Nisbett, R. and L. Ross. (1980) *Human Inference: Strategies and Shortcomings of Social Judgment* (Englewood Cliffs, NJ: Prentice-Hall).

Putnam, Hilary (1982), "Why Reason Can't be Naturalized," *Synthese* vol. 52, pp. 3–23.

Ruse, Michael (1986) *Taking Darwin Seriously* (Dordrecht-Holland, Reidel).

Shimony, Abner (1970) "Scientific Inference." In R. G. Colodny, ed. *The Nature and Function of Scientific Theories* (Pittsburgh: University of Pittsburgh Press).

Simon, H. A. (1945) *Administrative Behavior* (New York: The Free Press).

Simon, H. A. (1957) *Models of Man* (New York: John Wiley & Sons).

Simon, H. A. (1977) *Models of Discovery* (Dordrecht-Holland: D. Reidel).

Simon, H.A. (1983) *Models of Bounded Rationality*. 2 vols. (Cambridge, Mass.: M.I.T. Press).

Simon, H. A., P. Langley, G. Bradshaw and J. Zytkow. (1986) *Scientific Discovery* (Cambridge, Mass.: MIT Press).

Stegmueller, Wolfgang (1979) *The Structuralist View of Theories* (New York: Springer).

Tolman, E. C. (1948) "Cognitive Maps in Rats and Men." *Psychological Review* vol. 55, pp. 189–208.

Toulmin, Stephen (1972) *Human Knowledge* (Princeton: Princeton University Press).

van Fraassen, B. C. (1980) *The Scientific Image* (Oxford: Oxford University Press).

EXPLANATORY SUCCESS AND THE TRUTH OF THEORY

ERNAN MCMULLIN

WHEN Aristotle, in his account of the nature of *epistēmē* in the *Posterior Analytics*, looked for examples to illustrate the notion of demonstration which he claimed to be the defining characteristic of science, he turned to astronomy. On the face of it, this would not seem surprising, since astronomy might plausibly have been regarded as the most obvious candidate for a science of nature in his day. Yet the examples he chose were puzzling. Recall that demonstration had to proceed from premisses themselves known to be true; to "demonstrate" was to show how the observed properties of things proceeded necessarily from the essences of those things. One might have expected Aristotle to choose examples where these essences were familiar to us so that the truth of the premisses could readily be guaranteed.

Instead, however, he drew his examples from the sky. Why does the moon have phases, crescent, half, and so on? Because it is spherical. Why do the planets (unlike the stars) shine steadily? Because they are relatively near. To "demonstrate" waxing and waning as a property of the moon was to show how it followed necessarily from the spherical shape of the moon (and its being illuminated from afar by the sun). To "demonstrate" that planets would shine with a steady light, it was sufficient to show that nearby objects like them would shine in this way. But how were we to *know* that the moon is spherical or that the planets are relatively near? We would have to *know* this to be true if the premiss were to function as the starting-point of demonstration.

Aristotle's account is elliptical at this point[1] What he seems to suggest is that what assures us of the truth of the claim that the moon is spherical is that this does indeed explain the observed appearances of waxing and waning. We do not observe the shape of the moon directly, anymore than we do the nearness of the planets. But what allows us to take these properties as truly characterizing ones is that doing so enables us to explain the immediately observed properties. Explanatory power here serves as the criterion of truth in cases where the truth cannot be more directly ascertained. What is striking, once again, about Aristotle's choice of examples is that he deliberately chooses cases where the truth of the premisses could not be plausibly supposed to be known in advance. (One might have expected him instead to begin from a premiss like: man is rational; such an attribution on our part would not need to depend for its truth upon the success of the explanation this provided of other features of human behavior.)

Aristotle's account is problematic, of course, and later generations would struggle to make a consistent theory of demonstration out of it. I begin with it here in order to underline that the relation between explanatory power and truth has been a central issue in the understanding of science right from the beginning. In Aristotle's day, the hard cases were the physical properties of distant objects not open to immediate inspection, and therefore to be known only indirectly. In recent centuries, the objects of scientific interest have become far more remote, not so much in distance as in their relation to our intuitive understanding, the sort of understanding on which Aristotle's conception of science depended. The relation between explanatory power and truth, never an entirely easy one to explicate has thus become far more problematic.

I. Theory, Explanation, Truth

What I want to argue here, in the context of theory as we know it in natural science today, is that the only means open to us of judging the *truth* of a theory is through an assessment of the explanatory success of the theory. More important, this assessment *does* allow us, in favorable cases, to make a truth claim of a limited sort for the theory. To make my case, I should lay out very briefly first some

familiar distinctions in regard to the three key terms in my thesis: "theory," "explanatory success," and "truth." The first of these may best be approached by contrasting two types of general statement found in science. A "law" derives from an observed regularity, and is arrived at by a process of generalization or induction. The inference from observed singulars to such a law is ampliative, that is, it goes beyond the evidence. But it does so only in the sense that it takes the instances observed to be a fair sample of a broader class. And of course the adequacy of the language used is also presupposed. A law-statement, in short, postulates a correlation between observable factors, on the basis of some limited experience of such a correlation.

