of Truth

application to an
ed), and especially
uch a theory, the
finition 22b, seems
int of view, mainly
.[12]

ion of a law (not written
x) as a finite sequence of
tant, to define a *refuting*
a finite [and adequate]

APPENDIX

The Bucket and the Searchlight: Two Theories of Knowledge

THE purpose of this paper is to criticize a widely held view about the aims and methods of the natural sciences, and to put forward an alternative view.

I

I SHALL start with a brief exposition of the view I propose to examine, which I will call '*the bucket theory of science*' (or '*the bucket theory of the mind*'). The starting point of this theory is the persuasive doctrine that before we can know or say anything about the world, we must first have had perceptions—sense experiences. It is supposed to follow from this doctrine that our knowledge, our experience, consists either of accumulated perceptions (naïve empiricism) or else of assimilated, sorted, and classified perceptions (a view held by Bacon and, in a more radical form, by Kant).

The Greek atomists had a somewhat primitive notion of this process. They assumed that atoms break loose from the objects we perceive, and penetrate our sense organs, where they become perceptions; and out of these, in the course of time, our knowledge of the external world fits itself together [like a self-assembling jigsaw puzzle]. According to this view, then, our mind resembles a container—a kind of bucket—in which perceptions and knowledge accumulate. (Bacon speaks of perceptions as 'grapes, ripe and in

---

A lecture delivered (in German) at the European Forum of the Austrian College, Alpbach, Tyrol, in August 1948, and first published, in German, under the title '*Naturgesetze und theoretische Systeme*' in *Gesetz und Wirklichkeit*, edited by Simon Moser, 1949. Not previously published in English. [Textual additions made in this translation are put into square brackets, or indicated in the footnotes.]

The paper anticipates many of the ideas developed more fully in this volume and in *Conjectures and Refutations*, and in addition it contains some ideas which I have not published elsewhere. Most of the ideas, and the expressions: 'the bucket theory of the mind' and 'the searchlight theory of science [and of the mind]' go back to my New Zealand days and are first mentioned in my *Open Society*. I read a paper under the title 'The Bucket Theory of the Mind' in the Staff Club of the London School of Economics in 1946. This Appendix is especially closely related to Chapters 2 and 5 of the present volume.

season' which have to be gathered, patiently and industriously, and from which, if pressed, the pure wine of knowledge will flow.)

Strict empiricists advise us to interfere as little as possible with this process of accumulating knowledge. True knowledge is pure knowledge, uncontaminated by those prejudices which we are only too prone to add to, and mix with, our perceptions; these alone constitute experience pure and simple. The result of these additions, of our disturbing and interfering with the process of accumulating knowledge, is error. Kant opposes this theory: he denies that perceptions are ever pure, and asserts that our experience is the result of a process of assimilation and transformation—the combined product of sense perceptions and of certain ingredients added by our minds. The perceptions are the raw material, as it were, which flows from outside into the bucket, where it undergoes some (automatic) processing—something akin to digestion, or perhaps to systematic classification—in order to be turned in the end into something not so very different from Bacon's 'pure wine of experience'; let us say, perhaps, into fermented wine.

I do not think that either of these views suggests anything like an adequate picture of what I believe to be the actual process of acquiring experience, or the actual method used in research or discovery. Admittedly, Kant's view might be so interpreted that it comes much nearer to my own view than does pure empiricism. I grant, of course, that science is impossible without experience (but the notion of 'experience' has to be carefully considered). Though I grant this, I nevertheless hold that perceptions do not constitute anything like the raw material, as they do according to the 'bucket theory', out of which we construct either 'experience' or 'science'.

II

In science it is *observation* rather than perception which plays the decisive part. But observation is a process in which we play an intensely *active* part. An observation is a perception, but one which is planned and prepared. We do not 'have' an observation [as we may 'have' a sense experience] but we 'make' an observation. [A navigator even 'works' an observation.] An observation is always preceded by a particular interest, a question, or a problem—in short, by something theoretical.[1] After all, we can put every question

[1] By the word 'theoretical' I do not mean here the opposite of 'practical' (since our interest might very well be a practical one); it should rather be understood in the sense of 'speculative' [as with a speculative interest in a pre-existing problem] in contrast to 'perceptive'; or 'rational' as opposed to 'sensual'.

in the form of a hypothesis or conjecture to which we add: 'Is this so? Yes or no?' Thus we can assert that every observation is preceded by a problem, a hypothesis (or whatever we may call it); at any rate by something that interests us, by something theoretical or speculative. This is why observations are always selective, and why they presuppose something like a principle of selection.

Before elaborating these points any further I shall try to introduce, as a digression, a few remarks of a biological nature. Although these are not meant to constitute a basis or even an argument for the main thesis which I intend to propose later, they may perhaps be helpful in getting over, or in circumventing, certain objections to it, and in this way facilitate later its understanding.

III

We know that all living things, even the most primitive, react to certain stimuli. These reactions are specific; that is to say, for each organism (and for each type of organism) the number of possible reactions is limited. We can say that every organism possesses a certain innate set of possible reactions, or a certain disposition to react in this or that way. This set of dispositions may change with the advancing age of the organism (partly perhaps under the influence of sense impressions or perceptions) or it may remain constant; however this may be, at any instant in the life of the organism it is, we may assume, invested with such a set of possibilities and dispositions to react, and this set constitutes what may be called its [momentary] inner state.

