METHOD IN THE PHYSICAL SCIENCES

JOHN VON NEUMANN
Professor of Mathematics
Institute of Advanced Studies

Emphasis on methodology seems most often to arise when there are symptoms of trouble, when a realization of difficulties makes necessary a re–examination of some position inherited from the past. Because traditional attitudes in the sciences seem to have been firmer and more self-assured than in other disciplines, there has perhaps been less searching of the scientist's conscience and, therefore, less concern with methodology on his part. Yet he has not been without concern, for there have been within the experience of people now living at least three serious crises — or reverberations of earlier crises — which have caused him to reorientate his thinking. We can use these crises as prototypes for reference, and we can calibrate our statements with their help.

There have been two such crises in physics — namely, the conceptual soul-searching connected with the discovery of relativity and the conceptual difficulties connected with discoveries in quantum theory. In the case of relativity, the crisis was brief but violent. The second persisted over a longer period, during the almost thirty years in which the quantum theory took shape. The third crisis was in mathematics. It was a very serious conceptual crisis, dealing with rigor and the proper way to carry out a correct mathematical proof. In view of earlier notions of the absolute rigor of mathematics, it is surprising that such a thing could have happened, and even more surprising that it could have happened in these latter days when miracles are not supposed to take place. Yet it did happen.

Concerning the crisis in mathematics, Hermann Weyl, who had a much more direct part in it, can speak with more authority than I.

Published in "The Unity of Knowledge", ed. L. Leary (1955), pp. 157–164; and also John von Neumann Collected Works, ed. A. Taub, Vol. VI, pp. 491–498.

Therefore my discussion will be limited to the two crises in physics — and on these I shall be less specific about conceptual revisions because Niels Bohr has already touched upon this subject in his essay. I will further limit myself to saying a few things about procedure and method which will illustrate the general character of method in science. Not only for the sake of argument but also because I really believe it, I shall defend the thesis that the method in question is primarily opportunistic — also that outside the sciences, few people appreciate how utterly opportunistic it is.

I

To begin, we must emphasize a statement which I am sure you have heard before, but which must be repeated again and again. It is that the sciences do not try to explain, they hardly even try to interpret, they mainly make models. By a model is meant a mathematical construct which, with the addition of certain verbal interpretations, describes observed phenomena. The justification of such a mathematical construct is solely and precisely that it is expected to work — that is, correctly to describe phenomena from a reasonably wide area. Furthermore, it must satisfy certain esthetic criteria — that is, in relation to how much it describes, it must be rather simple. I think it is worth while insisting on these vague terms — for instance, on the use of the word rather. One cannot tell exactly how "simple" simple is. Some of the theories that we have adopted, some of the models with which we are very happy and of which we are very proud would probably not impress someone exposed to them for the first time as being particularly simple.

Simplicity is largely a matter of historical background, of previous conditioning, of antecedents, of customary procedures, and it is very much a function of what is explained by it. If the amount of material which is unambiguously explained — that is, explained with no added interpretations or commentaries — is extremely extensive, if it is also very heterogeneous, if one has clearly explained a large number of things in very different areas, then one will accept a good deal of complication and a good deal of deviation from stylistic beauty. If, on the other hand, only relatively little has been explained, one will absolutely insist that it should at least be done by very simple and direct means. It must also be said that the criterion, that a lot should be explained, has to be applied with a good deal of sophistication. Indeed, some of the nuances of all these requirements can probably only be appreciated intuitively.

The ability to describe — or to predict — correctly is important in such a model, but it need not be decisive per se. Also, in scientific prediction it does not matter enormously whether prediction occurs before or after the fact. Of course, it must be correct. However, as I mentioned above, it is considered very important that the material which has been correctly described or predicted should be heterogeneous. Let me analyze this requirement in somewhat more detail.

If possible, the confirmation should not all stem from one area alone. In this sense, it is considered particularly significant to find confirmations in areas which were not in the mind of anyone who invented the theory. Thus, if you discover that the theory, which was necessitated by difficulties in one area, describes things correctly in entirely different areas, this is highly significant. It is even more important, if things have not been previously very harmonious in these latter areas and there was no sense of optimism about them.

In this regard, the enormous authority of quantum mechanics is typical. It was probably strongly conditioned by the fact that quantum mechanics came into being because of various difficulties in spectroscopy and of various other problems of atoms and molecules which are variously connected with spectroscopy, but that it was then suddenly found capable of describing or predicting correctly various things in chemistry, in solid-state physics, and even to have some bearing on epistemology. These were hardly on anyone's mind at the beginning.

