[12] But excluding Leibniz, who "knew" there had to be some real, natural difference between right- and left-handedness.

[13] Iceland spar is an elegant example of how experimental phenomena persist even while theories about them undergo revolutions. Mariners brought calcite from Iceland to Scandinavia. Erasmus Bartholinus experimented with it and wrote it up in 1609. When you look through these beautiful crystals you see double, thanks to the so-called ordinary and extraordinary rays. Calcite is a natural polarizer. It was our entry to polarized light which for three hundred years was the chief route to improved theoretical and experimental understanding of light and then electromagnetism. The use of calcite in PEGGY II is a happy reminder of a great tradition.

[14] It also turns GaAs, a 3/4 to 1/4 left-hand/right-hand polarizer, into a 50-50 polarizer.

[15] I owe these examples to conversation with Roger Miller of SLAC.

[16] The concept of a "convincing experiment" is fundamental. Peter Gallison has done important work on this idea, studying European and American experiments on weak neutral currents conducted during the 1970s.

[17] I owe this information to Charles Sinclair of SLAC.

[18] Alexander Bain, *Logic, Deductive and Inductive* (London and New York, 1870), 362.

9

Explanation and Realism

Clark Glymour

I

One way to argue to a theory is to show that it provides a good explanation of a body of phenomena and, indeed, that it provides a better explanation than does any available alternative theory. This pattern of argument is not bounded by time or by subject matter. One can find such arguments in sociology, in psychometrics, in chemistry, and in astronomy, in the time of Copernicus and in the most recent scientific journals. The goodness of explanations is a ubiquitous criterion; in every scientific subject it forms one of the principal standards by which we decide what to believe. The ambition of philosophy of science is, or ought to be, to obtain from the literature of the sciences a plausible and precise theory of scientific reasoning and argument: a theory that will abstract general patterns from the concreta of debates over genes and spectra and fields and delinquency. A philosophical understanding of science should, therefore, give us an account of what explanations are and of why they are valued but, most important, it should also provide us with clear and plausible criteria for comparing the goodness of explanations. One half of the subject of this essay concerns a fragment of such a theory. I will try to describe, without gratuitous formality, some features that generate clear and powerful criteria for judging the goodness of explanations. That is half of my subject; the remainder concerns what such criteria determine about what we ought to believe. Both halves are prompted by Bas van Fraassen's recent and delightful book, *The Scientific Image*.

Wherever theorists have postulated features of the world that could not, at the time, be observed, debates have erupted over the scientific credentials of beliefs in the unobserved or unobservable. The scientific debates have more often than not been best articulated in throwaway lines: "If I were the Master," wrote Dumas in the 1830s,[1] "I would ban the word 'atom' from chemistry, for it goes beyond all experience, and never, in chemistry, should we go beyond experience." And in our own time, we have B. F. Skinner's argument by rhetorical question: When one has a secure equational linkage between observed variables, why introduce gratuitous unobserved variables? We also have more considered, philosophical discussions that lead to the same opinion, namely, that there is no warrant to be found in scientific observations for conclusions that concern unobserved or unobservable features of the world.

Call those who hold such views "antirealists," and those who hold the contrary view, "realists." Johannes Kepler, J. Dalton, and C. Spearman were realists. Andreas Osiander, F. Dumas, and Skinner were (and in the last case remains) antirealists. The antirealist case has been made with care and with something approaching logical precision by several writers, but still nowhere better than by C. G. Hempel who argued as follows:[2] Make the idealization which supposes that all the evidence pertinent to our theories can be formulated using only a part of the terminology of science. Call this part the "observational vocabulary." Then full-fledged theories whose vocabulary goes beyond this observational fragment seem to have no epistemological advantage over the collection of their consequences that can be stated in observational terms alone. For example, Hempel views explanation as a kind of deductive systematization, and the "observational consequences" of a theory are as well systematized—by Hempel's criteria anyway—as is the theory itself. (It must be noted that Hempel himself did not draw the antirealist conclusion, for he hoped that theories might be shown to afford an "inductive systematization" unobtainable without them.) Further, both Duhem and Quine[3] have argued that once we entertain hypotheses that transcend all possible experience, we enter into the realm of convention and arbitrariness, for the evidence and the canons of scientific method can only determine what we ought to believe about the observable, and the unobservable is indeterminate: a plethora of alternative conjectures will explain the evidence equally well. Claims of the same sort were made against Dalton's theory by his anti-atomist opponents—the properties of atoms, they said, are indeterminate and arbitrary.

Duhem, Quine, and others are right enough if we allow science only a sufficiently impoverished collection of principles of assessment and inference. But even the fragmentary criteria for comparing scientific explanations which I am able to display are strong enough to defeat their antirealist arguments. For although the criteria make no use of the notion of observability, they yield the result that sometimes the best explanation *does* go well beyond what is observed or observable. Or, to put it in something more like Hempel's fashion, the criteria yield the result that sometimes a theory with "nonobservation terms" has explanatory virtues that are unobtainable without such terms. Moreover, the criteria for comparing explanations do not, in leading us beyond the observable, enmesh us in interdeterminacy. On the contrary, in some contexts they suffice to determine a unique, best theory which explains the phenomena. Further, there is nothing about these criteria for the goodness of explanations which requires or presupposes that they be applied only to theories that contain claims about what is unobservable. The very same criteria are used to determine the best of competing theories about observable features of the world. The same criteria sometimes determine—at least as far as I understand the notion of 'observability'—that theories that are confined to observable features of the world are better explainers than competitors that postulate unobservable features. This all seems to me exactly as it should be. The result is simply that the same features of inference which lead to general conclusions about the observable also lead in other contexts to determinate conclusions about the unobservable. In consequence, I believe there are only two ways to maintain the antirealist position: either by impoverishing (perhaps I should say emasculating) the methods of science, and disallowing altogether explanatory criteria such as those I will describe, or by arbitrarily (and vaguely) restricting the scope of application of such principles to the realm of the observable.