Upon this inner state of the organism will depend how it will react to its external environment. This is why physically identical stimuli may at different times produce different reactions, whilst physically different stimuli may result in identical reactions.[2]

Now we shall say that an organism '*learns from experience*' only if its dispositions to react change in the course of time, and if we have reason to assume that these changes do not depend merely on innate [developmental] changes in the state of the organism but also on the changing state of its external environment. (This is a necessary condition, though not a sufficient one, for saying that the organism learns from experience.) In other words, we shall regard the process by which the organism learns as a certain kind of change, or modification, in its dispositions to react, and not, as would the

[2] Compare F. A. von Hayek, 'Scientism and the Study of Society', *Economica*, N.S. 9, 10, and 11 (1942, 1943, and 1944); [now also in his *The Counter-Revolution of Science*, 1952].

bucket theory, as an (ordered or classified or associated) accumulation of memory traces, left over by perceptions that are past.

These modifications in the organism's disposition to react, which go to make up the processes of learning, are closely connected with the important notion of an '*expectation*', and also with that of a '*disappointed expectation*'. We may characterize an expectation as a *disposition to react, or as a preparation for a reaction*, which is adapted to [or which anticipates] a state of the environment yet to come about. This characterization seems to be more adequate than one that describes an expectation in terms of states of consciousness; for we become conscious of many of our expectations only when they are disappointed, owing to their being unfulfilled. An example would be the encountering of an unexpected step in one's path: it is the unexpectedness of the step which may make us conscious of the fact that we expected to encounter an even surface. Such disappointments force us to *correct* our system of expectations. The process of learning consists largely in such corrections; that is, in the elimination of certain [disappointed] expectations.

IV

LET us now return to the problem of observation. An observation always presupposes the existence of some system of expectations. These expectations can be formulated in the form of queries; and the observation will be used to obtain either a confirming or a correcting answer to expectations thus formulated.

My thesis that the question, or the hypothesis, must precede the observation may at first have seemed paradoxical; but we can see now that it is not at all paradoxical to assume that expectations—that is, dispositions to react—must precede every observation and, indeed, every perception: for certain dispositions or propensities to react are innate in all organisms whereas perceptions and observations clearly are not innate. And although perceptions and, even more, observations, play an important part in the process of *modifying* our dispositions or propensities to react, some such dispositions or propensities must, of course, be present first, or they could not be modified.

These biological reflections are by no means to be understood as implying my acceptance of a behaviourist position. I do not deny that perceptions, observations, and other states of consciousness occur, but I assign to them a role very different from the one they are supposed to play according to the bucket theory. Nor are these biological reflections to be regarded as forming in any sense an

assumption on which my arguments will be based. But I hope that they will help towards a better understanding of these arguments. The same may be said of the following reflections, which are closely connected with these biological ones.

At every instant of our pre-scientific or scientific development we are living in the centre of what I usually call a '*horizon of expectations*'. By this I mean the sum total of our expectations, whether these are subconscious or conscious, or perhaps even explicitly stated in some language. Animals and babies have also their various and different horizons of expectations though no doubt on a lower level of consciousness than, say, a scientist whose horizon of expectations consists to a considerable extent of linguistically formulated theories or hypotheses.

The various horizons of expectations differ, of course, not only in their being more or less conscious, but also in their content. Yet in all these cases the horizon of expectations plays the part of a frame of reference: only their setting in this frame confers meaning or significance on our experiences, actions, and observations.

Observations, more especially, have a very peculiar function within this frame. They can, under certain circumstances, destroy even the frame itself, if they clash with certain of the expectations. In such a case they can have an effect upon our horizon of expectations like a bombshell. This bombshell may force us to reconstruct, or rebuild, our whole horizon of expectations; that is to say, we may have to correct our expectations and fit them together again into something like a consistent whole. We can say that in this way our horizon of expectations is raised to and reconstructed on a higher level, and that we reach in this way a new stage in the evolution of our experience; a stage in which those expectations which have not been hit by the bomb are somehow incorporated into the horizon, while those parts of the horizon which have suffered damage are repaired and rebuilt. This has to be done in such a manner that the damaging observations are no longer felt as disruptive, but are integrated with the rest of our expectations. If we succeed in this rebuilding, then we shall have created what is usually known as an *explanation* of those observed events [which created the disruption, the problem].

As to the question of the temporal relation between observation on the one hand and the horizon of expectations or theories on the other, we may well admit that a new explanation, or a new hypothesis, is generally preceded in time by *those* observations which destroyed the previous horizon of expectations and thus were the stimulus to our attempting a new explanation. Yet this must not be

understood as saying that observations generally precede expectations or hypotheses. On the contrary, each observation is preceded by expectations or hypotheses; by those expectations, more especially, which make up the horizon of expectations that lends those observations their significance; only in this way do they attain the status of real observations.

The question, 'What comes first, the hypothesis (*H*) or the observation (*O*)?' reminds one, of course, of that other famous question: 'What came first, the hen (*H*) or the egg (*O*)?' Both questions are soluble. The bucket theory asserts that [just as a primitive form of an egg (*O*), a unicellular organism, precedes the hen (*H*)] observation (*O*) always precedes every hypothesis (*H*); for the bucket theory regards the latter as arising from observations by generalization, or association, or classification. By contrast, we can now say that the hypothesis (or expectation, or theory, or whatever we may call it) precedes the observation, even though an observation that refutes a certain hypothesis may stimulate a new (and therefore a temporally later) hypothesis.