A theory, on the other hand, is proposed as tentative causal explanation of a law or laws, as a way of explaining why the law holds. The joint process of arriving at theory and validating it is called "retroduction" by Peirce, because it means moving backward in thought from an observed effect to a hypothetical cause. It is ampliative in a much stronger way than is induction since it involves some degree of invention and imagination. One has initially to guess at the cause. The concepts in which the effect is described will not ordinarily suffice to describe the cause, and so an innovation of language may be needed. There are no rules (as there are in the case of deduction) that would tell us whether we are making the inference successfully.[2] In fact, even the term "inference" itself is misleading, since it might suggest something rule-governed. Conjecturing a cause, or more broadly, postulating a theory, is not rule-governed, although it will be influenced, up to a point, by prior commitments as to what counts as plausible.

What is important to us here is that theory cannot be assessed, as deductive inferences are, by the extent to which certain rules have been followed in its formulation. The criteria employed in theory evaluation are quite complex, and do not depend on how the theory has been arrived at. They have been attracting a good deal of attention of late since the point was first made that they function as values rather than as rules.[3] Here are just a few of the more obvious ones.[4] First, of course, comes empirical adequacy (or predictive accuracy), that is, the ability of the theory to account for the data at hand in a deductive (or at least a quasi-deductive) way. A second is proven

fertility, the ability that the theory has already demonstrated of predicting "novel" facts, as well as the ability it has shown of dealing with anomaly by imaginative modifications suggested by the theory itself.[5] A third is unifying power or scope, the ability the theory has shown of bringing together domains hitherto taken to be disparate. A fourth would be consistency with background presuppositions and other accepted theories. Fifth is coherence, "naturalness," the absence of those *ad hoc* features that can so damage an hypothesis, even one that otherwise "works." More "internal" than the others, finally, is logical consistency. Each of these considerations would obviously require more detailed discussion, but this summary listing may suffice for our purposes here.

These are some of the "values" scientists look for in a theory, considered as an explanatory hypothesis. Taken together, they allow an estimate to be made of the *explanatory success* of the theory. A failure in any one of them is regarded as a failure in explanatory power. Empiricist philosophies of science have tended to assimilate explanation with prediction, and hence to focus on the first criterion only, as though one could evaluate a theory simply in terms of its predictive accuracy. This was especially appealing to the logical positivists because it seemed obvious to them that the warrant for a theory could lie only in its empirically verified logical consequences. The inductive strength of an hypothesis could come, they argued, only from its deductive relationships with the observational evidence. Other considerations might, indeed, intervene but they could be regarded as "aesthetic" or "pragmatic"; they were not to be counted towards the *truth* of the hypothesis. Nor could they, in consequence, be included in the requirements for explanation proper.

The difficulty with this approach was, of course, that the primary function of a theory is to explain, and that its failings as explanation would seem on the face of it to count against its credentials as a good theory. A theory with features that are perceived as *ad hoc*, for example, may be perfectly adequate as a means of accurate prediction, but these features will count against the theory as explanation and will prompt efforts to find an alternative. One could say that this is because it is anticipated that such a theory will ultimately fail as a means of prediction because of the presence of these *ad hoc* features, so that predictive accuracy might still be regarded as in some sense the primary value. But the fact remains that the evaluation of a theory does *not* simply reduce to asking about its present efficacy as a predictive device.

These criteria have been described above as a sort of "given," and of course this is much too simple. Have they changed over time? On what does their warrant *as* criteria of theory rest? On the consensus of practicing scientists? On their effectiveness as leading to predictively powerful theories? On their intrinsic logical merits? These issues have given rise to lively discussion among philosophers of science in recent years, and no resolution of the differences appears to be in sight. It is fair to say that these values *are*, in fact, employed in theory-evaluation in contemporary natural science. There would not, however, be agreement among working scientists as to precisely how they should be understood or how much weight to attach to each. It is, indeed, this lack of agreement that explains in great part (though not entirely) the persistent presence of controversy in natural science, something that on the older simpler accounts of the theory-evidence relationship was quite hard to understand.

For the moment, I want to set aside questions about the status of these epistemic values and simply to note that there would, in the case of longer-standing theories at least, be a considerable measure of agreement among working scientists as to the degree of explanatory success to attribute to the theories. How does this success bear on the *truth* of the theory? Much, of course, depends on what one means by "truth" in this context. In the period between the two world wars, there was a great deal of discussion of the different "theories" of truth, prompted in the first instance by the writings of Moore and Russell. Philosophers debated about correspondence and coherence and pragmatic success, and generated arguments in support of the claim that truth resided in one and only one of these. I shall return to this debate later, but for the moment let me recall Aristotle's simple dictum: "To say of what is that it is, is true; to say of what is that it is not, is false." It is easy enough to apply this to "the cat is on the mat," or "snow is white." But what about a scientific theory? This is a very complex sort of statement indeed. What would it mean to say of such a theory that it is "true"?

A theory explains in the first instance by postulating a hypothetical structure of some sort, a set of entities, processes and

relations that are causally sufficient to account for the observational data.[6] An example will help. Most of the major features of the earth's surface, such as the great mountain ranges, the undersea ridges, the volcanoes, as well as some less obvious features like the geological match between parts of the coasts of West Africa and eastern South America, and parallel magnetic striations on the ocean floor, can be explained by postulating the existence of immense plates, carrying continents and sea-floor alike, slowly moving relatively to one another, sometimes pulling apart and sometimes pressing together with force enough to thrust mountains into the sky.