II

In discussions of scientific explanation, and of the grounds such explanations may afford for belief in unobserved features of the world, it is important to keep actual cases in mind. Eventually, I will try to show that all of the following cases have something important in common:

1. The Copernican explanation of the regularities of the superior planets, in particular the regularity noted by Ptolemy, that if in a whole number of solar years a superior planet goes through a whole number of oppositions and also a whole number of revolutions of longitude, then the number of solar years equals the number of oppositions plus the number of revolutions of longitude. Copernican theory explains the regularity by noting that the number of solar years is equal to the number of revolutions the earth has made around the sun, and that the number of

oppositions is equal to the number of times the faster moving earth has overtaken the superior planet moving more slowly in an orbit outside the earth's orbit, and also that the number of revolutions of longitude equals the number of revolutions the superior planet has made about the sun.

2. The Daltonian explanation of the law of definite proportions. The law asserts that in any two cases in which quantities of the same pure chemical reagents combine to produce quantities of the same products, the ratios of the combining weights of the reactants are the same. Dalton's explanation is that any sample of a pure chemical substance consists of molecules, each having the same weight as any other, and that each molecule of a given kind is composed of the same numbers of atoms of various kinds. All atoms of the same kind have the same weight, and a chemical reaction is just a process in which the constituent atoms of the molecules in the reagents are recombined to form molecules of the products. Thus, the ratios of the numbers of molecules of the various reactants that combine are invariant and characteristic of the reaction. Two different samples of reactants with different weights will differ in the number of molecules they contain, but since the weight of any pure sample is just the sum of the weights of its constituent molecules, the ratio of the weights of the reactants will equal the ratio of the number of molecules they severally contain.

3. The general relativistic explanation of the anomalous motion of Mercury's perihelion, and of the deflection of starlight passing near the limb of the sun. The explanations go roughly as follows: for weak gravitational fields, such as those of the planets at astronomical distances, the theory of relativity closely approximates Newtonian gravitational theory. Hence, in the case of Mercury as well as in the case of the starlight, the only significant contribution to the gravitational field is that of the sun, which is known of to be close to spherical. Therefore, the gravitational field is, to good approximation, the gravitational field of a single mass point. From the field equations of the theory, the gravitational field, in empty space, of a single mass point is uniquely determined and describes the geometry of the surrounding space-time. The equations of motion of the theory imply that light rays move on null geodesics of this geometry and massive particles move on timelike geodesics. Light from a star may be described as a ray, and the planet Mercury reacts to gravity (to good approximation) as a point mass. Given the initial conditions, the phenomena follow.

4. Spearman's explanation of the correlations among intelligence tests. Although the history is complicated, it is probably fair to say that Spearman invented factor analysis when, in 1904, he published two papers, one on the measurement of correlations, and another on the measurement of intelligence.[4] Spearman obtained several assessments of populations of school children, including assessments of "intelligence" and of sensory discrimination. Analyzing his data from 1904 and later, he argued in his *Analysis of Human Abilities,* that for any four such measures, say X_1, X_2, X_3, X_4, the correlations among the measures have vanishing "tetrad differences." That is to say, empirically it is found that for any such quadruple of measures for his samples:

$$\rho_{12}\rho_{34} = \rho_{13}\rho_{24}$$
$$\rho_{13}\rho_{24} = \rho_{14}\rho_{23}$$

where ρ_{ij} is the correlation between X_i and X_j. Spearman claimed that the best explanation of these equations is that there is a common factor which determines the outcomes of scores on all of the measures, and he labeled that factor "general intelligence."

These aforementioned cases are not essential ones for any theory of explanation, but as cases they have important virtues. They represent a range of subject matters, they are explanations that have a genuine historical role, and they were not used to satisfy idle curiosity—as with questions about soap bubbles—but were instead crucial pieces of the arguments given for the respective theories. They also have no uniform connection with distinctions between what is observable and what is not. The Copernican theory postulated features—the motions of the planets in three-dimensional space—that could not be observed in the sixteenth century. So did competing theories. Yet, some of these features might arguably be regarded as observable today. Dalton's atomic theory certainly did concern features of the world that have served almost as paradigm cases of the unobservable. I have no idea whether the metric field of space-time ought to count as observable or as unobservable. Spearman's "general intelligence" is surely an unobservable feature of persons—more carefully, since one may doubt there is any such thing, if there were a factor such as Spearman's general intelligence, it would be unobservable. The phenomena explained in these cases can be viewed sometimes as regularities, sometimes as particular events, or sequences of events, and sometimes as both. The regularity of the superior planets is indeed a regularity, but Copernician theory also explains all of the instances of the phenomenon. The same holds of definite proportions and the motion of Mercury.