All this applies, more especially, to the formation of scientific hypotheses. For we learn only from our hypotheses what kind of observations we ought to make: whereto we ought to direct our attention; wherein to take an interest. Thus it is the hypothesis which becomes our guide, and which leads us to new observational results.

This is the view which I have called the '*searchlight theory*' (in contradistinction to the '*bucket theory*'). [According to the searchlight theory, observations are secondary to hypotheses.] Observations play, however, an important role as *tests* which a hypothesis must undergo in the course of our [critical] examination of it. If the hypothesis does not pass the examination, if it is falsified by our observations, then we have to look around for a new hypothesis. In this case the new hypothesis will come after those observations which led to the falsification or rejection of the old hypothesis. Yet what made the observations interesting and relevant and what altogether gave rise to our undertaking them in the first instance, was the earlier, the old [and now rejected] hypothesis.

In this way science appears clearly as a straightforward continuation of the pre-scientific repair work on our horizons of expectations. Science never starts from scratch; it can never be described as free from assumptions; for at every instant it presupposes a horizon of expectations—yesterday's horizon of expectations, as it were. Today's science is built upon yesterday's science [and so it is the result of yesterday's searchlight]; and

yesterday's science, in turn, is based on the science of the day before. And the oldest scientific theories are built on pre-scientific myths, and these, in their turn, on still older expectations. Ontogenetically (that is, with respect to the development of the individual organism) we thus regress to the state of the expectations of a newborn child; phylogenetically (with respect to the evolution of the race, the phylum) we get to the state of expectations of unicellular organisms. (There is no danger here of a vicious infinite regress—if for no other reason than that every organism is born with *some* horizon of expectations.) There is, as it were, only one step from the amoeba to Einstein.

Now if this is the way science evolves, what can be said to be the characteristic step which marks the transition from pre-science to science?

V

THE first beginnings of the evolution of something like a scientific method may be found, approximately at the turn of the sixth and fifth centuries B.C., in ancient Greece. What happened there? What is new in this evolution? How do the new ideas compare with the traditional myths, which came from the East and which, I think, provided many of the decisive suggestions for the new ideas?

Among the Babylonians and the Greeks and also among the Maoris in New Zealand—indeed, it would seem, among all peoples who invent cosmological myths—tales are told which deal with the beginning of things, and which try to understand or explain the structure of the Universe in terms of the story of its origin. These stories become traditional and are preserved in special schools. The tradition is often in the keeping of some separate or chosen class, the priests or medicine men, who guard it jealously. The stories change only little by little—mainly through inaccuracies in handing them on, through misunderstandings, and sometimes through the accretion of new myths, invented by prophets or poets.

Now what is new in Greek philosophy, what is newly added to all this, seems to me to consist not so much in the replacement of the myths by something more 'scientific', as in a *new attitude towards the myths*. That their character then begins to change seems to me to be merely a consequence of this new attitude.

The new attitude I have in mind is *the critical attitude. In the place of a dogmatic handing on of the doctrine* [in which the whole interest lies in the preservation of the authentic tradition] *we find a critical*

discussion of the doctrine. Some people begin to ask questions about it; they doubt the trustworthiness of the doctrine: its *truth.*

Doubt and criticism certainly existed before this stage. What is new, however, is that doubt and criticism now become, in their turn, part of the tradition of the school. A tradition of a higher order replaces the traditional preservation of the dogma: in the place of traditional theory—in place of the myth—we find the tradition of criticizing theories (which at first themselves are hardly more than myths). It is only in the course of this critical discussion that observation is called in as a witness.

It can hardly be a mere accident that Anaximander, the disciple of Thales, developed a theory which explicitly and consciously diverges from that of his master, and that Anaximenes, the disciple of Anaximander, diverges just as consciously from his master's doctrine. The only explanation seems to be that the founder of the school himself challenged his disciples to criticize his theory, and that they turned this new critical attitude of their master's into a new tradition.

It is interesting that this happened only once, so far as I know. The earlier Pythagorean school was almost certainly a school of the old kind: its tradition does not embrace the critical attitude but is confined to the task of preserving the doctrine of the master. It was undoubtedly only the influence of the critical school of the Ionians which later loosened the rigidity of the Pythagorean school tradition and so paved the way leading to the philosophical and scientific method of criticism.

The critical attitude of ancient Greek philosophy can hardly be better exemplified than by the famous lines of Xenophanes:

> Yet if cattle or horses or lions had hands and could draw
> And could sculpture like men, then the horses would draw their gods
> Like horses; and cattle like cattle; and each would then shape
> Bodies of gods in the likeness, each kind, of its own.

This is not only a critical challenge—it is a statement made in the full consciousness and mastery of a critical methodology.

Thus it seems to me that it is the tradition of criticism which constitutes what is new in science, and what is characteristic of science. On the other hand it seems to me that the task which science sets itself [that is, the explanation of the world] and the main ideas which it uses, are taken over without any break from prescientific mythmaking.

