Realism, Underdetermination, and a Causal Theory of Evidence[1]

RICHARD N. BOYD

CORNELL UNIVERSITY

I shall be concerned in this paper to defend scientific realism against the thesis that the structure of the scientific theories we accept is radically underdetermined by any possible experimental evidence. In the course of this discussion I will have occasion to advance an account of scientific evidence which, I believe, extends in interesting ways the considerations which have led some philosophers to advance "causal theories of knowledge."

By scientific realism I mean the doctrine that the sort of evidence which ordinarily counts in favor of the acceptance of a scientific law or theory is, ordinarily, evidence for the (at least approximate) truth of the law or theory as an account of the causal relations obtaining between the entities quantified over in the law or theory in question. On this view, experimental evidence for a theory which describes causal relations between "theoretical" (that is, unobservable) entities is evidence not only for the correctness of the observational consequences of the theory, but is also evidence that the particular causal relations in question *explain* the predicted regularities in the behavior of observable phenomena. Of course this does not mean that, in the general case, experimental evidence for a theory is evidence that the causal relations it describes between observable *or* theoretical entities exhaust those causal relations obtaining between them (although this might be the case in the case of theories which were suitably "complete"). But it does entail that experimental evidence for a theory is evidence that those causal relations it describes, *and not others incompatible with them*, operate to produce the regularities in observable phenomena which the theory predicts.

This last feature of scientific realism has been thought by many philosophers in the tradition of logical empiricism to embody a fatal weakness. They argue that, given any theory which contains

non-observational terms and is consistent, it is always possible to produce alternative theories which share with the original theory exactly the same set of observational consequences, and which advance what are clearly *incompatible* causal explanations at the theoretical level for those observational predictions. Since these theories all have the same observational consequences (and, they might add, comparable degrees of "simplicity"), and since experimental evidence for or against a scientific theory arises from the success or failure of one of its observational predictions, they argue that the choice between one or the other of these theories cannot be a matter of experimental evidence. Two such theories would be equally confirmed or disconfirmed by any possible experimental evidence, and thus—since they also offer incompatible accounts of the causal relations between theoretical entities —it is impossible that we should have (as realists suggest) experimental evidence for any particular account of the causal relations between unobservable entities.

It is this argument that I shall be concerned to refute here. Its force depends on an apparently innocent principle:

(1) If two theories have exactly the same deductive observational consequences, then any experimental evidence for or against one of them is evidence of the same force for or against the other.

I hope to show that, on every possible reading useful to the argument indicated above, (1) is false.

I should clarify something at the outset: I am concerned to defend scientific realism, not just the thesis that "ontological commitment" to theoretical entities is "methodologically" or epistemologically "legitimate." Philosophers who advance the sort of arguments I am criticizing here often grant the latter point—that ontological commitment to theoretical entities is methodologically legitimate, even efficacious, and claim that this is all that a sensible realist should ever have claimed. The difference between this position and scientific realism is this: scientific realism offers an *explanation* for the legitimacy of ontological commitment to theoretical entities. The point is this: to say that ontological commitment to theoretical entities is legitimate is to say at least the following: that *observational* evidence for a theory which contains *non*-observational terms is evidence for the truth of its as yet untested observational consequences even though the deduction of these consequences may crucially involve the non-observational portions of the theory. To the question why this

should be the case, scientific realism offers the broad outline of an answer: experimental evidence for a theory is evidence for the truth of even its non-observational laws and, hence, for the truth of observational predictions deduced from them, since deductive inference preserves truth.

I should make explicit another position which will underlie what I have to say here:

(2) Suppose that some principle of scientific methodology contributes to the reliability of that methodology in the following minimal sense: that its operation contributes to the likelihood that the observational consequences of accepted scientific theories will be (at least approximately) true. Then it is the business of scientific epistemology to *explain* the reliability of that principle.

I will not defend (2) at length here. I only wish to suggest that it has been tacitly accepted even by those philosophers who claim that their philosophy of science is purely descriptive, not normative or explanatory. Such considerations as (2) must, I suggest, underlie the decision that certain regular features of actual scientific practice must be reflected in one's "rational reconstruction" of scientific methodology, while others are to be treated as psychological or sociological artifacts, or even as examples of methodological errors in current or past scientific practice.

