

1

SCIENCE: CONJECTURES AND REFUTATIONS

Mr. Turnbull had predicted evil consequences, . . . and was now doing the best in his power to bring about the verification of his own prophecies.

ANTHONY TROLLOPE

I

When I received the list of participants in this course and realized that I had been asked to speak to philosophical colleagues I thought, after some hesitation and consultation, that you would probably prefer me to speak about those problems which interest me most, and about those developments with which I am most intimately acquainted. I therefore decided to do what I have never done before: to give you a report on my own work in the philosophy of science, since the autumn of 1919 when I first began to grapple with the problem, 'When should a

A lecture given at Peterhouse, Cambridge, Summer 1953, as part of a course on developments and trends in contemporary British philosophy, organized by the British Council; originally published under the title 'Philosophy of Science: a Personal Report' in British Philosophy in Mid-Century, ed. C. A. Mace, 1957.

theory be ranked as scientific?' or 'Is there a criterion for the scientific character or status of a theory?'

The problem which troubled me at the time was neither, 'When is a theory true?' nor, 'When is a theory acceptable?' My problem was different. I wished to distinguish between science and pseudo-science; knowing very well that science often errs, and that pseudo-science may happen to stumble on the truth.

I knew, of course, the most widely accepted answer to my problem: that science is distinguished from pseudo-science—or from 'metaphysics'—by its empirical method, which is essentially inductive, proceeding from observation or experiment. But this did not satisfy me. On the contrary, I often formulated my problem as one of distinguishing between a genuinely empirical method and a non-empirical or even a pseudo-empirical method—that is to say, a method which, although it appeals to observation and experiment, nevertheless does not come up to scientific standards. The latter method may be exemplified by astrology, with its stupendous mass of empirical evidence based on observation—on horoscopes and on biographies.

But as it was not the example of astrology which led me to my problem I should perhaps briefly describe the atmosphere in which my problem arose and the examples by which it was stimulated. After the collapse of the Austrian Empire there had been a revolution in Austria: the air was full of revolutionary slogans and ideas, and new and often wild theories. Among the theories which interested me Einstein's theory of relativity was no doubt by far the most important. Three others were Marx's theory of history, Freud's psycho-analysis, and Alfred Adler's so-called 'individual psychology'.

There was a lot of popular nonsense talked about these theories, and especially about relativity (as still happens even today), but I was fortunate in those who introduced me to the study of this theory. We all—the small circle of students to which I belonged—were thrilled with the result of Eddington's eclipse observations which in 1919 brought the first important confirmation of Einstein's theory of gravitation. It was a great experience for us, and one which had a lasting influence on my intellectual development.

The three other theories I have mentioned were also widely discussed among students at that time. I myself happened to come into

personal contact with Alfred Adler, and even to co-operate with him in his social work among the children and young people in the working-class districts of Vienna where he had established social guidance clinics.

It was during the summer of 1919 that I began to feel more and more dissatisfied with these three theories—the Marxist theory of history, psycho-analysis, and individual psychology; and I began to feel dubious about their claims to scientific status. My problem perhaps first took the simple form, ‘What is wrong with Marxism, psycho-analysis, and individual psychology? Why are they so different from physical theories, from Newton’s theory, and especially from the theory of relativity?’

To make this contrast clear I should explain that few of us at the time would have said that we believed in the truth of Einstein’s theory of gravitation. This shows that it was not my doubting the truth of those other three theories which bothered me, but something else. Yet neither was it that I merely felt mathematical physics to be more exact than the sociological or psychological type of theory. Thus what worried me was neither the problem of truth, at that stage at least, nor the problem of exactness or measurability. It was rather that I felt that these other three theories, though posing as sciences, had in fact more in common with primitive myths than with science; that they resembled astrology rather than astronomy.