VI

WHAT is the task of science? With this question I have ended my preliminary examination of biological and historical trends, and I now come to the logical analysis of science itself.

The task of science is partly theoretical—*explanation*—and partly practical—*prediction and technical application*. I shall try to show that these two aims are, in a way, two different aspects of one and the same activity.

I will first examine the idea of an explanation.

One often hears it said that an explanation is the reduction of the unknown to the known; but we are rarely told how this is to be done. At any rate, this notion of explanation is not one that was ever used in the actual practice of explanation in science. If we look at the history of science in order to see what kinds of explanation were used and accepted as satisfactory at one time or another, then we find a very different notion of explanation in practical use.

I gave a short sketch of this history (I do not mean the history of the concept of explanation but the history of the practice of explanation) to the philosophical seminar this morning.[3] Unfortunately, time prevents me from dealing with this question here again at length. Yet I should mention here one general result. In the course of the historical development of science many different methods and kinds of explanation have been regarded as acceptable; but they all have one aspect in common: the various methods of explanation all consist of a *logical deduction*; a deduction whose conclusion is the *explicandum*—a statement of the thing to be explained—and whose premisses consist of the *explicans* [a statement of the explaining laws and conditions]. The main changes that have occurred in the course of the history of science consist in the silent abandonment of certain implicit demands regarding the character of the *explicans* (that it can be intuitively grasped, that it is to be self-evident, etc.); demands which turn out not to be reconcilable with certain other demands whose crucial significance becomes more and more obvious as time goes on; in particular the demand for the independent testability of the *explicans* [which forms the premisses and thus the very heart of the explanation].

[3] (Added in the translation.) Part of the fuller story will be found (though somewhat condensed and with reduced emphasis upon what has been in actual practice accepted as an explanation) in my Venice lecture: 'Philosophy and Physics: Theories of the Structure of Matter', now contained in my book *Philosophy and Physics* (1972). Other parts are to be found in the first half of my *Conjectures and Refutations*, especially chapters 6, 3, and 4. (This last chapter will be found to overlap with, and expand, some parts of the present lecture.)

Thus an explanation is always the deduction of the *explicandum* from certain premisses, to be called the *explicans*.

Here is a somewhat gruesome example, just for the purpose of illustration.[4]

A dead rat has been discovered and we wish to know what has happened to it. The *explicandum* may be stated thus: 'This rat here has died recently.' This *explicandum* is definitely known to us—the fact lies before us in stark reality. If we want to explain it, we must try out some conjectural or hypothetical explanations (as the authors of detective stories do); that is to say, explanations which introduce something *unknown*, or at any rate much less known, to us. Such a hypothesis may be, for instance, that the rat died of a large dose of rat poison. This is useful as a hypothesis in so far as, firstly, it helps us to formulate an *explicans* from which the *explicandum* can be deduced; secondly, it suggests to us a number of independent tests—tests of the *explicans* which are quite independent of whether the *explicandum* is true or not.

Now the *explicans*—which is our hypothesis—does not only consist of the sentence 'This rat has eaten some bait containing a large dose of rat poison', for from this statement alone one cannot validly deduce the *explicandum*. Rather, we shall have to use, as *explicans*, two different kinds of premisses—*universal laws* and *initial conditions*. In our case the universal law might be put like this: 'If a rat eats at least eight grains of rat poison it will die within five minutes.' The (singular) initial condition (which is a singular statement) might be: 'This rat ate at least eighteen grains of rat poison, more than five minutes ago.' From these two premisses together we may now indeed deduce that this rat recently has died [that is, our *explicandum*].

Now all this may seem somewhat obvious. But consider one of my theses—the thesis, namely, that what I have called the '*initial conditions*' [the conditions pertaining to the individual case] never suffice by themselves as an explanation, and that we always need a general law as well. Now this thesis is by no means obvious; on the contrary, its truth is often not admitted. I even suspect that most of you would be inclined to accept a remark like 'this rat has eaten rat poison' as quite sufficient to explain its death, even if no explicit statement of the universal law regarding the effects of rat poison is added. But suppose for a moment that we were living in a world in which anybody (and also any rat) who eats a lot of that chemical called 'rat poison' will feel especially well and happy for a week to come and more lively than ever before. If a universal law like this were valid, could the statement 'This rat has eaten rat

[4] I have made the example slightly less gruesome in the translation.

poison' still be acceptable as an explanation of death? Obviously not.

Thus we have reached the important result, often overlooked, that any explanation that utilizes the singular initial conditions alone would be incomplete, and that *at least one universal law* is needed besides, even though this law is, in some cases, so well known that it is omitted as if it were redundant.

To sum up this point. We have found that an explanation is a deduction of the following kind:

U	(Universal Law)	} Premisses
I	(Specific Initial Conditions)	} (constituting the *Explicans*)
$\overline{E}$	(*Explicandum*)	Conclusion

VII

BUT are all explanations of this structure *satisfactory*? Is, for instance, our example (which explains the death of the rat by reference to rat poison) a satisfactory explanation? We do not know: the tests may show that whatever the rat may have died of, it was not rat poison.