To say of such a theory that it is *true* would seem most simply to be construed as saying of these plates "that they are." Or as Wilfrid Sellars puts it: "To have good reason for espousing a theory is *ipso facto* to have good reason for saying that the entities postulated by the theory really exist."[7] A theory in this reading is interpreted as a complex causal hypothesis, which is true if and only if the entities postulated in the theory actually exist and operate as they are supposed to do. The idea is that if the causal mechanism does *not* exist, the explanation does not hold, not as an explanation, at least.

But is there any other way a theory *might* hold? A much weaker thesis would be that a theory is true if its observable consequences, its predictions, all hold good. In this reading, the causal mechanisms postulated by the theory are assumed to be no more than heuristic devices whose function is to aid scientists in extending the range of their predictions. No existential claim is intended, and thus the truth of the theory does not require that anything more exist than the effects themselves.

But is "truth" the proper term to use in such a case? It would seem not, and indeed most of those who interpret theory in this way would reject the term entirely in this context. The classic instrumentalist view of theory was that it served as nothing more than an instrument of prediction. A *good* theory is one that gets all the predictions right. Such a theory is a tool, not a truth-claim, and thus the term "true" is not to be used of it at all. Bas van Fraassen proposes a view he calls "constructive empiricism" in which a theory is held to nothing more than its consequences.[8] A theory whose consequences are duly verified is said, not to be true, but to be "empirically adequate." In van Fraassen's view, no more may be asked of a theory

than this. In particular, its success as a theory cannot be held to warrant in any way the stronger claim that the entities postulated by it exist, that is, that it is true.

I shall argue the opposite thesis, namely, that the success of a theory as a theory suggests the *truth* of the theory. And I mean this last phrase to be taken in its strong Sellarsian sense. Notice, though, that I only said "*suggests*." I am not saying that explanatory success *implies* the truth of the theory. The rest of what I have to say will be a gloss on this.

II. Scientific realism

At this point, it would be helpful to ask what the relationship is of all this to the doctrine known as scientific realism. In point of fact, there seem to be almost as many scientific realisms as there are scientific realists. I will have to rely here on my own definition of that much-fought-over label.[9] Scientific realism, let me stipulate, is the view that there is good reason to believe in the existence of entities substantially like those postulated by theories that have been successful over the long term. There are several important qualifications built into this definition. Briefly, I take realism to commit one only to saying that there is "good reason" and not that there are compelling grounds. It is always *possible* for even a highly successful theory, to be false; it may be extremely unlikely but we have to hold the logical possibility open. And the theory may develop further, so that the present specifications of the causal mechanisms may have to be revised and sharpened. Finally, only theories that have already shown a considerable degree of explanatory power would qualify as having reliable ontological implications. Each of these three reservations would need to be specified in further detail but I will have to leave that task aside here.[10]

How, then, does scientific realism so defined relate to the question of whether scientific theories may be said to be *true*? And, even more to the point, how do both issues tie in with the notion of explanatory success? My argument may be reduced to two constituent claims: first, that a scientific realist is committed to attributing *some* sort of truth-value to successful theories, and second, that the measure of truth-likeness or likelihood to be attributed to the theory

is precisely the degree of explanatory success the theory is taken to have shown.

The first of these claims is easier to fill out than the second. The realist asserts that when a theory satisfies to a high degree the criteria just listed, this gives us good reason to believe in the existence of something like the causal mechanisms it postulates. But it is, in fact, the existence of these mechanisms that would make the theory true. Can we say, then, that a realist should hold that one can ascribe truth to a successful theory? This is indeed the way realists sometimes talk, but unfortunately the inference is a little too quick. If the theory really *is* a theory and not a covert description (such as the "theory" that the blood is circulated by the heart, a theory in Harvey's time, but no longer in ours), it is inferring to causal mechanisms that are themselves not directly accessible to us, at the moment, at least. The manner in which it does this, the concepts in terms of which these mechanisms are described, in short, the theory itself, is subject in principle to further development. Something left blank in the present model may get filled in. The concepts in terms of which the theoretical entities are specified may be modified, perhaps even considerably modified.

The success of the plate-tectonic model in the two decades since it first gained wide acceptance among geologists around 1965 has been striking in every regard. It has, for example, managed to explain whole ranges of data that were not originally even seen to be related to it. Would this lead us to say that the theory is *true*? Clearly not. It is still under development. We can be quite sure that there are still anomalies ahead that will lead to modifications in the original model, just as the plate-tectonic model itself involved a fundamental change in the continental drift model that prepared the way for it. If by "true," we mean "literally true," that is that the causal mechanisms are *exactly* as the present theory postulates them to be, then we can be fairly sure that the theory is *not* true. To say that it is true would be to assert that nothing further will be discovered in all the wide domain of geology that would be sufficiently anomalous, relative to the present model, to lead to a modification in this model. It does not require very much familiarity with the history of science to know that even the highest degree of explanatory success in the case of a large-scale theory would not warrant such confidence.