I found that those of my friends who were admirers of Marx, Freud, and Adler, were impressed by a number of points common to these theories, and especially by their apparent *explanatory power*. These theories appeared to be able to explain practically everything that happened within the fields to which they referred. The study of any of them seemed to have the effect of an intellectual conversion or revelation, opening your eyes to a new truth hidden from those not yet initiated. Once your eyes were thus opened you saw confirming instances everywhere: the world was full of *verifications* of the theory. Whatever happened always confirmed it. Thus its truth appeared manifest; and unbelievers were clearly people who did not want to see the manifest truth; who refused to see it, either because it was against their class interest, or because of their repressions which were still ‘un-analysed’ and crying out for treatment.

The most characteristic element in this situation seemed to me the incessant stream of confirmations, of observations which 'verified' the theories in question; and this point was constantly emphasized by their adherents. A Marxist could not open a newspaper without finding on every page confirming evidence for his interpretation of history; not only in the news, but also in its presentation—which revealed the class bias of the paper—and especially of course in what the paper did not say. The Freudian analysts emphasized that their theories were constantly verified by their 'clinical observations'. As for Adler, I was much impressed by a personal experience. Once, in 1919, I reported to him a case which to me did not seem particularly Adlerian, but which he found no difficulty in analysing in terms of his theory of inferiority feelings, although he had not even seen the child. Slightly shocked, I asked him how he could be so sure. 'Because of my thousandfold experience', he replied; whereupon I could not help saying: 'And with this new case, I suppose, your experience has become thousand-and-one-fold.'

What I had in mind was that his previous observations may not have been much sounder than this new one; that each in its turn had been interpreted in the light of 'previous experience', and at the same time counted as additional confirmation. What, I asked myself, did it confirm? No more than that a case could be interpreted in the light of the theory. But this meant very little, I reflected, since every conceivable case could be interpreted in the light of Adler's theory, or equally of Freud's. I may illustrate this by two very different examples of human behaviour: that of a man who pushes a child into the water with the intention of drowning it; and that of a man who sacrifices his life in an attempt to save the child. Each of these two cases can be explained with equal ease in Freudian and in Adlerian terms. According to Freud the first man suffered from repression (say, of some component of his Oedipus complex), while the second man had achieved sublimation. According to Adler the first man suffered from feelings of inferiority (producing perhaps the need to prove to himself that he dared to commit some crime), and so did the second man (whose need was to prove to himself that he dared to rescue the child). I could not think of any human behaviour which could not be interpreted in terms of either theory. It was precisely this fact—that they always fitted, that

they were always confirmed—which in the eyes of their admirers constituted the strongest argument in favour of these theories. It began to dawn on me that this apparent strength was in fact their weakness.

With Einstein's theory the situation was strikingly different. Take one typical instance—Einstein's prediction, just then confirmed by the findings of Eddington's expedition. Einstein's gravitational theory had led to the result that light must be attracted by heavy bodies (such as the sun), precisely as material bodies were attracted. As a consequence it could be calculated that light from a distant fixed star whose apparent position was close to the sun would reach the earth from such a direction that the star would seem to be slightly shifted away from the sun; or, in other words, that stars close to the sun would look as if they had moved a little away from the sun, and from one another. This is a thing which cannot normally be observed since such stars are rendered invisible in daytime by the sun's overwhelming brightness; but during an eclipse it is possible to take photographs of them. If the same constellation is photographed at night one can measure the distances on the two photographs, and check the predicted effect.

Now the impressive thing about this case is the risk involved in a prediction of this kind. If observation shows that the predicted effect is definitely absent, then the theory is simply refuted. The theory is incompatible with *certain possible results of observation*—in fact with results which everybody before Einstein would have expected.¹ This is quite different from the situation I have previously described, when it turned out that the theories in question were compatible with the most divergent human behaviour, so that it was practically impossible to describe any human behaviour that might not be claimed to be a verification of these theories.

These considerations led me in the winter of 1919–20 to conclusions which I may now reformulate as follows.

(1) It is easy to obtain confirmations, or verifications, for nearly every theory—if we look for confirmations.