The explanation of the deflection of light was, in fact, the explanation of a few events, but there is a corresponding regularity that the theory explains. Spearman's equations neither describe particular events nor express generalizations; I suppose they are best understood as expressing particular features of the population tested. Finally, for each of these explanations, there was available an alternative theory generating competing explanations. The gravamen of the scientific arguments was that the explanations cited were *better* than any others available. Copernican explanations were compared with Ptolemaic explanations, general relativistic explanations with classical ones, and one-factor accounts of correlations were compared with multifactor accounts.

III

I view one of the chief goals of a philosophical account of aspects of scientific explanation as that of providing a canon or canons for the assessment of scientific theories. To the best of my knowledge, the philosophical literature on scientific explanation provides no clear criteria that can serve to explain why the four cases I have cited might have been viewed as good explanations or, in particular, why they should have been viewed as better than competing explanations. Nonetheless, what I shall have to offer by way of criteria is closely connected with some philosophical articulations of the notion of scientific explanation.

Philosophical theories of scientific explanation can be roughly divided into three types: purely logical theories, which analyze explanation solely in terms of logical relations and truth conditions; theories with extra objective structure, which impose objective conditions on explanation beyond those of truth and logical structure; and theories with extra subjective structure, which impose on explanations psychological conditions of belief, interest, and so forth.[5]

The apparent aim of Hempel and Oppenheim[6] was to provide an account of the logical structure of 'explains' in much the way that the logical tradition of Frege, Russell, Whitehead, and Hilbert had provided accounts of the logical structure of "is a proof of." The conditions given by Hempel and by Oppenheim were not intended as an analysis of when it is appropriate to say that someone has explained something to someone; instead, they were intended to specify the logical structure that fully explicit, nonstatistical explanations in the natural sciences would typically have if there were any, and which actual explanations in the natural sciences typically abbreviate. If Hempel and Oppenheim had been successful in their aim, their criteria would have formulated a critical standard for judging theories and hypotheses. On the basis of logical structure alone, it would be fully determinate what singular sentences a given theory could potentially explain, and what sentences it could not possibly explain. Combined with a further account of how to use information about what theories can and cannot explain in order to assess those theories, a logical account such as Hempel and Oppenheim's promised to provide us with an understanding of how explanation is used in deciding what to accept and believe.

Hempel and Oppenheim's account of the logical structure of deductive explanations turned out to be altogether unequal to its task. In effect, given an *arbitrary* true (but not logically true) sentence *E* and an *arbitrary* universally quantified sentence *S* which is not a logical truth, there exists a logical consequence of *S* and a true singular sentence *C*, such that the logical consequence of *S* and *C* together constitute the premises of a potential explanation of *E*.[7] In effect, anything explains anything. What was important in Hempel and Oppenheim's work was not the execution but the vision. If one shared the vision, the natural response to the failure of their analysis is to attempt another of the same kind, either in entirely logical terms or, perhaps, with extra structure.

Several purely logical replacements for Hempel and Oppenheim's analysis have been published. While not all of them are trivial, they are, without exception, far too weak to provide useful criteria for theory assessment. Perhaps the best attempt is David Kaplan's,[8] who uses conditions of truth as well as logical structure to constrain explanation.

Accounts such as Kaplan's are deficient in tolerating as explanations a range of cases that we would dismiss as not explanatory. For any explanation of the form:

$$\begin{array}{c} \forall x\, Dx \\ \hline Da \end{array} \tag{1}$$

if *P* is any property which the individual *a* also has, then

$$\begin{array}{c} \forall x\,(Px \rightarrow Dx) \\ Pa \\ \hline Da \end{array} \tag{2}$$

is an explanation meeting Kaplan's criteria and likewise various generalizations of Kaplan's criteria. Again, for any explanation of the form:

$$\begin{array}{c} \forall x\,(Fx \rightarrow Gx) \\ Fa \\ \hline Ga \end{array} \tag{3}$$

if *P* is any property which the individual *a* also has, then

$$\begin{array}{c} \forall x\,(Fx \,\&\, Px \rightarrow Gx) \\ Fa \,\&\, Pa \\ \hline Ga \end{array} \tag{4}$$

is likewise an explanation meeting Kaplan's and related criteria. Thus, to use an example of Henry Kyburg's, we can explain the fact that a sample of table salt dissolves in water by citing the (true) claim that the sample of table salt has been hexed, and the (true) generalization that all hexed table salt dissolves in water.

One sensible response to these difficulties is that they show nothing more than the incompleteness of the analysis of explanation. Thus, to an account of Kaplan's kind, restricting the logical form of explanations, one should add criteria for comparing explanations with the appropriate logical form. For example: when an explanation of form (1) exists, any explanation of form (2) is defeated. Alternatively, one might reconstruct the logical conditions for explanation so that (2) and (4) will not count as explanations at all. Wesley Salmon proposes to regard deductive explanation as a limiting case of statistical explanation:[9] explaining a single event consists in assigning it to one member of a statistically homogeneous partition of the largest class of events that can be so partitioned. Very cleverly, it follows that when (1) obtains, (2) does not meet the conditions for explanation (unless *P* and *D* are coextensive), and analogously with (3) and (4).