A successful theory ought not, therefore, be described without qualification as "true." Aristotle long ago required science to generate eternal and necessary truth. We would be happy to settle for just plain truth. But even that is beyond our powers in all but the most limited explanatory contexts in natural science. Scientists recognize this fact and ordinarily are quite careful not to describe even their most cherished theories as "true," preferring weaker terms like "reliable" or "well-supported." And philosophers are similarly cautious, proposing substitutions like "warranted assertibility" for the older and more demanding term.

Someone who accepts a version of scientific realism would seem justified, however, in employing somewhat stronger language. A realist ordinarily allows that a high degree of explanatory success on the part of a theory gives good grounds for asserting the existence of entities of the general kind postulated by the theory. But this is sufficient to allow one to say that there are good grounds for asserting the theory to be "approximately true" or "highly likely." Critics have objected to the term "approximate" in this context, both because the term is vague and because the degree of approximation cannot, in principle, be specified. One cannot know in advance how much and in what directions a scientific theory may still have to be modified.

While this is true, there is nevertheless a sense of "approximate truth" that would seem perfectly defensible in this context. Consider again our geological example. The striking success of the plate-tectonic model in so many domains gives geologists good reason to believe it to be *true* that the continents and ocean floors are carried on massive rocky plates in slow relative motion. There is much about these plates and about the mechanisms repsonsible for their motion that geologists can as yet only guess at. In that sense, the theory is incomplete, open to further refinement. But suppose we focus not upon the specifics of the theory but upon the very general claim that there are large plates in relative motion. Geologists would, I think, say that we have very good grounds for believing *this* claim to be true.

Calling a theory "approximately true," then, would be a way of saying that entities of the general kind postulated by the theory exist. It is "approximate" because the theory is not definitive as an ex-

planation; more has yet to be said. But it already has a bearing on truth because we can say that it has allowed us discover that entities of a certain sort exist, entities that we could not (for the moment, at least) have known about without the aid of the theory.

To the extent that a theory postulating quite different sorts of explanatory entities is still an open possibility, we would want to qualify the claim made, and say that the original theory is "probably" approximately true, or less cumbrously, "highly likely." The notion of likelihood is useful here because it implicitly contains both types of qualification on the truth-claim being made, first that it is only "approximate" in the sense specified above, and second that there is a chance, though perhaps only a very small one, that entities of the general kind postulated do not exist after all.

Some other examples may help to fix the kind of claim that I am making. Astrophysics has developed steadily since the seventeenth century. At this point, we can be reasonably assured that the stars are large glowing masses of gas, containing chemical elements similar to those of earth. There is such an abundance of evidence converging on this general claim that it would be given at least as much credence as would the average report of a laboratory result. Much more than this can, of course, be said about the composition of individual stars, about the sources of their energy, about their likely past evolution and future development, about their grouping at very different distances from us, and so on. The relative assurance with which each of these further specifications can be asserted will depend on the explanatory power of the associated model and theory. But for the moment, we are interested only in the most basic claim, namely that the points of light in the night sky are either individual globes of incandescent gas like our own sun, or groups of such suns.

Any other mature structural science will serve to make the same point. The "billiard-ball" atom of the kinetic theory of gases, for example, has gradually taken on a more and more detailed structure. The existence of atoms themselves as structured individual entities of known mass and volume can now be asserted with assurance, even though much has still to be learned about the details of their structure. Our atomic theories are unlikely to remain in precisely their present form. But this does not prevent us from making a more general claim about atoms, basing it on the widest possible appeal to the explanatory success of a cluster of associated theories.[11]

III. Realism without theory-truth?

I have been arguing that scientific realists are committed to a truth-claim of *some* kind about the theories in which the entities whose existence they assert make their appearance. But would any realist actually *deny* this? Some would, at least, have doubts. In his recent book, *Representing and Intervening*,[12] Ian Hacking defends a form of realism that relies entirely on an experimental tracing of effect back to cause, avoiding any commitment to the truth-status of theories. What he asserts is not the existence of theoretical entities but of "phenomena" that can be produced experimentally. Since the photo-electric effect was first discovered in 1829 by Becquerel, he reminds us, one theory after another has been put forward to explain it. The current (photon) theory may be rejected in time to come, but "the supermarket doors (which depend on the photo-electric effect for their functioning) will still go on working."[13] Phenomena, once established, are not affected by theory-change. "The phenomena which we have created will still exist and the inventions will work." Once phenomena like the photo-electric effect are created in the laboratory, "they are phenomena thereafter, regardless of what happens." He calls this "experimental realism."

But who would disagree with this? Not the instrumentalist surely? Hacking is willing, it is true, to go beyond such "phenomena" to entities like electrons that "in principle cannot be 'observed'," as long as they can be experimentally "manipulated to produce a new phenomenon."[14] We come to know the mass and the charge of electrons and enough of their "causal powers" to enable us to "build devices that achieve well-understood effects in other parts of nature." Not only does a *theory* of electrons play no part in ensuring the reality of these experimental entities, it does not even serve to define or describe them either. There is, in fact, "no common core of theory," only a "a lot of theories, models, approximations, pictures . . . ."[15]

It is important to realize that Hacking opposes his experimental realism not to instrumentalism but to *idealism*. The theme of ex-

perimental intervention contrasts "doing" with "thinking," manipulation with explanation. In this context, to argue that something is "real" is to argue that it is not simply a creation of mind, that it imposes itself *on* mind. The category he perceives as threatened is not the unobserved entity but the "phenomena" themselves; the instrumentalist would find no difficulty in sharing this concern.