(2) Confirmations should count only if they are the result of risky predictions; that is to say, if, unenlightened by the theory in question, we

¹ This is a slight oversimplification, for about half of the Einstein effect may be derived from the classical theory, provided we assume a ballistic theory of light.

should have expected an event which was incompatible with the theory—an event which would have refuted the theory.

(3) Every ‘good’ scientific theory is a prohibition: it forbids certain things to happen. The more a theory forbids, the better it is.

(4) A theory which is not refutable by any conceivable event is non-scientific. Irrefutability is not a virtue of a theory (as people often think) but a vice.

(5) Every genuine test of a theory is an attempt to falsify it, or to refute it. Testability is falsifiability; but there are degrees of testability: some theories are more testable, more exposed to refutation, than others; they take, as it were, greater risks.

(6) Confirming evidence should not count *except* when it is the result of a genuine test of the theory; and this means that it can be presented as a serious but unsuccessful attempt to falsify the theory. (I now speak in such cases of ‘corroborating evidence’.)

(7) Some genuinely testable theories, when found to be false, are still upheld by their admirers—for example by introducing *ad hoc* some auxiliary assumption, or by re-interpreting the theory *ad hoc* in such a way that it escapes refutation. Such a procedure is always possible, but it rescues the theory from refutation only at the price of destroying, or at least lowering, its scientific status. (I later described such a rescuing operation as a ‘conventionalist twist’ or a ‘conventionalist stratagem’.)

One can sum up all this by saying that the criterion of the scientific status of a theory is its falsifiability, or refutability, or testability.

II

I may perhaps exemplify this with the help of the various theories so far mentioned. Einstein’s theory of gravitation clearly satisfied the criterion of falsifiability. Even if our measuring instruments at the time did not allow us to pronounce on the results of the tests with complete assurance, there was clearly a possibility of refuting the theory.

Astrology did not pass the test. Astrologers were greatly impressed, and misled, by what they believed to be confirming evidence—so much so that they were quite unimpressed by any unfavourable evidence. Moreover, by making their interpretations and prophecies sufficiently vague they were able to explain away anything that might

have been a refutation of the theory had the theory and the prophecies been more precise. In order to escape falsification they destroyed the testability of their theory. It is a typical soothsayer's trick to predict things so vaguely that the predictions can hardly fail: that they become irrefutable.

The Marxist theory of history, in spite of the serious efforts of some of its founders and followers, ultimately adopted this soothsaying practice. In some of its earlier formulations (for example in Marx's analysis of the character of the 'coming social revolution') their predictions were testable, and in fact falsified.² Yet instead of accepting the refutations the followers of Marx re-interpreted both the theory and the evidence in order to make them agree. In this way they rescued the theory from refutation; but they did so at the price of adopting a device which made it irrefutable. They thus gave a 'conventionalist twist' to the theory; and by this stratagem they destroyed its much advertised claim to scientific status.

The two psycho-analytic theories were in a different class. They were simply non-testable, irrefutable. There was no conceivable human behaviour which could contradict them. This does not mean that Freud and Adler were not seeing certain things correctly: I personally do not doubt that much of what they say is of considerable importance, and may well play its part one day in a psychological science which is testable. But it does mean that those 'clinical observations' which analysts naïvely believe confirm their theory cannot do this any more than the daily confirmations which astrologers find in their practice.³ And as

² See, for example, my *Open Society and Its Enemies*, ch. 15, section iii, and notes 13–14.

³ 'Clinical observations', like all other observations, are *interpretations in the light of theories* (see below, sections iv ff.); and for this reason alone they are apt to seem to support those theories in the light of which they were interpreted. But real support can be obtained only from observations undertaken as tests (by 'attempted refutations'); and for this purpose *criteria of refutation* have to be laid down beforehand: it must be agreed which observable situations, if actually observed, mean that the theory is refuted. But what kind of clinical responses would refute to the satisfaction of the analyst not merely a particular analytic diagnosis but psycho-analysis itself? And have such criteria ever been discussed or agreed upon by analysts? Is there not, on the contrary, a whole family of analytic concepts, such as 'ambivalence' (I do not suggest that there is no such thing as ambivalence), which would make it difficult, if not impossible, to agree upon such criteria? Moreover, how much headway has been made in investigating the question of the extent

for Freud's epic of the Ego, the Super-ego, and the Id, no substantially stronger claim to scientific status can be made for it than for Homer's collected stories from Olympus. These theories describe some facts, but in the manner of myths. They contain most interesting psychological suggestions, but not in a testable form.