None of this work on explanation provides any criteria that can help to account for the cases mentioned earlier, or for others like them. Ptolemaic and Copernican astronomical theories each provided deductions of sentences about the positions of the planets from general lawlike sentences; if that is what is required for potential explanation, then both theories provided it. The atomic theory explained each instance of the law of definite proportions, but so far as these logical criteria are concerned, equally good explanations were obtained from the view, popular in the nineteenth century, that one ought to abjure atoms and simply make tables of combining weights. Similar remarks hold for the other cases. Salmon's account of explanation, it is fair to say, is such that we simply do not know how to apply it in these sorts of contexts.

A simple way to see the limitations of the deductive criteria described so far is to consider what it would be to explain a regularity rather than a single event. One naturally supposes that to explain a regularity is to explain every instance of it, but if no more is required, then every regularity explains itself. If we require further that what explains a regularity not be logically implied by the regularity itself, we are still no better off, as Hempel and Oppenheim noted, since it will remain the case that by these criteria any regularity can be explained by conjoining that regularity with any other regularity logically independent of it.

I do not know of any writer who has attempted to account for the explanation of regularities in logical terms alone. Inevitably, considerations other than truth and logical form enter into the criteria, and that is quite proper and harmless so long as the further structure does not defeat the very goal of an account of explanation—namely, to help us to understand how explanation can be and is used to assess our hypotheses. Three recent discussions of explanation that appeal to extra structure are especially valuable, not so much for any technical advance they make, but for pointing to features of explanation easily overlooked, and for at least suggesting that such features might be formally tractable. One is Michael Friedman's proposal that what good theoretical explanations do is to somehow reduce the number of hypotheses that must be accepted independently of one another.[10] Another is Robert Causey's careful analysis of the structure of intertheoretical reductions,[11] with its emphasis on the identity of objects and properties described and postulated by the hypotheses to be explained with complexes of objects and properties postulated by the theory that provides the explanation. A third is Baruch Brody's attempt to bring Aristotle back to scientific explanation,[12] and to take seriously the idea that good scientific explanations show that the event or regularity to be explained is *necessary*, not just a necessary consequence of theoretical assumptions. None of these three accounts suffices to treat the sorts of cases I described earlier, and only Friedman's really constitutes an attempt to do so. However, Friedman's technical account came apart; Causey's analysis has technical difficulties and does not provide applicable comparative criteria; and Brody's account of explanation can be applied to compare competing claims to explain a phenomenon only in those situations where we know, independently, something about what properties are essential and what properties are accidental. Even so, in each of these three cases the motives seem sound in their own way, and the scheme that I present below can be viewed as an attempt to unify aspects of each of them.

IV

One of the things that many explanations seem to do, and which appeals to certain intuitions about explanation, is to demonstrate that one phenomenon is really just a variant of, a different manifestation of, another phenomenon. In some sciences, the explanations provided in textbooks are very largely of this kind. In classical thermodynamics, for example, the standard exercise repeated throughout textbooks of physical chemistry consists in using thermodynamic principles to derive empirical regularities from one another. The first and second laws are used to derive Joule's law from the ideal gas law, or to derive from the gas law the fact that specific heat at constant pressure minus specific heat at constant volume is proportional to the gas constant. Intuitively, an explanation produces a form of understanding if it shows us that the phenomenon we want explained is a manifestation of a different phenomenon we already know about. The understanding consists in a grasp of the unity of pattern or substance behind the apparently disparate phenomena. In physical contexts, the mechanics of such explanations is familiar to anyone who has suffered through problem sets. In the thermodynamic case again, one starts, say with $C_p - C_v = R$ and explains this relation by relating a subformula of it (e.g., $C_p - C_v$) to thermodynamic quantities, applying the thermodynamic laws to these quantities, and deriving the result that $C_p - C_v = PV/nT$.[13]

As a first approximation to a formal account of this feature of explanation, consider theories represented in the predicate calculus. For convenience, let us suppose that all sentences considered are in prenex form. Let a *subformula* of a sentence S be any well-formed subformula occurring in the matrix of the prenex form which is not logically equivalent to the entire sentence. A subformula of a sentence may have several occurrences in a sentence. By a term in S we will mean any term occurring in the matrix of S. Let us say that where T, H, and K are sentences, *T explains H as a result of K* if for (nonvalid) subformulas $H_1, \ldots, H_n$ of H, no one of which occurs in H as a subformula of any of the others, and terms $t_1, \ldots, t_k$ of H, not occurring in $H_1, \ldots, H_n$, there are formulas $K_1, \ldots, K_n$, terms $s_1, \ldots, s_k$, and subtheories $T_1, \ldots, T_{n+k}$, of T such that

(i) $T_i \vdash H_i \leftrightarrow K_i$
(ii) $T_{n+j} \vdash t_j = s_j$
(iii) $T_i \vdash H$ for any $i < n+k$
(iv) If $H(H_i\, t_j / K_i\, s_j)$ is the result of simultaneously substituting, for all i and j, K_i for every occurrence of H_i in H, and s_j for every occurrence of t_j in H, then $K \vdash H(H_i\, t_j / K_i\, s_j)$.

The third condition above is both natural and necessary to eliminate certain spurious explanations. For example, one could not use any theory including the ideal gas law to explain that very law as a form of the relation $T = T$ (where T is the temperature) by starting with $PV = nRT$, and noting that from the theory it follows that $PV/nR = T$ and that on substituting T for PV/nR, in the ideal gas law, one obtains $T = T$. Such an explanation is excluded by condition (iii). The condition that a term substituted for not occur as part of a subformula substituted for is perhaps unduly restrictive, but is intended to avoid complications.