If some friend should be sceptical of our explanation and ask, 'How do you know that this rat ate poison?', it will obviously not be sufficient to answer, 'How can you doubt it, seeing that it is dead?'. Indeed, any reason which we may state in support of any hypothesis must be other than, and independent of, the *explicandum*. If we can only adduce the *explicandum* itself as evidence, we feel that our explanation is circular, and therefore quite *unsatisfactory*. If, on the other hand, we are able to reply, 'Analyse the contents of its stomach, and you will find a lot of poison', and if this prediction (which is new—that is, not entailed by the *explicandum* alone) proves true, we shall at least consider our explanation a fairly good hypothesis.

But I have to add something. For our sceptical friend may also question the truth of the universal law. He may say, for instance, 'Granted, this rat has eaten a certain chemical; but why should it have died of it?'. Again, we must not answer: 'But don't you see it is dead? That just shows you how dangerous it is to eat this chemical.' For this again would make our explanation circular and unsatisfactory. In order to make it satisfactory we should have to submit the universal law to test cases which are independent of our *explicandum*.

With this, my analysis of the formal schema of explanation may be regarded as concluded, but I shall add some further remarks and analyses to the general schema I have outlined.

First, an observation about the ideas of cause and effect. The state of affairs described by the singular *initial conditions* can be called the '*cause*', and the one described by the *explicandum* the '*effect*'. I feel, however, that these terms, encumbered as they are with associations from their history, are better avoided. If we still want to use them, we should always remember that they acquire a meaning only relative to a theory or a universal law. It is the theory or the law which constitutes the *logical link* between cause and effect, and the statement '*A* is the cause of *B*' should be analysed thus: 'There is a theory *T* which can be, and has been, independently tested, and from which, in conjunction with an independently tested description *A*, of a specific situation, we can logically deduce a description, *B*, of another specific situation.' (That the existence of such a *logical* link between 'cause' and 'effect' is presupposed in the very use of these terms has been overlooked by many philosophers, including Hume.)[5]

VIII

THE task of science is not confined to searching for purely theoretical explanations; it also has its practical sides: prediction-making as well as technical applications. Both of these can be analysed by means of the same logical schema which we introduced to analyse explanation.

(1) *The derivation of predictions.* Whereas in the search for an explanation the *explicandum* is given—or known—and a suitable *explicans* has to be found, the derivation of predictions proceeds in the opposite direction. Here the theory is given, or assumed to be known (perhaps from textbooks), and so are the specific initial conditions (they are known, or assumed to be known, by observation). What remain to be found are the logical consequences: certain logical conclusions which are not yet known to us from observation. These are the *predictions*. In this case, the prediction *P* takes the place of the *explicandum E* in our logical schema.

(2) *Technical application.* Consider the task of building a bridge which has to comply with certain practical requirements, laid down in a list of specifications. What we are given are the specifications, *S*, which describe a certain required state of affairs—the bridge to be built. (*S* are the customer's specifications, which are given prior to,

[5] (Added in translation.) I made these comments on the notions 'cause' and 'effect' first in section 12 of my *Logik der Forschung* (*The Logic of Scientific Discovery*). See also my *Poverty of Historicism*, pp. 122 f.; my *Open Society and Its Enemies*, especially note 9 to chapter 25; and 'What can Logic do for Philosophy?', *Aristotelian Society, Supplementary Volume*, **22**, 1948, pp. 145 ff.

and are distinct from, the architect's specifications.) We are given, further, the relevant physical theories (including certain rules of thumb). What are to be found are certain initial conditions which may be realized technically and which are of such a nature that the specifications may be deduced from them, together with the theory. So in this case, S takes the place of E in our logical schema.[6]

This makes it clear how, from a logical point of view, both the derivation of predictions and the technical application of scientific theories may be regarded as mere inversions of the basic schema of scientific explanation.

The use of our schema, however, is still not exhausted: it may also serve to analyse the *procedure of testing our explicans*. The testing procedure consists in the derivation from the *explicans* of a prediction, P, and in comparing it with an actual, observable, situation. If a prediction does not agree with the observed situation, then the *explicans* is shown to be false; it is falsified. In this case we still do not know whether it is the universal *theory* which is false, or whether the *initial conditions* describe a situation which does not correspond with the real situation—so that the initial conditions are false. [Of course, it may well be that the theory *and* the initial conditions are false.]

The falsification of the prediction shows that the *explicans* is false, yet the reverse of this does not hold: it is incorrect and grossly misleading to think that we can interpret the 'verification' of the prediction as 'verifying' the *explicans* or even a part of it. For a true prediction may easily have been validly deduced from an *explicans* that is false. It is even quite misleading to regard *every* 'verification' of a prediction as something like a practical *corroboration* of the *explicans*: it would be more correct to say that only such 'verifications' of predictions which are 'unexpected' [without the theory under examination] may be regarded as corroborations of the *explicans*, and so of the theory. This means that a prediction can be used to corroborate a theory only if its comparison with observations might be regarded as a serious attempt at testing the *explicans*—a serious

[6] (Added in the translation.) This analysis must not be interpreted as implying that the technologist or the engineer is concerned only with '*applying*' theories which are supplied by the pure scientist. On the contrary, the technologist and the engineer are constantly faced with *problems to be solved*. These problems are of various degrees of abstraction, but are usually, at least in part, of a theoretical character; and in trying to solve them, the technologist or engineer uses, like everybody else, the method of conjecture, or trial, and testing, or refutation, or error elimination. This is well explained on p. 43 of J. T. Davies, *The Scientific Approach*, 1965, a book in which many good applications and illustrations of the searchlight theory of science can be found.

attempt at refuting it. A ['risky'] prediction of this kind may be called 'relevant to a test of the theory'.[7] After all, it is fairly obvious that the passing of an examination can give an idea of the qualities of the student only if the examination which he passes is sufficiently severe, and that an examination can be designed which even the weakest student will pass easily.[8]

In addition to all this, our logical scheme permits us, finally, to analyse the difference between the tasks of a *theoretical* and of a *historical* explanation.