To return to the main theme, in order to assess (1) we must first establish how (1) is supposed to be interpreted with respect to the employment of "auxiliary hypotheses" in the deduction of observational consequences from the theories in question. When the antecedent of (1) says that two theories have the same observational deductive consequences, what additional theories or laws, if any, are we to suppose are employed in making the relevant deductions? This question arises in view of the widely acknowledged fact that most of the important theories arising in the physical sciences have, unless other laws are employed with them as "auxiliary hypotheses," *no* non-trivial observational consequences. This is, for example, true of any two consistent theories which are stated entirely in non-observational terms. Principle (1) is absurd if the antecedent is taken as referring to the observational consequences of the theories by themselves (*i.e.*, with auxiliary hypotheses not employed in the deductions), since it would claim that the experimental evidence for classical mechanics is exactly as good as that for special relativity, if only both theories are stated abstractly enough. So presumably (1) is to be interpreted so that

the sets of observational consequences compared in the antecedent are to be taken relative to the theories in question *together with* some auxiliary hypotheses or other. Which auxiliary hypotheses are these to be? Only three possibilities suggest themselves, and they give rise to these three versions of (1):

(1′) If T and T' are each consistent and have exactly the same observational consequences no matter which set of possible auxiliary hypotheses is employed with both in the course of the deductions, provided only that the auxiliary hypotheses are consistent with T and T', then T and T' are equally supported or disconfirmed by any possible experimental evidence.

(1″) If T and T' are each consistent and have the same observational consequences when one is allowed to employ with each of them as auxiliary hypotheses any set of *currently accepted* laws or generalizations which forms together with the theory a consistent set, then T and T' are equally supported or disconfirmed by any possible experimental evidence.

(1‴) If T and T' are each consistent, and if, when one is allowed to employ with each of them as auxiliary hypotheses whatever laws or generalizations *will eventually be accepted* (and not thereafter rejected) in the course of scientific research, T and T' have the same observational consequences, then T and T' are equally supported by any possible experimental evidence.

Version (1′) is certainly true provided that at least one of the theories has some non-observational terms and provided that the set of observational consequences which these theories each yield with no auxiliary hypotheses leaves some observational question unsettled (*i.e.*, is not a complete subset of the set of all observation statements). Version (1′) is true under these conditions only because, subject to these restrictions, two theories satisfying the antecedent of (1′) must be exactly the same theory (their deductive closures must be identical). But this means that (1′) cannot be employed to defend the radical underdetermination of theoretical structure by any possible observational evidence.

Version (1″) is patently false since the truth of the antecedent does not preclude the possibility of experimental evidence suitable for discovery of new laws or generalizations, not currently accepted, which could be, in turn, employed as auxiliary hypotheses to derive contradictory, observationally testable predictions from T and T'.

Version (1‴) might be offered to avoid the difficulties involved

in (1″). It is difficult to say whether it even makes sense, or whether there is a non-empty set of laws or generalizations which will eventually be accepted and never rejected. At any rate, even if (1‴) is meaningful and true, it is of no use to the defender of radical underdetermination, since, barring precognition, we know of no technique of logic, or historical prediction, for showing the antecedent of (1‴) true except in the case where the two theories are identical.

Thus it would appear that no version of (1) is true which is also useful for the defender of radical underdetermination.

I think that the argument just presented is sound, but that it may fail to attack directly the intuition upon which the plausibility of radical underdetermination rests. That intuition rests upon another false version of (1), as we shall see presently.

By way of getting at this remaining intuition, consider the example of experimentally indistinguishable, causally incompatible theories which has been paradigmatic at least since the publication of Reichenbach's *Philosophy of Space and Time* [3]. Let F be current physical theory, and, in particular, let F contain a "catalogue" of the sorts of forces which operate in physical systems. Let G be the geometrical principles which are true if "straight line" is taken as "trajectory of an (idealized) point mass upon which the resultant of the forces acknowledged by F is zero." Let G' be an alternative set of (suitably comparable) geometrical axioms, and let F' be the physical theory which results from the addition to F of laws governing an additional universal force f' with the following property: f' is so defined (rigged, as it were) that G' is the correct physical geometry if the physical interpretation of "straight line′ " is amended so that the relevant trajectories are those of point masses upon which the forces acknowledged by F together with the force f' have resultant zero.

Now, the two theories "F and G" and "F' and G' " have exactly the same observational consequences when taken together with those currently accepted theories with which they are respectively consistent. Furthermore, since F' adds an additional force to our catalogue of physical causes, it would appear that they offer incompatible causal accounts at the theoretical level. Thus they are cited as providing examples of the sort of radical underdetermination being discussed. The example is particularly striking because it seems to show that there is no difference in experimental evidence between adopting a geometrical convention on the one hand, and discovering a new force on the other.