His second target is the extreme form of empiricism that leads van Fraassen to deny that we can "observe" through a microscope and to reject, in consequence, the reality of entities attested to only by such instruments. Hacking uses his notion of intervening once again to argue quite persuasively that the entities viewed in the microscope can be experimentally tested for "reality" in numerous ways, that is, shown not to be artifacts of the experimental techniques employed. The truth of the reality-claim in their regard does not depend in any way (Hacking argues) on having a *theory* about them. We may have no understanding at all of the organism we are observing under the high-power microscope, yet we might still be quite sure that it is real.

This example brings out both the strength and the limitation of Hacking's argument. The entities "observed" by scientists are ordinarily identified by a term, like "electron," that carries with it a "theory" of the entity involved. When a scientist claims to have observed an electron in a cloud-chamber, there is not only a causal tracing but an implicit commitment to the theory that enabled this identification to be made in the first place. If the theory were to be rejected, scientists would still suppose they had observed *something* but they would no longer report this observation in the same way. Calling the object an "electron" enables them to relate it to other entities in a network of relationships prescribed by theory. The case of the investigator who notes, for example, a hitherto unidentified reddish-colored needle-like entity on his microscope slide and suspects it to be (say) the agent that caused Legionnaire's disease is untypical. Once, however, this entity is classified as a virus or as a bacterium, a whole set of theoretical expectations immediately come into play.

What Hacking is telling us is that our reason for believing in the existence of the unidentified organism has nothing to do with a *theory* of that organism. Even in the case of the electron, our reason

for believing that the electron exists is primarily the causal tracing. The theory is needed only for identification's sake and not for the assurance of existence. The weight of an existence-claim must rest on intervening, he insists, even though an account of the outcome of the intervening will surely require some degree of representing.

Like Kuhn, he finds theory too weak a reed to support reality claims. But his causal argument for the existence of such an entity as an electron will hold (he reminds us) even if the *theory* of the electron should change drastically. Of course, the meaning of the existence-claim will change in such a case, but Hacking is not worried about this. Intervening makes it possible to know *that* the electron is, even though there is no similar assurance as to *what* it is. But once again, unless the "what" can be identified—and how else except in terms of theory?—the "that" would seem to be empty. It is not clear, then, that reference to explanatory success can be dispensed with, even in an argument that relies mainly on experimental intervention.

Hacking's argument for realism leaves aside most of the entities and processes that the theoretical sciences deal with. When chemists postulate hydrogen bonds between the component atoms of a molecule, for example, it is not on the grounds that these are "observed," even in the extended sense of that term advocated by Hacking.[16] The warrant for entities such as these is the success of the theories of which they form a part, nothing else. And what about objects distant from us in space (e.g. galaxies), or in time (e.g. dinosaurs)? Causal processes link us to them, it is true, and in the former case (though not the latter) we might even say that they are "observed." But no "intervening" is possible, so they do not appear to qualify for support under Hacking's argument. It is, however, this profusion of theoretical structures that lie beyond instrumental intervention that gives science much of its interest. Hacking does not deny their existence; he merely expresses his doubts as to whether valid arguments can be found to support existence-claims on their behalf by appeals to explanatory power alone.[17]

One conclusion we can draw from all this is that a broadly "realist" approach to the explanatory entities of the scientist can be defended in two quite different ways. An "intervention" type of argument is applicable only to a sub-class of these entities; it has the

advantage that it affords a relatively direct inference to existence. The argument from explanatory success is much more general in its application but is more vulnerable because of the complex ways in which theoretical concepts refer and the ever-present likelihood that the theories may change. This argument takes the same form whether or not the entities it postulates are "manipulable," in Hacking's sense, or not. The two arguments can, of course, be combined for "manipulable" theoretical entities, affording a greater degree of assurance in their regard. What the "intervention" type of argument does is to ensure that there *is* an entity there to begin with, even though it may have little to say about its nature other than that it causes certain effects. The argument from explanatory success goes on to fill in the nature of the entity, so that the two complement one another. One can see this dual process at work in the history of the electron concept, where from the beginning the electron was "whatever it is that is causally responsible for these experimentally-determined results," as well as a complex construct in a sequence of explanatory theories.

Where reference depends on theory alone, an argument from explanatory success is required in order to support the initial claim that an entity of this general sort exists. The familiar stories of phlogiston and caloric remind us that such assertions are defeasible. Note that the claim here is for an entity *of this general sort*, not of this *precise* sort. Theoretical entities are idealizations, possessing only the attributes given them by the theory. The success of the theory does not warrant the claim that something *exactly* corresponding to this construct exists; such an entity would have no "material" aspect, no further properties to explore.[18] The success of a theory can at best warrant one only in claiming that an entity possessing *among others* the properties attributed to it by the theory exists. And since the theory is unlikely to be definitive, it is more appropriate to say that certain properties the referent of the construct possesses are likely to be of the general sort prescribed by the theory.