The theoretician is interested in finding, and testing, universal laws. In the course of testing them he uses other laws, of the most diverse kinds (many of them quite unconsciously) as well as diverse specific initial conditions.

The historian on the other hand, is interested in finding descriptions of states of affairs in certain finite, specific spatio-temporal regions—that is to say, what I have called specific initial conditions—and in testing or checking their adequacy or accuracy. In this kind of testing he uses, in addition to other specific initial conditions, universal laws of all kinds—usually rather obvious ones—which belong to his horizon of expectations, though, as a rule, he is not conscious that he uses them. In this he resembles the theoretician. [Their difference, however, is very marked: it lies in the difference between their various interests, or problems: in the difference of what each regards as problematic.]

In a logical schema [similar to our previous ones] the procedure of the theoretician may be represented in the following manner:

$$\begin{array}{cccc} U_0 & U_0 & U_0 & \ldots \\ U_1 & U_2 & U_3 & \ldots \\ \underline{I_1} & \underline{I_2} & \underline{I_3} & \ldots \\ P_1 & P_2 & P_3 & \ldots \end{array}$$

[7] A relevant prediction corresponds, in a certain sense, to an acid test, or to an '*experimentum crucis*'; for in order that a prediction P may be relevant to a test of a theory T, it must be possible to state a prediction P' which does not contradict the initial condition and the remainder of our horizon of expectations for the time being, other than T (assumptions, theories, etc.), and which, combined with the initial conditions and the remainder of the horizon of expectations, contradicts P. This is what is meant if we say that $P(= E)$ ought to be (without T) 'unexpected'.

[8] Experienced examiners will feel that the word 'easily' is somewhat unrealistic. As the President of a Governmental Board of Examiners in Vienna sometimes said musingly: 'If a student, in answering the examination question "How much is 5 plus 7?" puts down "eighteen", then we give him a pass. But if he answers "green", I then sometimes think afterwards that we really ought to have ploughed him.'

U_0 is here the universal law, the universal hypothesis, which is under examination. It is kept constant throughout the tests, and used, together with various other laws U_1, U_2, ... and various other initial conditions I_1, I_2, ... in order to derive various predictions P_1, P_2, .. which may then be compared with observable actual facts.

The procedure of the historian may be represented by the following schema:

$$\begin{array}{llll} U_1 & U_2 & U_3 & \ldots \\ I_1 & I_2 & I_3 & \ldots \\ \underline{I_0} & \underline{I_0} & \underline{I_0} & \ldots \\ P_1 & P_2 & P_3 & \ldots \end{array}$$

Here, I_0 is the historical hypothesis, the historical description, which is to be examined or tested. It is kept constant throughout the tests; and it is combined with various (mostly obvious) laws, U_1, U_2, ... and with corresponding initial conditions, I_1, I_2, ... for deriving various predictions, P_1, P_2, etc.

Both our schemata are, of course, highly idealized and oversimplified.

IX

EARLIER I have tried to show that an explanation will be *satisfactory* only if its universal laws, its theory, can be tested independently of the *explicandum*. But this means that any satisfactory explanatory theory must always assert *more* than what was already contained in the *explicanda* which originally led us to put it forward. In other words, satisfactory theories must, as a matter of principle, transcend the empirical instances which gave rise to them; otherwise they would, as we have seen, merely lead to explanations which are circular.

Here we have a methodological principle which stands in direct contradiction to all positivistic and naïvely empiricist [or inductivist] tendencies. It is a principle which demands that we should dare to put forward bold hypotheses that open up, if possible, new domains of observations, rather than those careful generalizations from 'given' observations which have remained [ever since Bacon] the idols of all naïve empiricists.

Our view that it is the task of science to put forward explanations, or (what leads essentially to the same logical situation)[9] to create the theoretical foundations for predictions and other applications—this view has led us to the methodoiogical demand that our theories should be testable. Yet there are *degrees of testability*. Some theories are *better* testable than others. If we strengthen our methodological demand and aim at *better and better testable* theories, then we arrive at a methodological principle—or a statement of the task of science—whose [unconscious] adoption in the past would rationally explain a great number of events in the history of science: it would explain them as steps towards carrying out the task of science. (At the same time it gives us a statement of the task of science, telling us what should in science be regarded as *progress*; for in contrast to most other human activities—art and music in particular—there really is, in science, such a thing as progress.)

An analysis and comparison of the degrees of testability of different theories shows that the testability of a theory grows with its *degree of universality* as well as with its *degree of definiteness, or precision.*

The situation is fairly simple. Along with the degree of universality of a theory goes an increase in the range of those events about which the theory can make predictions and thereby also the domain of possible falsifications. But a theory which is more easily falsified is at the same time one that is better testable.