The reply to this example which is suggested by the arguments given here so far is to observe that it relies upon an application of (1″), and to argue that there might be subsequent experimental evidence for additional theories or laws which could be employed to distinguish "*F* and *G*" and "*F′* and *G′*" experimentally. In this case one might suggest, for example, the possibility that there would arise evidence for a general theory *T* of forces, say, one which portrayed all forces as arising from fields associated with particular particles, or from the motion of those particles. Theory *T* would either contradict *F′* outright, or allow the deduction of an experimentally falsifiable observational prediction from *F′*.

At this point, I think, we can see how the remaining underdeterminist intuition functions. The philosopher who thinks of "postulating" a new force as analogous to adopting a new convention will not accept the possibility outlined above so easily. He will focus his objections on consideration of the role that received theoretical knowledge plays in the assessment of new data, saying something like this:

"There is something wrong with the proposal that we might discover some such general theory of forces as *T*. It is certainly true that, given that we have adopted '*F* and *G*' there could arise experimental data *D* which, assessed in the light of what we then accepted, would be evidence for a theory of forces such as *T*. But, if we instead had accepted the (currently) experimentally indistinguishable theory '*F′* and *G′*,' then the same experimental data *D*, once uncovered, would, in the light of the body of theory then accepted, be evidence not for *T*, but instead for another theory *T′*, which is like *T* except that it asserts about *forces other than f′* what *T* asserts about all forces. And *T′* no more gives rise to disconfirmation of '*F′* and *G′*' than *T* gives rise to disconfirmation of '*F* and *G*.' If, as we have argued, the choice between '*F* and *G*' and '*F′* and *G′*' is not a matter of experimental evidence but instead of something like convention, then the later choices of *T* or *T′*, respectively, would simply extend the relevant convention in the light of new data. Thus, even without precognition, we are in a position to know something like the antecedent of (1‴) in the case at hand. We can know that scientific research governed by the *conventional* adoption of '*F* and *G*' will lead to the refutation of *F*, under all and only those circumstances in which scientific research governed by the conventional adoption of '*F′* and *G′*' would lead to the refutation of *F′*. Thus, if the scientific realist cannot show '*F* and *G*' and '*F′* and *G′*' to be currently distinguish-

able on experimental evidence despite having the same observational consequences, there is little hope of his showing that future research could distinguish them on the basis of experimental evidence."

This sort of argument suggests that the choice of one or another of two causally incompatible but experimentally indistinguishable theories could give rise to experimentally indistinguishable scientific traditions, which, when developed, continue to offer radically different accounts of the causal relations among theoretical entities. Furthermore, the argument depends on a methodological principle—that the interpretation of new experimental data should reflect the current state of theoretical knowledge—which a scientific realist certainly cannot reject. Scientific realism is, after all, offered in part as an explanation for the legitimacy of such "intertheoretic" considerations in scientific methodology. Perhaps it is true that a scientific realist must insist that "*F* and *G*" and "*F′* and *G′*" are currently distinguishable on experimental evidence despite having the same observational consequences when taken together with currently accepted theories. That would require the rejection of the following weakened version of (1″):

(1″a) If *T* and *T′* are each consistent and have the same observational consequences when one is allowed to employ with each of them, as auxiliary hypotheses, any set of currently accepted laws or generalizations which forms, together with the theory, a consistent set, then *T* and *T′* are equally supported by any possible experimental evidence, provided that this evidence does not dictate the acceptance of some new law, or the disconfirmation of an old one.

Version (1″a) is, I believe, also false, and its falsity points to a realistic reply to the new argument for radical underdetermination which we have just examined. In order to see that (1″a) is false *enough* to serve the realist's purpose, however, we will have to look quite closely at the notion of experimental evidence. This is so because what a realist should say is that *right now* there is some reason to *reject* "*F′* and *G′*" in favor of "*F* and *G*" and that the reason is somehow a matter of experimental evidence. He should say this:

Even though "*F* and *G*" and "*F′* and *G′*" have the same observational consequences (in the light of currently accepted theories), they are not equally supported or disconfirmed by any

possible experimental evidence. Indeed, *nothing* could count as experimental evidence for "F' and G'" in the light of current knowledge. This is so because the force f' required by F' is dramatically unlike all those forces about which we now know—for instance, it fails to arise as the resultant of fields originating in matter or in the motions of matter. Therefore, it is, in the light of current knowledge, highly implausible that such a force as f' exists.

