA: If III is successful, seek independent evidence for H′

Stage IIIB: If IIIA is successful, ask whether H′ can replace H in other successful schemata of prediction and explanation.

Stage IV: Ask whether we can save O′ by adopting both a new H′ and a new A′ (likewise exploring the counterpart stages above for its scrutiny)

Stage V: Ascertain which complex, among (H&A′), (H′&A), and (H′&A′) is both better tested and more complete.

Worst Case Scenario: Unable to develop any complex meeting the required demands, suspend belief about both H and A.

1. I do not mean to imply that Bayesians could not produce reasons for moving beyond Stage I in the theory evaluation process. After all, Bayesians can give reasons for almost

have found a way to reconcile the recalcitrant anomaly, O′, and our favorite hypothesis, H. As long as we were highly confident about H going into the experiment (and less confident about A), then the experiment should not cause us to hesitate about affirming H—even though we have just learned that there is some relevant phenomenon in the domain which, so far as we know, H does not explain.[2] Mayo is quite right that "the Bayesian Way out of Duhem's problem is really no way out at all. . . . It does not illuminate the problem because if does not accord with how Duhem's problem is or should be grappled with in science." Although, as we shall see, Mayo and I disagree about some of these later stages, we are of one mind in believing that the touted Bayesian solution to Duhem's problem would be—if not associated with so many very bright folks—a bad joke. Its deepest failing is that it utterly fails to realize that the occurrence of the anomaly, O′, demands a grander gesture than a minor readjustment downwards in the probability of a preferred hypothesis. Always supposing (as the Bayesians themselves do) that the A was itself a plausible auxiliary (whose initial probability as the Bayesians figure it was greater than one half), the rational retention of H and the abandonment of A requires more than minor tinkering with their respective probabilities as they stood prior to the occurrence of the anomaly. One must come to terms with O′ and that means finding some way of accounting for it using whatever resources we have to hand or can invent. Leaving O′ dangling—as the Bayesian "solution" empowers us to do—is not only not standard scientific practice, it is bad scientific practice.

Ponder some of its implications. For one thing, it means that one can never be forced to abandon a deeply entrenched theory by a single anomaly, no matter how powerful the latter. Advocates of prima facie threatened hypotheses are under no epistemic obligation, according to the Bayesians, to show that their favorite hypothesis can function in a complex explaining the anomaly. The anomaly is simply shelved, ignored on the grounds that our acceptance of H is in no way hostage to H's capacity to account for O′. This, of course, was Kuhn's view about how normal science proceeds. Lakatos, who was scarcely less sympathetic than Kuhn to dogmatism about "hard-core" assumptions, understood that an H—which had functioned in a complex leading to a falsified prediction—must re-establish its credentials by offering a

any course of action. My charge, rather, is that, in their willingness to embrace H at Stage I, they are giving short shrift to considerations of explanatory scope in the evaluation of theories.

2. Philip Kitcher (1993) spends the bulk of a very long chapter interestingly exploring how scientists explore the "escape tree" associated with stages I–IV.

change in the "protective belt" which would bring H into an explanatory connection with O′. Failure to do so was, for Lakatos, sign of a degenerating research program. Bayesians apparently recognize no such presumption.

Consider how this works using one of Hellman own examples. In the 1840s, astronomers had become persuaded that Uranus' motion exhibited unusual deviations from the path predicted for it by Newtonian mechanics. Or rather, there were irregularities supposing (and here is the relevant auxiliary) that there was no large mass in the vicinity of Uranus. Now, according to the Bayesians, the rational thing to do in these circumstances (since the H here was more probable than the A) was simply to assume that Newtonian mechanics was right and that A was wrong. (This judgment follows on Stage I in our schema.) But that, of course, was not what either Adams or Leverrier did. Instead, they tackled the inverse perturbation problem, attempting to find out whether Newtonian mechanics could, utilizing a different auxiliary, find a way of explaining the recalcitrant behavior of Uranus. (Stage II) Nor did they stop there. Having computed the parameters of the required orbit of the hypothetical planet, they sought independent evidence for the auxiliary—in this case, ocular evidence of the existence of Neptune, which was found by Galle in 1846. (Stages IIA and IIB) The subjectivist Bayesian would have so say that all this work was unnecessary since everybody knew all along that Newtonian mechanics was right. Pursue the case a bit further. Suppose Neptune had not been found or that its orbit had parameters very different from those predicted from Newtonian mechanics. By any standards, this would have been telling prima facie evidence against Newtonian mechanics and scientists would have proceeded to stages III and IV. But a Bayesian scientist need never, except by inadvertence, discover such evidence since he is not obliged to show the explanatory prowess of his theory, once it reaches a certain level of probability.

The point is that over and over again in the history of the physical sciences, much of the most powerful evidence, both positive and negative, that has emerged about theories has been developed through researches in Stages II, III and IV. The general moral is that we can decide that deflecting the refutatory force of an anomaly away from a prized hypothesis is reasonable only *after,* not before, finding a new, independently testable auxiliary which will allow H to explain O′. (And even then, failure to pursue stages III and IV leaves any judgments about H more precarious than they should be.) Mayo is quite right to insist that anomaly adjustment mechanisms must be independently scrutinized.