At the same time I realized that such myths may be developed, and become testable; that historically speaking all—or very nearly all—scientific theories originate from myths, and that a myth may contain important anticipations of scientific theories. Examples are Empedocles' theory of evolution by trial and error, or Parmenides' myth of the unchanging block universe in which nothing ever happens and which, if we add another dimension, becomes Einstein's block universe (in which, too, nothing ever happens, since everything is, four-dimensionally speaking, determined and laid down from the beginning). I thus felt that if a theory is found to be non-scientific, or 'metaphysical' (as we might say), it is not thereby found to be unimportant, or insignificant, or 'meaningless', or 'nonsensical'.⁴ But

to which the (conscious or unconscious) expectations and theories held by the analyst influence the 'clinical responses' of the patient? (To say nothing about the conscious attempts to influence the patient by proposing interpretations to him, etc.) Years ago I introduced the term 'Oedipus effect' to describe the influence of a theory or expectation or prediction upon the event which it predicts or describes: it will be remembered that the causal chain leading to Oedipus' parricide was started by the oracle's prediction of this event. This is a characteristic and recurrent theme of such myths, but one which seems to have failed to attract the interest of the analysts, perhaps not accidentally. (The problem of confirmatory dreams suggested by the analyst is discussed by Freud, for example in *Gesammelte Schriften*, III, 1925, where he says on p. 314: 'If anybody asserts that most of the dreams which can be utilized in an analysis . . . owe their origin to [the analyst's] suggestion, then no objection can be made from the point of view of analytic theory. Yet there is nothing in this fact', he surprisingly adds, 'which would detract from the reliability of our results.')

⁴ The case of astrology, nowadays a typical pseudo-science, may illustrate this point. It was attacked, by Aristotelians and other rationalists, down to Newton's day, for the wrong reason—for its now accepted assertion that the planets had an 'influence' upon terrestrial ('sublunar') events. In fact Newton's theory of gravity, and especially the lunar theory of the tides, was historically speaking an offspring of astrological lore. Newton, it seems, was most reluctant to adopt a theory which came from the same stable as for example the theory that 'influenza' epidemics are due to an astral 'influence'. And Galileo, no doubt for the same reason, actually rejected the lunar theory of the tides; and his misgivings about Kepler may easily be explained by his misgivings about astrology.

it cannot claim to be backed by empirical evidence in the scientific sense—although it may easily be, in some genetic sense, the ‘result of observation’.

(There were a great many other theories of this pre-scientific or pseudo-scientific character, some of them, unfortunately, as influential as the Marxist interpretation of history; for example, the racist interpretation of history—another of those impressive and all-explanatory theories which act upon weak minds like revelations.)

Thus the problem which I tried to solve by proposing the criterion of falsifiability was neither a problem of meaningfulness or significance, nor a problem of truth or acceptability. It was the problem of drawing a line (as well as this can be done) between the statements, or systems of statements, of the empirical sciences, and all other statements—whether they are of a religious or of a metaphysical character, or simply pseudo-scientific. Years later—it must have been in 1928 or 1929—I called this first problem of mine the ‘problem of demarcation’. The criterion of falsifiability is a solution to this problem of demarcation, for it says that statements or systems of statements, in order to be ranked as scientific, must be capable of conflicting with possible, **or conceivable, observations.**

III

Today I know, of course, that this criterion of demarcation—the criterion of testability, or falsifiability, or refutability—is far from obvious; for even now its significance is seldom realized. At that time, in 1920, it seemed to me almost trivial, although it solved for me an intellectual problem which had worried me deeply, and one which also had obvious practical consequences (for example, political ones). But I did not yet realize its full implications, or its philosophical significance. When I explained it to a fellow student of the Mathematics Department (now a distinguished mathematician in Great Britain), he suggested that I should publish it. At the time I thought this absurd; for I was convinced that my problem, since it was so important for me, must have agitated many scientists and philosophers who would surely have reached my rather obvious solution. That this was not the case I learnt from Wittgenstein’s work, and from its reception; and so I published my

results thirteen years later in the form of a criticism of Wittgenstein's criterion of meaningfulness.