Similar considerations apply to Newtonian mechanics and its still more monumental degree of authority. The latter is largely due to the fact that the Newtonian system was introduced in order to describe the behavior of the sun's planets. It then turned out that it also described, with only small and perfectly plausible additions, many things in extensive areas in very varied parts of physics.

There are also other aspects of the matter which must be kept in mind. It is important that the phenomena which are correctly described should vary considerably, not only qualitatively, but also in their quantitative aspects. Thus it is one of the most impressive traits of Newtonian theory, the classical theory of gravity, that it explains phenomena on the human scale as well as on the planetary scale. And many outside the sciences are not aware of how limited are the scales on which theories usually work. The ratio of the linear sizes of the largest and the smallest objects that have figured in physical theory — the hypothesized universe on the one hand and the smallest particles on the other — is about 10^{40}. In other words, all our physical experience, no matter how recondite, is covered in its linear scale by 40 powers of 10. This amounts to 133 powers of 2–133 octaves. No theory to data has had something valid to say all along the scale. Any theory which can make statements referring to widely separated portions of the scale enjoys very great authority, even if the statements are quite weak. For the Newtonian system, confirmation was developed over 25, or possibly 30 powers of 10 along this scale. For most other existing theories, the area of confirmation in this sense is still more restricted.

II

In evaluating what these models do, one should also emphasize how little of directly interpretive element need be attached to them. In this respect, it is instructive to look at a classical example. In other areas, even in some which are within science — for instance, biology — it is, or was, considered very important to which one of two major types the view that one takes of the area belongs. Specifically, whether the view is causal or teleological. In using the word causal, I do not have the contrast of causal and statistical in mind, but the other contrast of causal and teleological. Causal means that if you know the state of the system now, then you can use this knowledge to predict its state immediately thereafter. Immediately means a very short time; the prediction may not be exact for any finite time, but the shorter the time, the more exact it gets, and that at an accelerated pace, so that it can be extended by the usual process of integration. Thus, one can extend such a prediction by successive steps to any point in the future with any desired degree of precision. Hence, complete knowledge of the system now permits one to calculate unambiguously everything about it at any time in the future. In most causal systems one can also proceed similarly to any time in the past.

This is one of the major ways of looking at nature. Classical Newtonian mechanics is usually quoted as the archetype of this kind of insight and procedure. Under this dispensation, if you know the state of a system now, you can calculate what it will be at any moment thereafter and also, usually at any moment before. One has to be careful, however, in defining the concept of a state. The state is specified if one has a complete description. However, one must consider that this is to a certain degree begging the principle, since by a complete description one means one which comprises precisely as much information as is needed in order to perform the causal progression to the future.

In classical mechanics one knows how much information such a complete description must contain. One has to know, not only where all parts of the system are (all coordinates), but also how rapidly these are moving (all velocities). Then classical mechanics permits calculations of what positions and velocities it will have at any later time. One needs precisely these positions and these velocities. Nothing less will do, and there is no need for other things that might, a priori, seem equally important, like accelerations. The reason why a state in classical mechanics is described by specifying position and velocity, and not position alone, or not position and velocity and certain accelerations, is of course that the Newtonian system is closed just at this point. It is just this amount of information, position plus velocity, which is hereditary in that theory and which can be propagated into the future by unambiguous calculations.

Another aspect is the teleological one. Here one has to take a whole expanse of the history of the system, between two moments which are definitely apart in time, for example, between now and an hour from now, or between now and a trillionth of a second from now, or now and a millennium from now. Taking such a finite stretch of history as the subject of inquiry, a teleological theory asserts that this entire historical process must satisfy certain criteria which are usually stated in terms of optimizing (maximizing) a suitable function of the process. The use of the word optimizing again illustrates the opportunism that even reflects itself in the terminology. By optimizing one only means that one makes some quantity as large as possible. Whether that quantity is particularly desirable or not does not matter. By changing its sign one could transform the criterion in making it as small as possible. Thus, optimizing, maximizing, and minimizing are all neutral mathematical terms, to be substituted for each other on the basis of mathematical convenience and taste.