We find a similar situation if we consider the degree of definiteness or precision. A precise statement can be more easily refuted than a vague one, and it can therefore be better tested. This consideration also allows us to explain the demand that qualitative statements should if possible be replaced by quantitative ones by our principle of increasing the degree of testability of our theories. (In this way we can also explain the part played by *measurement* in the testing of theories; it is a device which becomes increasingly important in the course of scientific progress, but which should not be used [as it often is] as a characterizing feature of science, or the formation of theories, in general. For we must not overlook the fact that measuring procedures began to be used only at a fairly late stage in the development of some of the sciences, and that they are even now not used

[9] (Added in the translation.) I have in later years (from 1950 on) made a sharper distinction between the theoretical or explanatory and the practical or 'instrumental' tasks of science, and I have stressed the logical priority of the theoretical task over the instrumental task. I have tried to stress, more especially, that predictions have not only an instrumental aspect, but also, and mainly, a theoretical one, as they play a decisive role in testing a theory (as shown earlier in the present lecture). See my *Conjectures and Refutations*, especially chapter 3.

in all of them; and we must also not overlook the fact that all measurement is dependent on theoretical assumptions.)

x

A GOOD example from the history of science that may be used to illustrate my analysis is the transition from the theories of Kepler and Galileo to the theory of Newton.

That this transition has nothing whatever to do with induction, and that Newton's theory cannot be regarded as anything like a generalization of those two earlier theories may be seen from the undeniable [and important] fact that Newton's theory *contradicts* them. *Thus Kepler's laws cannot be deduced from Newton's* [although it has been often asserted that they can be so deduced, and even that Newton's can be deduced from Kepler's]: Kepler's laws can be obtained from Newton's only *approximately*, by making the [false] assumption that the masses of the various planets are negligible compared with the mass of the sun. Similarly, Galileo's law of free falling bodies cannot be deduced from Newton's theory: on the contrary, it contradicts it. Only by making the [false] assumption that the total length of all falls is negligible compared with the length of the radius of the earth can we obtain Galileo's law *approximately* from Newton's theory.

This shows, of course, that Newton's theory cannot be a generalization obtained by induction [or deduction] but that it is a new hypothesis which can irradiate the way to the falsification of the old theories: it can irradiate, and point the way to those domains in which, according to the new theory, the old theories fail to yield good approximations. (In Kepler's case this is the domain of the theory of perturbations, and in Galileo's case it is the theory of variable accelerations, since according to Newton gravitational accelerations vary inversely with the square of the distance.)

Had Newton's theory achieved no more than the union of Kepler's laws with Galileo's, it would have been only a *circular explanation of these laws* and therefore unsatisfactory as an explanation. Yet its power of illumination and its power of convincing people consisted just in its power to throw light on the way to independent tests, leading us to [successful] predictions which were incompatible with the two older theories. It was the way to new empirical discoveries.

Newton's theory is an example of an attempt to explain certain older theories of a lower degree of universality, which not only leads to a kind of unification of these older theories but at the same time to their falsification (and so to their correction by restricting or

determining the domain within which they are, in good approximation, valid).[10] A case which occurs perhaps more often is this: an old theory is first falsified; and the new theory arises later, as an attempt to explain the partial success of the old theory as well as its failure.

XI

In connection with my analysis of the notion (or rather the practice) of explanation a further point seems significant. From Descartes [and perhaps even from Copernicus] to Maxwell, most physicists tried to explain all newly discovered relations by means of *mechanical models*; that is to say, they tried to reduce them to laws of push or pressure with which we are acquainted from handling everyday physical things—things belonging to the realm of 'physical bodies of medium size'. Descartes made this into a kind of programme for all the sciences; he even demanded that we should restrict ourselves to models that work merely by push or pressure. This programme suffered its first defeat with the success of Newton's theory; but this defeat (which was a serious affliction to Newton and his generation) was soon forgotten, and gravitational attraction was admitted into the programme on equal terms with push and pressure. Maxwell, too, first tried to develop his theory of the

[10] (Added in the translation.) The incompatibility of Newton's theory with that of Kepler was stressed by Pierre Duhem, who wrote of Newton's '*principles of universal gravity*' that it is '*very far from being derivable by generalization and induction from the observational laws of Kepler*' in that it '*formally contradicts these laws. If Newton's theory is correct, Kepler's laws are necessarily false.*' (The quotation is from p. 193 of P. P. Wiener's translation of Duhem's *The Aim and Structure of Physical Theory*, 1954. The term '*observational*' applied here to the '*laws of Kepler*' should be taken with a good grain of salt: Kepler's laws were wild conjectures, just as much as Newton's were: they cannot be induced from Tycho's observations—no more than Newton's can from Kepler's laws.) Duhem's analysis is based on the fact that our solar system contains *many* heavy planets for whose mutual attraction allowance has to be made in accordance with Newton's theory of perturbation. We can, however, go beyond Duhem: even if we take Kepler's laws as holding for a set of *two-body systems*, each of them containing a central body of the mass of the sun and *one* planet (of varying mass and distance in the various different systems belonging to the set), even then Kepler's third laws fails if Newton's laws are true, as I have shown briefly in *Conjectures and Refutations*, note 28 to chapter 1 (p. 62) and in some detail in my paper 'The Aim of Science', (1957), now Chapter 5 of the present volume, also in *Theorie und Realität*, edited by Hans Albert, 1964, chapter 1, pp. 73 ff., especially pp. 82 f. In this paper I say a little more about explanations which *correct their* (*apparently* '*known*' *or* '*given*') *explicanda while explaining them approximately*. This is a view which I have developed fairly fully in my lectures since 1940 (first in a course of lectures given to the Christchurch branch of the Royal Society of New Zealand; cp. the footnote on pp. 134 f. of my *Poverty of Historicism*).

electromagnetic field in the form of a mechanical model of the ether; but in the end he gave up the attempt. With this, the mechanical model lost most of its significance: only the equations which originally were meant to describe the mechanical model of the ether remained. [They were interpreted as describing certain non-mechanical properties of the ether.]