In practice, we do not usually consider first-order formulas but rather equations representing theories which are to provide the explanations. In such a context, T explains H as a result of K, where $H(X_1, \ldots, X_n) = 0$ and $K(Y_1, \ldots, Y_m) = 0$ are equations and T a system of equations, if there are consequences T_i of T, not entailing H, and each T_i entails an equation between a combination of the Xs occurring in H and a combination of Ys, and when the combinations of Ys are substituted in H for the corresponding combination of Xs, one obtains an equation that is satisfied whenever K is.

As I have presented these conditions, they do not satisfy the condition of logical equivalence; that is to say, T can explain H as a result of K while for logically equivalent formulas T', H', and K', T' fails to explain H' as a result of K'. Of course, what I mean is that T explains H as a result of K if each of these sentences has equivalents meeting the formal conditions. In fact, the equivalence is generally much wider than logical equivalence, for we count an equation H as explained as a result of K if we can find mathematical equivalents of H and K meeting the conditions. Doubtless, the formal proposal has other defects as well; it is, at best, a first try. I hope that veterans of the textbook tradition in physics and chemistry at least will be tingled by a reminiscence that much of the work in the explanations presented and the problems assigned was to find the right representations of equations, and the right subtheories, in order to derive one thing from another.

A theory may entail each member of a class of regularities without explaining any one of them as a result of any other of them. If, for example, the regularities in question are logically (mathematically) independent, the theory consisting of their conjunction will have this property. Thus, to borrow an example of Friedman's, the theory consisting of the conjunction of Kepler's laws, Boyle's law, Galileo's law of falling bodies, and the law of the pendulum will not explain any one of these regularities as the result of any other. Again, the conjunction of the additivity of masses and the law of definite proportions does not explain either of these regularities by the other. By contrast, the atomic theory explains the law of definite proportions as a result of the additivity of masses, and that unification is one of the great virtues of the theory. In some contexts in which theories are being assessed and compared, there is a reasonably well-established collection of logically independent empirical regularities pertinent to the assessment. Contending theories, such as the atomic theory and the theory of equivalents, may well entail a common class of regularities, but the classes of regularities pertinent to the respective theories need not be coextensive: the additivity of masses has nothing to do with the theory of equivalents but a great deal to do with the theory of atoms. In general, we prefer theories that explain entailed regularities as the result of other established regularities to theories that do not. More exactly, I propose the following rather modest principle of comparison:

1. *Ceteris Paribus*, if T and Q are theories and for every established pair of regularities, H, K, such that Q explains H as a result of K, T also explains H as a result of K, but there exist established regularities, L, J, such that T explains J as a result of L but Q does not explain J as the result of any other established regularity, T is preferable to Q.

The statement is cumbersome, but the idea is almost trivial, and it seems to me to be part of what is correct about Friedman's suggestion that good theories reduce the number of independently acceptable hypotheses. It might be interesting to explore further principles of comparison that would extend to circumstances in which competing theories explain a common set of regularities as the result of other regularities, but disagree about which regularities are the result of which other regularities. I leave the matter aside here, except for a special case to be discussed later.

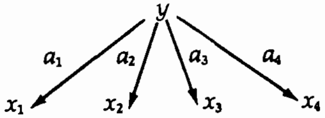
There are two points to emphasize. First, according to principle 1, certain theories entailing empirical regularities may be entirely preferable to the conjunction of those regularities alone, and may be preferable as well to the conjunction of the regularities entailed with any regularities from which these are claimed to result. The atomic theory illustrates the point. Second, according to principle 1, Copernican theory is preferable to its principal rival, Ptolemaic astronomy, and Spearman's explanation of the correlation equations among four measured variables in terms of a single common factor is preferable to many alternative explanations that equally save the phenomena. To see this, it will help to return to Brody's idea.

At least part of Brody's idea is that we may explain a state of affairs by showing it to be necessary, and that such a demonstration is to be given by showing that the state of affairs is a logical consequence of essential properties of the entities concerned, and of laws governing these essential properties. I suggest a different kind of demonstration of necessity: we may show that a regularity has a kind of necessity by explaining it as a result of a necessary truth. In particular, in the schema for 'T explains H as a result of K', nothing prohibits that K be a logical or mathematical, and thus a necessary, truth. Logical and mathematical truths, in general, may be considered regularities, and they may certainly be established and used in the assessment of theories. In the case of Copernican positional astronomy, and in the case of Spearman's tetrad equations, regularities or equations of an empirical kind are explained as the result of mathematical truths.

It is a mathematical truth that if two objects move in closed orbits

about a common center, one orbit inside the other, and the inner object moves faster than the outer object, then the number of revolutions of the inner object equals the number of revolutions of the outer object plus the number of times the inner object has overtaken the outer object, whenever all of these numbers are whole. According to Copernican theory, the Earth moves in an orbit interior to the orbit of any superior planet, and moves faster in its orbit than does a superior planet. The number of solar years equals the number of revolutions of the Earth in its orbit; the number of oppositions or cycles of anomaly equals the number of times the superior planet is overtaken by the Earth, and (over any long period) the number of revolutions of longitude of a superior planet equals the number of revolutions in its orbit. Thus, the planetary regularity noted by Ptolemy is explained by Copernican theory as the result of a mathematical truth. It is not explained in Ptolemaic theory as the result of any other independently established regularity.