Furthermore, this estimate of the implausibility of "F' and G'" reflects *experimental* evidence against "F' and G'," even though this theory has no falsified observational consequences. This is so because the experimental evidence which led to the adoption of our current theories of force is evidence that there really are electrical, gravitational, magnetic, and other such forces *and* that they *all* do result from such matter-dependent fields as have been described. But, then, this fact—that all hitherto-discovered forces arise from such fields—is, in turn, evidence that *all* forces have such an origin. So the experimental evidence for our current theories of force is indirect experimental evidence that no such force as f' exists—and that "F' and G'" must be false.

I think that this is the argument the realist should offer. However, it relies on a principle for the assessment of the plausibility of a theory which says that new theories should, *prima facie*, resemble current theories with respect to their accounts of causal relations among theoretical entities. The radical underdeterminist also accepts this principle—but he claims that it is *not* a matter of experimental evidence at all, *but* that it is merely another example of the sort of convention which, he says, gives rise to experimentally indistinguishable but causally incompatible scientific traditions in the first place.

Thus we appear to have come full circle—we must decide whether inter-theoretic criteria of plausibility are to be counted as reflecting experimental evidence or merely convention so that we can decide whether to adopt scientific realism or radical underdetermination. But it turns out that we will count such criteria as reflecting experimental evidence relevant to the acceptance of a proposed theory *if and only if* we have already adopted a realistic position with respect to the experimental evidence for the currently accepted body of scientific theories.

But we have still made some progress—we have isolated as central to the problem of scientific realism the question whether

or not certain inter-theoretic judgments of likelihood or plausibility are matters of experimental evidence about causal relations or merely the reflection of arbitrary conventions. *And* we have uncovered an interesting fact about the realist's answer to this question: given the assumption that the theories we accept at any one time constitute a roughly accurate picture of causal relations among theoretical entities, then evaluating experimental evidence for proposed theories *in the light of plausibility judgments based on this collateral causal information* (*e.g.*, counting confirmation of experimental predictions as evidence only for *plausible* theories) might function to make it likely that a new theory, if accepted, would *also* provide a good account of such causal relations. We have not shown this claim beyond offering the single example of ruling out implausible theories as candidates for confirmation, but more compelling examples will follow.

Anyway, given all this, the question still remains how we can decide whether plausibility estimates based on scientific theories should be understood as reflecting experimental evidence for the truth of the causal claims made by those theories, or should, instead, be thought of as reflecting relatively arbitrary conventions.

Here I propose to appeal to (2) and to adopt the following strategy:

(a) find a methodological principle P which involves inter-theoretic considerations of plausibility of the sort we are investigating;
(b) show that the employment of P contributes to the likelihood that accepted scientific theories will be good predictors of the behavior of observables; and
(c) argue that the only plausible explanation for the reliability of P lies in the assumption that it operates with respect to background theories which themselves reflect the actual causal relations among theoretical entities in such a way as to make it likely, in turn, that newly accepted theories will also provide approximately true causal accounts at the theoretical as well as the observational level.

This sort of strategy is suggested by remarks of Feigl [1] in which he discusses the heuristic role of such inter-theoretic considerations in the development of the atomic theory of gases. His examples focus on these inter-theoretic considerations as they affect the discovery and development of scientific theories. I shall avoid the difficult area of "context of discovery" and work

instead with a quite mundane methodological principle central to the "context of justification."

Before we examine this principle, however, it is interesting to observe that the strategy involved here is closely related to the efforts of Goldman and others to articulate a causal theory of knowledge. Goldman [2] suggests that there are cases in which in order to know something about an event, it is necessary that one correctly reconstruct the causal chains connecting the event to the phenomena which serve as one's evidence. This means, in particular, that acquiring new knowledge is possible only if certain of one's background beliefs about causal relations are already true. The strategy employed here is to show such a result about the principle P. That is, to show that P would fail to contribute to the likelihood that accepted scientific theories have (approximately) true observational consequences unless the "collateral" theories with respect to which plausibility judgments are made are approximately true (as causal accounts) and unless P contributed to the likelihood that accepted theories are likewise true.

The principle P we will use as an example is this one:

(P) a proposed theory T must be experimentally tested under situations representative of those in which, in the light of collateral information, it is most likely that T will fail, if it's going to fail at all.