The importance of these qualifications lies in the fact that theories develop and are modified, sometimes in fundamental ways. In this section I have been arguing that a realist is committed to attributing likelihood to successful theories in science. But realism

does not amount to holding that entities exactly and exhaustively described by the constructs of successful theory exist. Such a thesis would be defeated by the smallest theory-change. The continuity of reference of successful constructs is a much looser affair, as case-studies have shown. Such constructs suggest certain articulations of Nature, but these articulations may alter as theories develop and converge. The Dirac electron is completely specified by Dirac's theory. But the *real* electron, the entity whose nature has gradually become clearer, thanks to a confluence of theories and a strong intervention-type argument, is not exhausted by any theory we now have or for that matter ever may have. These are notoriously complicated issues; limitations of space prevent a fuller treatment.

IV. Inference from best explanation

My second thesis is that the likelihood of a theory depends on the degree of explanatory success the theory enjoys, construing "explanatory success" here in the broadest possible way. This resembles what Harman has called "inference to the best explanation":

> In making this inference, one infers from the fact that a certain hypothesis would explain the evidence to the truth of the hypothesis. In general, there will be several hypotheses which might explain the evidence, so one must be able to reject all such alternative hypotheses before one is warranted in making the inference. Thus one infers, from the premise that a given hypothesis would provide a "better" explanation for the evidence than would any other hypothesis, to the conclusion that the given hypothesis is true.[19]

Harman's label is not an entirely happy one. If he were right in saying that one must be able to reject *all* alternative hypotheses (something that is, in principle, hardly ever possible in science) before the inference were warranted, then we would have inference not just to the best, but to the *only*, explanation. Furthermore, we do not *infer* to a hypothetical explanation, strictly speaking, we *conjecture* it. Where inference comes in is in concluding that it *is* the best available explanation. We do not infer to the best explanation; we

infer that a given explanation *is* the best available explanation. In the passage quoted, Harman himself says, not that we infer to the best explanation, but *from* the claim that a given hypothesis is the best explanation *to* the conclusion that this hypothesis is true.

We have already seen that this last is far too strong. From the premise that an hypothesis is the best available explanation, we *cannot* infer that it is true, only at best that it is likely. Moving from best explanation to truth is the troublesome issue. Harman's problem is not with the inference that something is the best explanation (indeed, he dismisses this in a few lines), but rather with what one can infer *from* the assertion that a given hypothesis is the best explanation. Hence, inference *from* the best explanation (or even better, from explanatory success) would be a more appropriate label. The older term coined by Peirce, "retroduction," is in any event already available for this purpose.

How is inference from best explanation to the likelihood of the corresponding hypothesis to be supported? Earlier, we listed the three so-called "theories" of truth: correspondence, coherence, and pragmatic effectiveness. It has frequently been noted that the latter two are *criteria* of truth and are much less plausible as means of defining the nature of truth. Whereas correspondence quite obviously cannot be a criterion; it can at best only have something to do with the nature of truth.

Most of the difficulties that have been raised about truth-as-correspondence have been due to regarding it as a sort of *picturing* relationship. Wittgenstein and Rorty are only two of the many who have shown that the notion of picturing will not do in this context. But if this metaphor be set aside, one can retain the idea that it is the way the world is that determines what is true ("to say of what is that it is is true"). And the main criteria that allow us decide whether we have grasped the way the world is are clearly the coherence of the explanations we give and the accuracy of the predictions they enable us to make. These two generic criteria recall the criteria of theory-assessment already discussed. Coherence groups together a number of desirable values: internal consistency, consistency with other well-established theories, lack of *ad hoc* features, and so on. We also want a theory that is predictively accurate, whose practical applications work out as they should, and which enables a fruitful

research program to be carried on. The criteria that scientists would regard as appropriate to the assessment of a theory resemble, therefore, to a striking extent the criteria traditionally thought to be symptoms of truth generally. The values that constitute a theory as "best explanation" are, it would seem, similar to those that would qualify a statement as "true."

I can put this in a more concrete way. Suppose, as the realist supposes, that a scientific theory may be said to warrant some sort of qualified truth-claim in regard to the entities constituting it as explanatory. How ought a theory so understood be evaluated? How, for example, would one assess the claim that the core of the earth is made up of molten iron? I submit that it would be (so far as we can tell) by applying exactly the criteria that scientists *do* apply to theory. These are very much the criteria one would expect scientists to use if theories are given a realist construal. They are *not* the criteria that an instrumentalist construal would lead one to anticipate. The coherence criteria, in particular, are realist in their implications.

V. The opposition

Why, then, would there be any objection to linking explanation and truth in the way that realists do? I will end by sketching very briefly the reasons why some prominent philosophers of science reject any such link, and add a few comments. van Fraassen, as we have already seen, takes the strongest line. "The belief involved in accepting a scientific theory is only that it 'saves the phenomena', that is, correctly describes what is observable."[20] Other criteria that affect theory-acceptance are purely pragmatic, and "pragmatic virtues do not give us any reason over and above the evidence of the empirical data for thinking that a theory is true." The aim of science is the construction of models with a view to prediction, "and not the discovery of truth concerning the unobservable." [21] His basic quarrel with realism is with its "inflationary metaphysics,"[22] against which he invokes the traditional "nominalist response."[23] Regularities in the world are brute fact; whether or not they can be explained in terms of unobservables, it "does not matter to the goodness of the theory, nor to our understanding of the world."[24]

There are some fundamental difficulties, I think, with the notion that the regularities in the world around us are brute fact; one remembers the sad tale of "grue," and how Goodman had to fall back on a notion of "entrenchment" to mark the boundary between lawlike and accidental generalizations. But what "entrenches" a generalization is its explainability in terms of a broader theory. This is surely a long way from brute fact. van Fraassen attempts a detailed nominalist response to this and other objections to his view, many of which were already formulated by critics of positivism.