Wittgenstein, as you all know, tried to show in the *Tractatus* (see for example his propositions 6.53; 6.54; and 5) that all so-called philosophical or metaphysical propositions were actually non-propositions or pseudo-propositions: that they were senseless or meaningless. All genuine (or meaningful) propositions were truth functions of the elementary or atomic propositions which described 'atomic facts'—i.e., facts which can in principle be ascertained by observation. In other words, meaningful propositions were fully reducible to elementary or atomic propositions which were simple statements describing possible states of affairs, and which could in principle be established or rejected by observation. If we call a statement an 'observation statement' not only if it states an actual observation but also if it states anything that may be observed, we shall have to say (according to the *Tractatus*, 5 and 4.52) that every genuine proposition must be a truth-function of, and therefore deducible from, observation statements. All other apparent propositions will be meaningless pseudo-propositions; in fact they will be nothing but nonsensical gibberish.

This idea was used by Wittgenstein for a characterization of science, as opposed to philosophy. We read (for example in 4.11, where natural science is taken to stand in opposition to philosophy): 'The totality of true propositions is the total natural science (or the totality of the natural sciences).' This means that the propositions which belong to science are those deducible from true observation statements; they are those propositions which can be verified by true observation statements. Could we know all true observation statements, we should also know all that may be asserted by natural science.

This amounts to a crude verifiability criterion of demarcation. To make it slightly less crude, it could be amended thus: 'The statements which may possibly fall within the province of science are those which may possibly be verified by observation statements; and these statements, again, coincide with the class of *all* genuine or meaningful statements.' For this approach, then, *verifiability*, *meaningfulness*, and *scientific character* all coincide.

I personally was never interested in the so-called problem of meaning; on the contrary, it appeared to me a verbal problem, a typical

pseudo-problem, I was interested only in the problem of demarcation, i.e. in finding a criterion of the scientific character of theories. It was just this interest which made me see at once that Wittgenstein's verifiability criterion of meaning was intended to play the part of a criterion of demarcation as well; and which made me see that, as such, it was totally inadequate, even if all misgivings about the dubious concept of meaning were set aside. For Wittgenstein's criterion of demarcation—to use my own terminology in this context—is verifiability, or deducibility from observation statements. But this criterion is too narrow (and too wide): it excludes from science practically everything that is, in fact, characteristic of it (while failing in effect to exclude astrology). No scientific theory can ever be deduced from observation statements, or be described as a truth-function of observation statements.

All this I pointed out on various occasions to Wittgensteinians and members of the Vienna Circle. In 1931–2 I summarized my ideas in a largish book (read by several members of the Circle but never published; although part of it was incorporated in my *Logic of Scientific Discovery*); and in 1933 I published a letter to the Editor of *Erkenntnis* in which I tried to compress into two pages my ideas on the problems of demarcation and induction.⁵ In this letter and elsewhere I described the problem of meaning as a pseudo-problem, in contrast to the problem of demarcation. But my contribution was classified by members of

⁵ My *Logic of Scientific Discovery* (1959, 1960, 1961), here usually referred to as L.Sc.D., is the translation of *Logik der Forschung* (1934), with a number of additional notes and appendices, including (on pp. 312–14) the letter to the Editor of *Erkenntnis* mentioned here in the text which was first published in *Erkenntnis*, 3, 1933, pp. 426 f.