Spearman assumes (tacitly) that the relations between measured variables and unmeasured factors are linear. Assuming that measurement error averages to zero and that errors in measured variables are uncorrelated with one another, it follows that for standardized variables, the correlation between two variables is equal to the product of the linear coefficients relating each variable of the pair to the common factor. Thus, with Spearman's causal scheme



we have that

$$\rho_{ij} = a_i a_j.$$

On substituting the right-hand side of these equations into Spearman's supposed empirical equations:

$$\rho_{12}\rho_{34} = \rho_{13}\rho_{24}$$
$$\rho_{13}\rho_{24} = \rho_{14}\rho_{23}$$

we obtain instances of trivial truths of real algebra. Thus, Spearman's tetrad equations are explained as results of mathematical truths. Typically, other arrangements of unmeasured factors can be found which are consistent with the tetrad equations but do not explain the tetrad equations as the results of mathematical truths.

One might wonder whether, in the Copernican and psychometrical cases, an empirical regularity has been entirely explained as the result of necessary truths, for the Copernican relations between numbers of orbits and revolutions in longitude and solar years, and so forth are at least arguably not necessary truths. Again, psychometric equations relating correlation coefficients to "factor leadings" scarcely seem to be necessary truths. There are cases and contexts, however, in which an empirical relation seems literally to be explained *as* a necessary truth. I have in mind cases in which a regularity is explained as the result of a mathematical truth, and the equations used to connect quantities in the regularity with other quantities (as in [ii] and [iii] of the scheme) are identifications, as Professor Causey requires for microreductions. One case in which there is a tradition—correct or not—providing such an interpretation is that of general relativity. Both the motion of Mercury and the motion of light rays near the limb of the sun provide special cases of the general relativistic law of motion. That law of motion is, in turn, a consequence of the general relativistic conservation laws, which require that the covariant divergence of the energy-momentum tensor vanish. The field equations of the theory give the energy-momentum tensor as a function of the metric field; as early as 1920, Eddington proposed that the field equations be understood as an identification and not as a contingent relation: these equations, on Eddington's construal, specify what matter-energy *is*. When a function of the metric field which equals the momentum-energy tensor is substituted for the momentum-energy tensor in the conservation laws, the conservation laws are transformed into mathematical truths of the tensor calculus.

I surmise that, other things being equal, we prefer explanations that explain regularities in terms of necessary truths to explanations that explain those same regularities in terms of other empirical generalizations. I suppose, further, that we find particularly virtuous those explanations that explain regularities as the result of necessary truths and do so by means of connections that are themselves necessary. When such explanations succeed, they are complete, and nothing remains in need of explanation. I do not mean to suggest that any or all of these considerations exhaust the criteria by which we judge explanations. That is why there are *ceteris paribus* clauses. It was, of course, true that one of the major reasons for the inferiority of the Newtonian explanations of the perihelion advance and light deflection was that all of these explanations required hypothetical circumstances—hidden masses and concentrations of ether—for which there was no evidence whatsoever.

V

The explanatory principle provides a reason why, in each of the cases described in section II, the explanation is particularly virtuous, and why it is better than competing explanations. In some of these cases, there is even a little historical evidence that these virtues were recognized in something like the form they are given here: Kepler praised Copernican theory in contrast to Ptolemaic astronomy because the former makes the regularities of the planets a "mathematical necessity." In each of the cases considered, the explanatory power is obtained by introducing theoretical connections and structure not explicit in the regularities to be explained, and in each of these cases the connections among empirical regularities could not be made at all without additional theoretical structure. The regularity of the superior planets simply cannot be explained as the result of a mathematical truth, unless the quantities occurring in the statement of that regularity are somehow related to other quantities distinct from solar years, oppositions, and revolutions of longitude.

The same is true of Spearman's tetrad equations and of the motion of Mercury, and so on. The law of definite proportions cannot be explained as a result of the additivity of mass, unless one introduces objects other than the macroscopic samples of elements that were used in the nineteenth-century laboratory. In some of these cases, the additional structure that must be introduced to obtain the explanatory connections is structure that has turned out to be observable; in other cases, that structure has not turned out to be unobservable, or at least arguably unobservable. But observability or nonobservability has nothing to do with the explanatory virtue in question. There seem to be no grounds, save arbitrary ones, for using a principle of preference to give credence to theoretical claims when the claims turn out to be about observable features of the world, but to withhold credence when the theoretical claims obtained by the very same principle turn out to be about unobservable features of the world. The argument for atoms is as good as the argument for orbits.

I know of three objections to this argument: first, that explanation is subjective and context-dependent and therefore cannot be an objective and interpersonal ground for belief; second, that the principle of preference presented in section IV may be acceptable as a principle governing what theories we should prefer to accept or display or work with, but it cannot be acceptable as a principle governing what theories we should prefer to *believe*; and third, that the introduction of theoretical structure beyond the bounds of observation generates underdetermination without limit.