3. Gloom and Duhem: Mayo-ist Ambiguities about Falsification and

Testing. In general, I am very sympathetic to Mayo's advocacy of the techniques of error statistics. Large chunks of applied science (which is not for me a term of abuse), as well as much of the logic of experimental design and interpretation, depend crucially on error statistical techniques. Like her, I do not believe that degrees of belief have anything to do with the evaluation of theories. Like her, I am not much taken with the subjectivism of Bayesian personalism nor with the epistemological question-begging which typically accompanies Bayesian's attempt to grapple with the problems of the catchall hypothesis. I likewise agree with her that the severity of tests is a central notion in science, although she and I would almost certainly differ over the details about how that key concept is understood (of which more below). Having indicated how far Mayo and I are in agreement on the larger questions of scientific epistemology, I nonetheless want to take exception to the specific treatment she offers of the Duhem problem, not least because I believe that—in her formulation of that problem—she has, at least partially, allowed herself to be taken in by its Bayesian formulation. We can best see that by recalling the familiar problems associated with the Bayesian catchall hypothesis. Non-Bayesians, of which I am one, regard it as a debilitating obstacle to the realization of the Bayesian project. Ardent Bayesians tend to regard it as a minor nuisance, nothing that a bit of clever hand-waving cannot dispel or at least repair. Because that terrain is already so familiar, I shall not rehearse it here. I mention it at all only because the avowed anti-Bayesian, Deborah Mayo, *seems*—in her solution to the Duhem problem—to fall back on the statisticians' equivalent of the probability theorists' catchall hypothesis.

Recall that the key element in Mayo's resolution of the Duhem problem is the demand that A must pass (or ~A must fail) a severe test, before we are entitled to retain or to reject H. For her, T is a severe test of H if and only if there is a very low probability that H would pass T *if* H *were false*. Now, if there are difficulties, as many (including Mayo) allege, with the Bayesians hope of calculating the value of the catchall hypothesis (viz., the probability of e, explained by H, if H is false), Mayo appears to run squarely into the same problems with her demand that we must calculate the probability that H would pass test T *if H were false*. With the Mayo-ists, as with the Bayesians, we seem to be confronted by a probability calculation involving an indefinitely long disjunction of rival hypotheses to H. Such, at any rate, is the upshot of a recent critique of Mayo's notion of test severity by John Earman (1992).

In her splendid *Error and the Growth of Knowledge* (1996), Mayo responds to this criticism by insisting that what is up for test in empirical research are hypotheses quite different from what has been usually supposed. Consider everyone's favorite example: the light-bending experiment and GTR. Mayo argues that we do indeed have here a severe test. But, as she acknowledges, it could not possibly be a severe test of GTR. Why not? Because, so far as we know, there are multiple, rival theories of space and time besides Einstein's all of which would predict a larger-than-Newtonian deflection of light passing close to the Sun. Under such circumstances, one cannot say that GTR is unlikely to have passed the test in question if it were false. Why, then does Mayo hold that the phenomenon provided a severe test? Because, as she plausibly argues, those test results are a severe test (in Mayo's technical sense) not of GTR but of the hypothesis that "there is a deflection of light approximately equal to that predicted by Einstein's theory."

But suddenly we seem to have drastically changed the subject. According to Mayo, a test, even a severe test, of the light-bending hypothesis leaves us in the dark about the ability of GTR to stand up to tests of different ranges of its implications. For instance, should GTR's success in the light-bending experiments lend plausibility to GTR's claims about gravity waves or black holes? Mayo's strictures about the limited scope of severity seem to preclude a positive answer to that question. For her, all the light-bending experiments showed was that the light-bending type implication of GTR had been severely tested—nothing more. This is changing the subject because we were initially seeking a way to avoid the apparent ambiguity of falsification as it affects the principal theories of the physical sciences—theories like the GTR. Mayo's characterization of the notion of a severe test appears to concede, even to require, that such theories generally cannot be severely tested. But in that case, the Duhemian ambiguities return with a vengeance. If we ask "Was the detection of Neptune a severe test of Newtonian mechanics?", the Mayo-ist answer is no. Was the determination that light moves faster in air than in water a severe test of the wave theory of light? Was the prediction of residual radiation in space a severe test of Big Bang cosmology? In every case, Mayo has to say no; and that in turn means that in all such cases, the Duhemian ambiguities of falsification go unresolved at the level of general theory.

In defense of her error-testing strategy, Mayo claims that, even if we cannot severely test many of the classic theories of the science, we can nonetheless say "that we have learned about (i.e., severely tested) one facet or one [sub]hypothesis of some more global theories such as GTR" (1996, 190). Her idea seems to be that, by such piecemeal testing of one after another of the subhypotheses making up a general theory, we can eventually ascend to a judgment about the general theory itself:

"When enough is learned from piecemeal studies, severe tests of higher level theories are possible" (1996, 191).