Concerning my never published book mentioned here in the text, see R. Carnap's paper 'Ueber Protokollsätze' (On Protocol-Sentences), *Erkenntnis*, 3, 1932, pp. 215–28 where he gives an outline of my theory on pp. 223–8, and accepts it. He calls my theory 'procedure B', and says (p. 224, top): 'Starting from a point of view different from Neurath's' (who developed what Carnap calls on p. 223 'procedure A'), 'Popper developed procedure B as part of his system.' And after describing in detail my theory of tests, Carnap sums up his views as follows (p. 228): 'After weighing the various arguments here discussed, it appears to me that the second language form with procedure B—that is in the form here described—is the most adequate among the forms of scientific language at present advocated . . . in the . . . theory of knowledge.' This paper of Carnap's contained the first published report of my theory of critical testing. (See also my critical remarks in L.Sc.D., note 1 to section 29, p. 104, where the date '1933' should read '1932'; and ch. 11, below, text to note 39.).

the Circle as a proposal to replace the verifiability criterion of *meaning* by a falsifiability criterion of *meaning*—which effectively made nonsense of my views.⁶ My protests that I was trying to solve, not their pseudo-problem of meaning, but the problem of demarcation, were of no avail.

My attacks upon verification had some effect, however. They soon led to complete confusion in the camp of the verificationist philosophers of sense and nonsense. The original proposal of verifiability as the criterion of meaning was at least clear, simple, and forceful. The modifications and shifts which were now introduced were the very opposite.⁷ This, I should say, is now seen even by the participants. But since I am usually quoted as one of them I wish to repeat that although I created this confusion I never participated in it. Neither falsifiability nor testability were proposed by me as criteria of meaning; and although I may plead guilty to having introduced both terms into the discussion, it was not I who introduced them into the theory of meaning.

Criticism of my alleged views was widespread and highly successful. I have yet to meet a criticism of my views.⁸ Meanwhile, testability is being widely accepted as a criterion of demarcation.

⁶ Wittgenstein's example of a nonsensical pseudo-proposition is: 'Socrates is identical'. Obviously, 'Socrates is not identical' must also be nonsense. Thus the negation of any nonsense will be nonsense, and that of a meaningful statement will be meaningful. But the negation of a testable (or falsifiable) statement need not be testable, as was pointed out, first in my *L.Sc.D.*, (e.g. pp. 38 f.) and later by my critics. The confusion caused by taking testability as a criterion of *meaning* rather than of *demarcation* can easily be imagined.

⁷ The most recent example of the way in which the history of this problem is misunderstood is A. R. White's 'Note on Meaning and Verification', *Mind*, **63**, 1954, pp. 66 ff. J. L. Evans's article, *Mind*, **62**, 1953, pp. 1 ff., which Mr. White criticizes, is excellent in my opinion, and unusually perceptive. Understandably enough, neither of the authors can quite reconstruct the story. (Some hints may be found in my *Open Society*, notes 46, 51 and 52 to ch. 11; and a fuller analysis in ch. 11 of the present volume).

⁸ In *L.Sc.D.* I discussed, and replied to, some likely objections which afterwards were indeed raised, without reference to my replies. One of them is the contention that the falsification of a natural law is just as impossible as its verification. The answer is that this objection mixes two entirely different levels of analysis (like the objection that mathematical demonstrations are impossible since checking, no matter how often repeated, can never make it quite certain that we have not overlooked a mistake). On the first level, there is a logical asymmetry: one singular statement—say about the perihelion of Mercury—can formally falsify Kepler's laws; but these cannot be formally verified by any

IV

I have discussed the problem of demarcation in some detail because I believe that its solution is the key to most of the fundamental problems of the philosophy of science. I am going to give you later a list of some of these other problems, but only one of them—the problem of induction—can be discussed here at any length.

I had become interested in the problem of induction in 1923. Although this problem is very closely connected with the problem of demarcation, I did not fully appreciate the connection for about five years.