At any rate, by a single optimization, that is, a single maximization, the total history between two points in time is determined. The real course of events turns out to be that one for which the particular quantity referred to above is made as large as possible. In other words, one develops a complete historical evolution in a single act, by a single insertion between the known points at the beginning and at the end. It is not developed stepwise, progressing from the beginning forward in time, as it would be in a causal theory.

This contrast is very well known from biology. It is also familiar in a number of fields increasingly removed from science. It is usually considered as a very fundamental contrast: the causal and the teleological procedures are viewed as mutually exclusive, as highly antithetical ways of explaining phenomena. It is therefore very important and very characteristic that in science there need not be any meaningful difference between these two descriptions. Indeed, in classical mechanics there are two absolutely equivalent ways to state the same theory, and one of them is causal and the other one is teleological. Both describe the same thing, Newtonian mechanics. Newton's description is causal and d'Alembert's description is teleological. This has been known for well over two hundred years. All the difference between the two is a purely mathematical transformation. In principle such a transformation is no more profound than choosing to say four instead of saying two times two. In other words, by purely mathematical manipulation one can show that each of these two ways gives exactly the same results as the other.

Thus whether one chooses to say that classical mechanics is causal or teleological is purely a matter of literary inclination at the moment of talking. This is very important, since it proves, that if one has really technically penetrated a subject, things that previously seemed in complete contrast, might be purely mathematical transformations of each other. Things which appear to represent deep differences of principle and of interpretation, in this way may turn out not to affect any significant statements and any predictions. They mean nothing to the content of the theory.

III

Thus we have an example where alternative interpretations of the same theory are possible, but where the question of whether one uses one or the other is decided in a manner quite different from what is generally believed to be the valid way. Indeed, the criterion is one of mathematical convenience or taste.

There is also another example where this is the case, but only up to a certain point: beyond that point serious, substantively relevant differences of interpretation arise. The example is quantum mechanics. I will limit myself to that part of the theory which refers to the electronic shells of the atoms. For these a theory is known which seems to be entirely satisfactory at present. This theory can be described in two different ways which differ quite importantly, somewhat in the manner of the causal and teleological interpretations of Newtonian mechanics discussed above, though the difference in this case is not quite as striking and profound as there. The two descriptions are, first, the original procedure of Erwin Schrödinger which describes this part of quantum mechanics by an analogue with optics, and second, the method of Werner Heisenberg which describes this area in completely probabilistic terms.

Since these descriptions were first formulated, a great deal of work has been done on both and they have been further elaborated. In the process it was demonstrated that they are mathematically equivalent. The prevalent taste is today, and has been for more than twenty years, rather in favor of one of the two interpretations, namely, the statistical one. (It must be said, however, that there have been in the last few years some interesting attempts to revive the other interpretation.) It was, moreover, quite clear all along, that ultimately the motive for choosing one or the other attitude would be connected with the fact that quantum mechanics, in spite of all its successes, is contiguous with areas in which the theory is not satisfactory, specifically, with the quantum theory of electrodynamics and subsequently with the quantum theory of particles like mesons and their successors.

About all of these we know a great deal less than about the original area of quantum mechanics, and we are here in the midst of grave difficulties. The reason for preferring one version of quantum theory over the other has usually been the intuitive hope that one or the other would give better heuristic guidance in extending the theory into those areas which are not yet properly explained or not yet properly theoreticized and controlled. Throughout the last twenty years this has been prevalently believed to be a matter of finding correct formal extensions of the existing theory. If this ultimately proves to be the case, it will determine the final choice. Questions of form, even when the mathematical contexts are equivalent, can therefore have great heuristic and guiding importance, and in the end determine the outcome.

There have been some individual exceptions to this rule. Some physicists certainly had definite subjective preferences for one description or the other. However, there can be hardly any doubt that scientific "public opinion" in the end will only accept that variant which succeeds in pointing the way to explaining wider areas with greater power. In other words, while there appears to be a serious philosophical controversy between the interpretations of Schrödinger and Heisenberg, it is quite likely that the controversy will be settled in quite an unphilosophical way. The decision is likely to be opportunistic in the end. The theory that lends itself better to formalistic extension towards valid new theories will overcome the other, no matter what our preference up to that point might have been. It must be emphasized that this is not a question of accepting the correct theory and rejecting the false one. It is a matter of accepting that theory which shows greater formal adaptability for a correct extension. This is a formalistic, esthetic criterion, with a highly opportunistic flavor.