Instead of pressing these objections, I will content myself with one very general counter-argument. Retroduction is constantly employed in everyday contexts, as well as in science, to infer to the existence of postulated causes of the effects to be explained. The existence of these causes is frequently verified directly at a later time thus (so far as one can see) justifying, to some extent at least, the confidence extended to the retroduction at an earlier stage. A favorite example of mine comes form the First Day of Galileo's *Dialogo*. Only the existence of lunar mountains, he argues in some detail, can explain the phenomena of changing shadows observed on the lunar surface. These mountains began, therefore, as theoretical entities, the fruit of inference from best explanation. Now that men have walked on them, their existence has presumably been established. Was belief in their existence unwarranted until the 1960's? Or was there not at least a partial warrant beforehand? The nominalist has difficulties in finding a consistent answer to the objection that direct observational tests very frequently bear out existence-claims originally derived from successful theoretical explanation. Does this not afford a reason to attach some credence to such claims even in the absence of a direct check? What is so special about direct observation that the likelihood of existence of mountains on the moon is negligible until the spaceships land? And if it is *not* negligible, it can have been based only on retroduction. The burden of proof is surely on the person who denies *any* connection between truth and explanatory power.[25]

Another philosopher who appears to deny this connection is Nancy Cartwright. In her recent book, *How the Laws of Physics Lie*, she argues that the explanatory power of scientific laws is gained at the expense of truth, so that in practice truth and explanatory success actually tend to *exclude* one another.[26] In arriving at this provocative conclusion, she makes a number of questionable assumptions about the nature of explanation in science. Though critical of the positivist *DN* model of explanation, she shares in at least some of its presuppositions. She assumes, for example, that the basic explanatory function in science is carried by laws, not by theories, and that the criterion of explanation is what she calls "organizing power." Explaining is thus a matter of subsuming under a lawlike generalization. The basic laws in physics are idealizations intended to "explain" in this sense, she goes on, but when they are applied to complicated real-world situations, many of them have to be amended in *ad hoc* ways. They are thus in the strict sense false; it is the attempt to give them a high degree of explanatory power that leads to this.

There are many difficulties, to my mind, in this argument. The basic explanatory statements in science are theories, not lawlike generalizations; retroduction rather than induction lies at the basis of scientific explanation. Furthermore, the techniques of idealization and of composition of causes do not, for the most part, falsify in the way she supposes.[27] Although there are instances where idealized laws have to be modified in *ad hoc* ways to make them fit concrete cases, these do not seem to be frequent, certainly not as frequent in physics generally as she assumes. And where such instances occur, the original law is regarded as defective even from the explanatory standpoint. The *ad hoc* character of its application to concrete cases—think of the renomalization techniques used in the quantum theory of the electron thirty years ago—count against it not only as truth, but also (in the scientist's eyes) as explanation.

The most serious challenge to the linking of explanatory power with truth may well be the Kuhnian one, based on the historical testimony of theory-change and theory-replacement. Kuhn comments on the "implausibility" of the claim that successive theories in science come closer to the truth or that truth lies in a match between theoretical entities and their counterparts in nature.[28] Laudan has developed this argument in some detail. He takes the realist to be linking truth and explanatory power in two respects, holding first that if a theory is approximately true, it will be explanatorily successful, and second that if it is explanatorily successful, it is pro-

bably approximately true.[29] He denies both of these conditionals, and argues that the notion of approximate truth is too vague in any case to carry weight in an argument of the sort the realist proposes. I have tried above to fill in this notion a little. It is worth recalling again that a *formal* measure of the degree of approximate truth of a successful theory is, in principle, impossible to specify.

Realists would not, I think, be committed to holding that if a theory is approximately true, it will *necessarily* be explanatorily successful. They might say that a theory is more *likely* to be explanatorily successful if it is approximately true than if it is false, that is, if nothing like the entities it proposes exists. But what the realist is more interested in is the claim that if a theory is explanatorily successful over a considerable time, it is probably approximately true. Laudan objects to this on historical grounds, adducing examples of successful theories which were later abandoned and thus cannot be held to have been approximately true.

My response to this objection would come in three stages, which cannot be developed here in the detail needed.[30] First, the realist insists on the qualification "probably," thus allowing for the possibility in some cases that a successful theory may be abandoned. Second, the historical instances Laudan cites are of very uneven value. In some cases (such as the crystalline spheres of medieval natural philosophers, or the theory of circular inertia) there was no long record of explanatory success in the sense in which this criterion was defined above. In others (such as the caloric theory), the earlier theory was reinterpreted, retaining much of the formalism. Each case would require separate discussion, and the value of Laudan's paper is that he has provoked such discussion.