But this optimistic surmise flies in the face of Mayo's earlier acknowledgment that there are, so far as we know, indefinitely many general theories all of which "fit" whatever lower-level hypotheses have actually passed severe tests. If that is true, the prospects for severity to flow upwards by a process of gradual concatenation appear doomed to failure.

4. Comparativism to the Rescue. Is there some way to factor in these joint demands for well-testedness and scope without running afoul of the problems of the catchall hypothesis and its Mayo-ist counterpart, severity? I want to make a proposal which may move us partway to a solution. Let us drop the pretense, dear to the hearts of both Bayesians and error statisticians, that our evaluations of hypotheses are absolute. Instead, let us say explicitly what scientific practice already forces to acknowledge implicitly, viz., that the evaluation of a theory or hypothesis is relative to its *extant* rivals. To accept H is to hold that it is more reliable than its known rivals; to reject H implies that it is worse than at least one of its known rivals. Within such a comparativist perspective, we can say that a theory has been severely tested provided that it has survived tests its known rivals have failed to pass (and not vice versa). We need not require it to have passed all the tests to which it has been submitted, thus acknowledging that all theories characteristically exhibit some anomalies. But we do demand of an acceptable theory that it exhibit no anomalies not afflicting its rival or rivals. Now, on this analysis, when we ask if GTR can be rationally accepted, we are not asking whether it has passed tests which it would almost certainly fail if it were false. As Mayo herself acknowledges, we can rarely if ever make such judgments about most of the general theories of the sciences. But we can ask "Has GTR passed tests which none of its known rivals have passed, while failing none which those rivals have passed?" Answering such a question requires no herculean enumeration of all the possible hypotheses for explaining the events in a domain. The fact that there are in principle non-GTR theories which could pass all the same tests that GTR has passed is, from a comparativist perspective, neither here nor there—until such time as these in-principle theories are given flesh-and-blood in the form of a clearly articulated formulation. By relativizing severity to the class of extant theories, one can determine severity without stumbling over the problems of the catchall hypothesis.

Another feature of comparativism is worthy of mention. By demanding that an acceptable theory, T, must pass tests which none of its rivals have passed (and not vice versa), we simultaneously insure not only that T is better supported than its rivals, we also build a scope or completeness requirement into the appraisal of T. We guarantee that there is no fact in the domain of investigation which a rival can accommodate (in a non-ad hoc fashion) which T cannot.

According to Mayo, the Duhem problem is that of determining "which of the hypotheses used to derive a predicted consequence should be rejected." For the comparativist, the problem rather is that of deciding which, if any, of the hypotheses implicated in the failed prediction should be retained. This may seem akin to the glass half-full-half-empty conundrum, since Mayo will surely say that once we know what to reject, we will inevitably know what to accept (viz., what has not been rejected from the original complex). It is not so simple as that. Even if we have decided (say) that A is flawed, it obviously does not follow H is exonerated. For the Bayesians, what justifies the retention of H after blame has been localized to A is the fact that H went into the test with a high probability. But Mayo emphatically does not believe in such things. Besides, H's continued acceptability after O' hinges on whether H can account for O′. Deciding to jettison A—which Mayo thinks terminates the Duhemian discussion—still leaves us in doubt about the explanatory resources of H. Until those doubts have been laid to rest, we may have to suspend belief about H as well.

More importantly, comparativism keeps the focus on general theories, without diverting it away, Mayo-fashion, onto their subordinate parts. The comparativist believes that, if a theory like GTR explains or predicts phenomena which its known rivals have not and apparently cannot, then we have good grounds for preferring GTR to its known rivals. He insists on the point, which Mayo explicitly denies, that testing or confirming one "part" of a general theory provides, defeasibly, an evaluation of all of it. Thus, the observation of Neptune provided grounds for a global claim about the superiority of Newtonian mechanics to its known rivals, not merely support for the Newtonian subhypothesis that there was a massive object in the vicinity of Uranus.

In attempting to free us from the Duhemian ambiguities of testing, Mayo seems to have so balkanized the testing process that global or otherwise very general theories can rarely if ever be said to be well-tested. Faced with this Hobson's choice between the Bayesian catchall hypothesis and Mayo's fragmented testing of subhypotheses, the proper response is to retain our focus on general theories, while abandoning the pretense that our evaluations are grounded in a comparison of all possible rivals. Comparative theory evaluations may offer the tools for doing precisely that.

REFERENCES

Duhem, P. (1969), *To Save the Phenomena.* Chicago: University of Chicago Press.
Earman, J. (1992), *Bayes or Bust?* Cambridge, MA: MIT Press.
Kitcher, P. (1993), *The Advancement of Science.* Oxford: Oxford University Press.
Laudan, L. (1977), *Progress and Its Problems.* Berkeley: University of California Press.
Mayo, D. (1996), *Error and the Growth of Experimental Knowledge.* Chicago: University of Chicago Press.
van Fraassen, B. (1980), *The Scientific Image.* Oxford: Oxford University Press.