Consider the following example:[2] A set L of laws is proposed which specifies the lethal effects of an antibiotic A on bacterial species in some class C. It is proposed by L that, by some chemical mechanism M, A dissolves the cell walls of bacteria in C. From L, together with appropriate chemical laws and facts about population growth in C, it is possible to predict the population of bacteria in a certain environment as a function of their initial population, the dosage of A to which they have been exposed, and the time elapsed since exposure. How could one decide which sorts of experiments are crucial to establishing the acceptability of L? What sorts of considerations involving collateral information dictate which experiments are crucial?

Example I: Suppose that a drug somewhat similar to A is known to affect those bacteria to which it is fatal *not* by dissolving cell walls *but* only by interfering with the development of new cell walls after mitosis. This suggests that a *plausible* alternative to M might be a mechanism similar to this one. Supposing this

possibility not to be ruled out by other collateral information, it might be crucial to test the predictions of the theory under circumstances involving a time much smaller than that required for the typical bacterial cell of the sort in question to divide, together with a large dosage which L predicts will be fatal to most bacteria in this time interval (assuming that L yields such convenient predictions). If the alternative mechanism actually explained the lethal effects of A, one would expect the predictions of L to be falsified here, since by the time the brief interval had elapsed, most of the original bacteria would remain as yet undivided, with their cell walls, and their health, intact. Note that in the absence of such collateral information as we have postulated which makes this particular alternative plausible, the particular conditions here crucial to the testing of L might be of no particular significance.

Example II: Suppose it is known that certain bacteria in the class in question are particularly prone to mutations affecting the structure of the cell walls. Unless additional collateral information ruled it out, this might suggest the plausibility of the suspicion that the predictions of L might fail under circumstances in which the bacteria survived in numbers for enough generations to stand a chance of producing a mutation whose cell wall was resistant to M. Thus, it would be crucial to test the predictions of L under conditions of low dosage and over time intervals appropriately long so that such mutations would have a chance to occur. Once again, in the absence of the sort of collateral information considered, the same measurements might be of no particular importance.

The examples can be multiplied indefinitely, for this and other cases. The point is that among the criteria for the adequacy of the experimental testing of a theory is this one: that it should be inquired, in the light of available theoretical knowledge, under what circumstances the causal claims made by the theory might plausibly go wrong, either because alternative causal mechanisms plausible in the light of existing knowledge might be operating instead of those indicated by the theory, or because causal mechanisms of sorts already known might plausibly be expected to interfere with those required by the theory in ways which the theory does not anticipate.

It need hardly be argued that the operation of this principle is central to the adequacy of scientific methodology in the sense of (2).

But, I suggest, the only explanation for the reliability of this principle lies in a realistic understanding of the relevant collateral

theories. Suppose you always "guess" where theories are most likely to go wrong experimentally by asking where they are most likely to be false as accounts of causal relations, given the assumption that currently accepted laws represent probable causal knowledge. And suppose your guessing procedure works—that theories really are most likely to go wrong—to yield false experimental predictions—just where a realist would expect them to. And suppose that these guesses are so good that they are central to the success of experimental method. What explanation beside scientific realism is possible? Certainly *not* the mere effect of conventionally or arbitrarily adopted scientific traditions. It might be conventional to test theories according to P, but unless, as no empiricist would suggest, the world is molded by our *conventions*, there is no way that the reliability of this principle could merely be a matter of convention. It must be the case that, where P functions reliably, its reliability rests upon the accuracy of the causal claims represented by the collateral information.

But then inter-theoretic judgments of plausibility of the sort we have been discussing must be understood as constituting reliable "inductive inferences" from collateral laws representing probable causal knowledge. Since the inference from existing theories of force to the probable falsity of "F' and G' " is of the same sort, we must count that inference as showing that the experimental evidence for the existing theories is also evidence against the theory "F' and G' ". So the strong rejection of (1″a) necessary to refute the last radical underdeterminist argument is sound.

Hence, there is no version of (1) available to the defender of radical underdetermination. Indeed, principles contrary to (1) are absolutely essential to the experimental testing of scientific theories.

References

[1] Feigl, Herbert, "Logical Reconstruction, Realism, and Pure Semiotic," *Philosophy of Science* 17 (1950).

[2] Goldman, Alvin, "A Causal Theory of Knowing," *Journal of Philosophy* 64 (1967): 357-372.

[3] Reichenbach, Hans, *The Philosophy of Space and Time*, translated by Maria Reichenbach and John Freund (New York: Dover Publications, 1958).

Notes

[1] I would like to thank Ned Block, Harold Hodes, Barbara Koslowski, Mark Pastin, and Hilary Putnam for many helpful discussions about the topics discussed here.

[2] I wish to thank Robert Becklen for helpful criticisms of an earlier version of this example.