But my main argument against Laudan's blanket rejection of the explanatory success-to-truth mode of argumentation is that it proves too much. Or as the medieval adage puts it: *Quod nimis probat, nihil probat*. (What proves too much proves nothing.) It is one thing to use the historical record to urge caution on anyone trying to formulate the realist claim precisely. But it is quite another to use it to reject the realist claim entirely, i.e. to maintain that explanatory success is *no* indication of truth. For this leaves us unable to make sense of the manifold instances when retroduction can be shown to have led to correct assessments of the truth. It prevents us from giving any sort of credence to existence-claims about even the best-supported theoretical entities. The historical record simply does not sustain so extreme a principle.

NOTES

1. A fuller discussion of Aristotle's enigmatic astronomical examples will be found in a preliminary version of this paper, "Truth and Explanatory Success", *Proceedings of the American Catholic Philosophical Association*, vol. 59 (1985), pp. 206-231.

2. Induction is an intermediate case. Scientists make use of guidelines (e.g. about curve-fitting) in the formulation of empirical generalizations, but the guidelines do not prescribe a unique form for the inferred law.

3. The point was, I think, first clearly made by Thomas Kuhn in his *Structure of Scientific Revolutions* (Chicago: University of Chicago Press, 1962).

4. For a fuller discussion, see my "Values in Science," in P.S. Asquith and T. Nickles (eds.), *PSA 1982* (E. Lansing: Philosophy of Science Association, 1983), pp. 3-25. See also Paul Thagard, "The Best Explanation: Criteria for Theory Choice," *The Journal of Philosophy*, vol. 75 (1978), pp. 76-92.

5. It is important to distinguish *proven* from *unproven* fertility as criteria. The former is a criterion of success, the latter a criterion of future promise. See my "The Fertility of Theory and the Unit for Appraisal in Science," in R. S. Cohen *et. al.* (eds.), *Boston Studies in the Philosophy of Science*, vol. 39 (Dordrecht: Reidel, 1976), pp. 395-432.

6. See my "Structural Explanation," *American Philosophical Quarterly*, vol. 15 (1978), pp. 139-147.

7. *Science, Perception and Reality* (New York: Humanities, 1962), p. 97. I would not agree with the further inference Sellars draws from this: "On the view I propose, the assertion that the micro-entities of physical theory really exist, goes hand in hand with the assertion that the macro-entities of the perceptible world do not really exist" (p. 96).

8. See *The Scientific Image* (Oxford: Clarendon Press, 1980), chapter 2.

9. In the introduction to his anthology, *Scientific Realism* (Berkeley: University of California Press, 1984), Jarrett Leplin lists ten theses associated with positions commonly called "scientific realism." He notes (p. 1) that no realist is likely to endorse a majority of them, even if they were subjected to reasonable qualification. And yet each of them can be found in the work of some realist.

10. I have developed this definition of scientific realism in some detail in "A Case for Scientific Realism," Leplin, *op. cit.*, pp. 8-40.

11. Wesley Salmon has developed this as an argument from a common cause; see "Why Ask: 'Why?'", *Proceedings of the American Philosophical Association*, vol. 51, (1978), pp. 683-705.

12. Cambridge: Cambridge University Press, 1983.

13. These quotations are from "Five Parables" in R. Rorty, J. B. Schneewind and Q. Skinner (eds.) *Philosophy in History,* (Cambridge: Cambridge University Press, 1984), pp. 118-9, where Hacking summarizes the thesis of his 1983 book.

14. *Representing and Intervening* (Cambridge: Cambridge University Press, 1983.) p. 262.

15. *Ibid.*, p. 254.

16. Following Dudley Shapere, "The Concept of Observation in Science and Philosophy", *Philosophy of Science*, vol. 49 (1982), pp. 485-525.

17. In a personal communication to the author, Hacking underlines this: "Since I'm saddled with my witticism, 'If you can spray them, then they are real', I might as well stay saddled. But do not attribute to me the thesis that if they are real, then you can spray them. I only mean to express a certain inductive scepticism—that if we never get around to intervening, then maybe what theory leads us to believe will in the end not pan out. That is a vastly weaker thesis, and one compatible with saying that we do have (imperfect) evidence for the existence of all sorts of things that we cannot use to do anything with, yet."

18. E. McMullin, "Galilean Idealization," *Studies in the History and Philosophy of Science*, vol. 16 (1985), pp. 247-273; see sec. 4

19. Gilbert Harman, "The Inference to the Best Explanation," *The Philosophical Review*, vol. 74 (1965), pp. 88-95; see p. 89.

20. *The Scientific Image*, *op. cit.*, p. 4.

21. *Ibid.*, p. 5.

22. *Ibid.*, p. 73.

23. *Ibid.*, p. 24.

24. *Ibid.*

25. The other source of van Fraassen's anti-realism is quantum mechanics. But it can surely be admitted that the realist thesis is more difficult to interpret and to apply in that very special domain without compromising the overall realist thrust of retroductive inference elsewhere in natural science.

26. New York: Oxford University Press, 1983, p. 56.

27. See McMullin, "Galilean Idealization", *op. cit.*

28. *The Structure of Scientific Revolutions, op. cit.*, p. 206.

29. "A Confutation of Convergent Realism," in Leplin (ed.), *op cit.*, pp. 218-249; see p. 228.

30. For a fuller criticism, see Clyde Hardin and Alexander Rosenberg, "In Defense of Convergent Realism," *Philosophy of Science*, vol. 49 (1982), pp. 604-615. See also the papers by Boyd, Leplin, and McMullin, in Leplin, *op. cit.*