Many writers have maintained that explanation is basically a pragmatic notion. The fundamental explanatory relation is held to be something of the form '*X* explains *Y* for person *P*'. Of course, what will explain *Y* for person *P* will depend on what *P* already knows about *Y*, does not know about *Y*, wants to know about *Y*, and so on. This view of explanation is surrounded by examples trading on the ambiguities of "why questions," and the context dependence of relevant replies to questions of the kind "what causes *Y*?" Undoubtedly, understanding is a pragmatic notion, and also, undoubtedly, we legitimately call "explanations" those replies to "why questions" (and questions of other forms) that produce understanding in an audience. None of this has any pertinence to whether or not there are objective, nonpragmatic relations between theories and statements of empirical phenomena, relations which when apprehended produce understanding and give grounds for belief. The claim that explanation is *entirely* interest-dependent has exactly as much merit as the claim that 'proof' is entirely interest-dependent: based on what one's mathematical audience knows or believes, one says some things in giving proofs, but not other things; there are arguments that meet the formal conditions for proof but which we do not call proofs because they are uninteresting (for example, because we have no reason to believe the premises, as in the "proof" of the Poincaré conjecture from the premises that (1) Glymour weighs more than 190 pounds and (2) if Glymour weighs more than 190 pounds then the Poincaré conjecture is true). An account of "proof" entirely tied to speech acts and eschewing characterization of ideal logical form would be blinding; so would a like account of explanation.

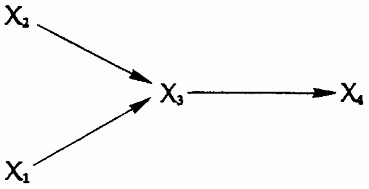
Bas van Fraassen has argued that preferences of the kind to which I have appealed in principle 1 cannot be credences or preferences as to what to believe. The reason is as follows: According to principles such as 1, Copernican theory, for example, is to be preferred to the conjunction of the empirical laws which it explains. But the conjunction of the empirical laws that Copernican theory explains is a *logical consequence* of Copernican theory. We cannot reasonably give a claim more credence than its logical consequences. Hence, the preferences in principle 1 cannot be credences or preferences for belief.

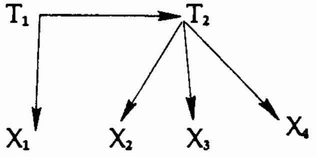
Van Fraassen is right, so far as this goes, but the preferences may, nonetheless, be tied to belief. It can be irrational to believe a collection of empirical laws, to believe that a particular theory entails these regularities, is tested by them, and explains them better than any other possible theory, and still not believe the theory. The irrationality is akin to that of believing certain premises, believing that some other sentence is a logical consequence of these premises, and refusing nonetheless to believe the further sentence. There is inductive closure quite as much as deductive closure. The preference stated in principle 1 is a preference for a certain candidate for the inductive closure of the regularities explained. If theory *T* is preferable to theory *Q*, according to principle 1, that is not to be understood to mean that *T* is to be believed rather than *Q*, or that there is reason to believe *T* rather than *Q*, for *Q* may be a logical consequence of *T*. Rather, the preference is to be understood to mean that the inductive closure of the laws explained is (the deductive closure of) theory *T* rather than (the deductive closure of) theory *Q*. More simply but less exactly, the preference is for believing *T* rather than believing *merely* *Q*, when *T* and *Q* are consistent, or for believing *T* rather than *Q* when *T* and *Q* are not consistent.

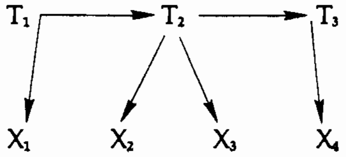
Some years ago, Quine argued that once we introduce language which does not describe observable features of the world, underdetermination is rampant. "Rampant" means infinite here, for if only a finite number of theories are consistent with all possible evidence of some kind and satisfy our inductive canons, then there is a preferred conclusion, namely their disjunction. Such sweeping views are unconvincing, for they lack both a convincing account of theoretical equivalence and a plausible characterization of inductive canons. In more restricted cases, where a characterization of the class of possible theories is available and clear and applicable criteria for sameness of theories are at hand, familiar arguments for underdetermination fail when inductive canons are applied. For example, H. Reichenbach's well-known arguments for the underdetermination of space-time geometry fail when the inductive canons I prefer are applied to space-time theories.

It is certainly the case that the principle of explanation that I have described, even with elaborations about necessity and identity, is insufficient as a foundation for inductive logic, and if that principle were the whole of the matter, underdetermination would indeed be rampant. Nonetheless, principle 1 in conjunction with other available inductive principles, can be extremely powerful, and can in some contexts virtually eliminate underdetermination—or so I conjecture. Consider the context provided by Spearman's kind of theory, where the data consist of correlations among some determinate set of variables. The theories consist of any set of digraphs without cycles, connecting the measured variables with one another or with unmeasured variables, and the associated linear equations. Under the assumptions about error made before the equation, they can be reduced to equations relating the "path coefficients" to one another and to the measured correlations. The correlations may satisfy equations such as Spearman's tetrad equations, or analogous equations involving triples of correlation coefficients, quadruples of correlation coefficients, and so on. Given a set of data satisfying a determinate set of equations of this kind, of the infinity of digraphs consistent with this data, a proper subset will reduce to mathematical truths all and only the equations of the set satisfied by the data. This set of digraphs will be those preferred by principle 1.