With this transition from a mechanical to an *abstract theory* a stage is reached in the evolution of science at which in practice no more is demanded of explanatory theories than that they can be tested independently; we are ready to work with theories which can be intuitively represented by diagrams such as pictures [or by 'picturable' or 'visualizable' mechanical models], if they are obtainable: this yields 'concrete' theories; or else, if these are not obtainable, we are ready to work with 'abstract' mathematical theories [which, however, may be quite 'understandable' in a sense I have analysed elsewhere].[11]

Our general analysis of the notion of explanation is of course unaffected by the failures of any particular picture or model. It applies to all kinds of abstract theories in the same manner as it applies to mechanical and other models. In fact, models are, from our point of view, nothing but attempts to explain new laws in terms of old laws which have already been tested [together with assumptions about typical initial conditions, or the occurrence of a typical structure—that is to say, the model in a narrower sense]. Models often play important parts in the extension and elaboration of theories; but it is necessary to distinguish a new model in a setting of old theoretical assumptions from a new theory—that is, from a new system of theoretical assumptions.

XII

I HOPE that some of my formulations which at the beginning of this lecture may have seemed to you far-fetched or even paradoxical will now appear less so.

There is no road, royal or otherwise, which leads of necessity from a 'given' set of specific facts to any universal law. What we call 'laws' are hypotheses or conjectures which always form a part of some larger system of theories [in fact, of a whole horizon of expectations] and which, therefore, can never be tested in isolation. The progress of science consists in trials, in the elimination of errors, and in further trials guided by the experience acquired in the course

[11] (Added in the translation.) A fuller analysis of 'understanding' is given in Chapter 4 of the present volume.

of previous trials and errors. No particular theory may ever be regarded as absolutely certain: every theory may become problematical, no matter how well corroborated it may seem now. No scientific theory is sacrosanct or beyond criticism. This fact has often been forgotten, particularly during the last century, when we were impressed by the often repeated and truly magnificent corroborations of certain mechanical theories, which eventually came to be regarded as indubitably true. The stormy development of physics since the turn of the century has taught us better; and we have now come to see that it is the task of the scientist to subject his theory to ever new tests, and that no theory must be pronounced final. Testing proceeds by taking the theory to be tested and combining it with all possible kinds of initial conditions as well as with other theories, and then comparing the resulting predictions with reality. If this leads to disappointed expectations, to refutations, then we have to rebuild our theory.

The disappointment of some of the expectations with which we once eagerly approached reality plays a most significant part in this procedure. It may be compared with the experience of a blind man who touches, or runs into, an obstacle, and so becomes aware of its existence. *It is through the falsification of our suppositions that we actually get in touch with 'reality'*. It is the discovery and elimination of our errors which alone constitute that 'positive' experience which we gain from reality.

It is of course always possible to save a falsified theory by means of supplementary hypotheses [like those of epicycles]. But this is not the way of progress in the sciences. The proper reaction to falsification is to search for new theories which seem likely to offer us a better grasp of the facts. Science is not interested in having the last word if this means shutting off our minds from falsifying experiences, but rather in learning from our experience; that is, in learning from our mistakes.

There is a way of formulating scientific theories which points with particular clarity to the possibility of their falsification: we can formulate them in the form of prohibitions [or *negative existential statements*] such as, for example, 'There does not exist a closed physical system, such that energy changes in one part of it without compensating changes occurring in another part' (first law of thermodynamics). Or, 'There does not exist a machine which is 100 per cent efficient'(second law). It can be shown that universal statements and negative existential statements are logically equivalent. This makes it possible to formulate all universal laws in the manner indicated; that is to say, as prohibitions. However, these are

prohibitions intended only for the technicians and not for the scientist. They tell the former how to proceed if he does not want to squander his energies. But to the scientist they are a challenge to test and to falsify; they stimulate him to try to discover those states of affairs whose existence they prohibit, or deny.

Thus we have reached a point from which we can see science as a magnificent adventure of the human spirit. It is the invention of ever new theories, and the indefatigable examination of their power to throw light on experience. The principles of scientific progress are very simple. They demand that we give up the ancient idea that we may attain certainty [or even a high degree of 'probability' in the sense of the probability calculus] with the propositions and theories of science (an idea which derives from the association of science with magic and of the scientist with the magician): the aim of the scientist is not to discover absolute certainty, but to discover better and better theories [or to invent more and more powerful searchlights] capable of being put to more and more severe tests [and thereby leading us to, and illuminating for us, ever new experiences]. But this means that these theories must be falsifiable: it is through their falsification that science progresses.