In some cases the set of preferred digraphs will, I believe, be finite and small. For example, in addition to the digraph illustrated in connection with Spearman's explanation of his correlation equations, I know of only three kinds of directed graphs that satisfy the principle for the three tetrad equations Spearman assumes holds empirically. They are







The first digraph, which contains no unmeasured variables, really stands for an entire set of digraphs, since any two variables can be interchanged in the digraph without affecting the derivation of the tetrad equations. Any of the digraphs of this kind can be distinguished from those which introduce latent variables, since digraphs of this kind imply restrictions on partial correlations that are not implied by digraphs with latent variables. Various other empirical constraints may distinguish the three alternatives with unmeasured variables from one another in the context of a larger theory. Even when other empirical constraints are not available, other methodological constraints may be. In *Theory and Evidence*, I described a strategy for testing systems of algebraic equations in real variables with known coefficients. In the context of linear theories of the kind under consideration here, principles founded on bootstrap testing may not themselves lead us to prefer theories with unmeasured or latent variables, but they may lead to determinate preferences among a set of theories all having latent variables. Starting with Spearman's supposed tetrad

equations, we have four classes of theoretical digraphs that satisfy principle 1. Partial correlations can either establish or eliminate the class of digraphs containing only measured variables. Of the remaining three digraphs, Spearman's original model is the one preferred on bootstrap principles in the absence of other information. That is because the other two digraphs implicitly contain theoretical coefficients which cannot be evaluated from measured correlations even assuming the correctness of the causal relations they postulate. In the econometricians' jargon, they are not "identifiable."

This is only an example, and a conjectural example at that. I mean it to illustrate nothing more than that there may well be many surprises in store when issues of underdetermination are examined in more modest and more detailed ways. Certainly, there is at present no particular reason to believe that underdetermination is rampant once we move beyond the observable. Underdetermination is, I expect, indestructible in another way, for logic does not determine a unique set of inductive canons, perhaps not any set of inductive canons. Thus, antirealists may still have recourse to denying principles of preference or inference such as those I have tried to describe here. My point has not only been to give that description, but to argue that the principles I have described are principles we use in our sciences to draw conclusions about the observable as well as about the unobservable. If such principles are abandoned *tout court*, the result will not be a simple scientific antirealism about the unobservable; it will be instead a not so simple skepticism.

NOTES

1. J. Dumas, *Lecons de Philosophie Chemique* (Paris: Gauthier Villars, 1937), 178-179. This edition is a reprinting (with slight alteration of title) of the 1837 edition.

2. C. Hempel, "The Theoretician's Dilemma," in *Aspects of Scientific Explanation* (Glencoe: Free Press, 1965).

3. Cf. P. Duhem, *The Aim and Structure of Physical Theory* (New York: Atheneum, 1974), and W. V. O. Quine, "Two Dogmas of Empiricism," in *From A Logical Point of View* (Cambridge: Harvard University Press, 1953).

4. C. Spearman, "General Intelligence Objectively Determined and Measured," *American Journal of Psychology* 15 (1904): 201-293.

5. In the first class fall the theories of Hempel and Oppenheim, D. Kaplan, J. Kim, and more recently, J. Thorpe, B. Cupples, and others. In the second fall theories such as those of Salmon, Brody, Friedman, and Causey, and the third class includes theories such as those of van Fraassen, B. Skyrms (I believe), P. Achinstein and, recently, Putnam.

6. C. Hempel and P. Oppenheim, "Studies in the Logic of Explanation," in *Aspects of Scientific Explanation*.

7. Cf. R. Eberle, D. Kaplan, and R. Montague, "Hempel and Oppenheim on Explanation," *Philosophy of Science* 28 (1961): 418-428.

8. D. Kaplan, "Explanation Revisited," *Philosophy of Science* 28 (1961): 429-436.

9. Cf. W. Salmon, *Statistical Explanation and Statistical Relevance* (Pittsburgh: University of Pittsburgh Press, 1971).

10. M. Friedman, "Explanation and Scientific Understanding," *Journal of Philosophy* 72 (1974): 5-19.

11. R. Causey, *Unity of Science* (Dordrecht: D. Reidel, 1979).

12. B. Brody, *Identity and Essence* (Princeton: Princeton University Press, 1980).

13. I am indebted to my student, Jeanne Kim, for suggesting this example to me.

10

Truth and Scientific Progress

Jarrett Leplin

Philosophers of science are increasingly taken with the following apparent paradox: theories that excelled under criteria now employed in evaluating theories ultimately proved unacceptable; therefore, it is likely that the best current theories, and even better ones that we might imagine overcoming what defects we now recognize in current theories, will prove unacceptable. Thus, we have a strong inductive argument against the ultimate acceptability of theories that we have strong inductive grounds to accept.[1] It is common to summarize this situation by some such equally paradoxical remark as: "We really know that all of our scientific beliefs are false."

Of course the "paradox" is not genuine. Eschewing induction blocks inference to the fate of current theories from the record of past theories. But then, acceptance of any particular theory on the basis of evidence bearing individually on it is equally blocked, and the negative, general implication as to the prospects for scientific knowledge is sustained. Philosophers of science on many fronts are understandably in retreat from the traditional view that the growth of science delivers ever closer approximations to fundamental truths about the physical world.

Part I of the present paper examines some examples of this movement so as to establish a general perspective on the problem of connecting truth with progressive theory change. Part II defends the traditional view that progress attests to truth by developing the thesis that science exhibits certain forms of progress for which realism with respect to scientific theories