I approached the problem of induction through Hume. Hume, I felt, was perfectly right in pointing out that induction cannot be logically justified. He held that there can be no valid logical⁹ arguments allowing us to establish ‘that those instances, of which we have had no experience, resemble those, of which we have had experience’. Consequently ‘even after the observation of the frequent or constant conjunction of objects, we have no reason to draw any inference concerning any object beyond those of which we have had experience’. For ‘shou’d it be said that we have experience’¹⁰—experience teaching us that objects constantly conjoined with certain other objects continue to be so conjoined—then, Hume says, ‘I wou’d renew my question, why from this experience we form any conclusion beyond those past instances, of which we have had

number of singular statements. The attempt to minimize this asymmetry can only lead to confusion. On another level, we may hesitate to accept any statement, even the simplest observation statement; and we may point out that every statement involves interpretation in the light of theories, and that it is therefore uncertain. This does not affect the fundamental asymmetry, but it is important: most dissectors of the heart before Harvey observed the wrong things—those, which they expected to see. There can never be anything like a completely safe observation, free from the dangers of misinterpretation. (This is one of the reasons why the theory of induction does not work.) The ‘empirical basis’ consists largely of a mixture of theories of lower degree of universality (of ‘reproducible effects’). But the fact remains that, relative to whatever basis the investigator may accept (at his peril), he can test his theory only by trying to refute it.

⁹ Hume does not say ‘logical’ but ‘demonstrative’, a terminology which, I think, is a little misleading. The following two quotations are from the *Treatise of Human Nature*, Book I, Part III, sections vi and xii. (The italics are all Hume’s.)

¹⁰ This and the next quotation are from *loc. cit.*, section vi. See also Hume’s *Enquiry Concerning Human Understanding*, section iv, Part II, and his *Abstract*, edited 1938 by J. M. Keynes and P. Sraffa, p. 15, and quoted in *L.Sc.D.*, new appendix *vii, text to note 6.

experience'. This 'renew'd question' indicates that an attempt to justify the practice of induction by an appeal to experience must lead to an infinite regress. As a result we can say that theories can never be inferred from observation statements, or rationally justified by them.

I found Hume's refutation of inductive inference clear and conclusive. But I felt completely dissatisfied with his psychological explanation of induction in terms of custom or habit.

It has often been noticed that this explanation of Hume's is philosophically not very satisfactory. Hume, however, without doubt intended it as a *psychological* rather than a philosophical theory; for it tries to give a causal explanation of a psychological fact—the fact that we believe in laws, in statements asserting regularities or constantly conjoined kinds of events. Hume explains this fact by asserting that it is due to (i.e. constantly conjoined with) custom or habit. But even this reformulation of Hume's theory is unacceptable; for what I have just called a 'psychological fact' may itself be described as a custom or habit—our custom or our habit of believing in laws or regularities. It is neither surprising nor enlightening to hear that such a custom or habit can be explained as due to custom or habit, or conjoined with a custom or habit (even though a different one). Only when we remember that the words 'custom' and 'habit' are used by Hume, as they are in ordinary language, not merely to describe regular behaviour, but rather to theorize about its origin (ascribed to frequent repetition), can we reformulate his psychological theory in a more satisfactory way. Hume's theory becomes then the thesis that, like other habits, *our habit of believing in laws is the product of frequent repetition*—of the repeated observation that things of a certain kind are constantly conjoined with things of another kind.

This genetic-psychological theory is, as indicated, incorporated in ordinary language, and it is therefore hardly as revolutionary as Hume thought. It is no doubt an extremely popular psychological theory—part of 'common sense', one might say. But in spite of my love of both common sense and Hume, I felt convinced that this psychological theory was mistaken; and that it was in fact refutable on purely logical grounds.

Hume's psychology, which is the popular psychology, was mistaken, I felt, about at least three different things: (a) the typical result of repetition; (b) the genesis of habits; and especially (c) the character of

those experiences or modes of behaviour which may be described as 'believing in a law' or 'expecting a law-like succession of events'.

(a) The typical result of repetition—say, of repeating a difficult passage on the piano—is that movements which at first needed attention are in the end executed without attention. We might say that the process becomes radically abbreviated, and ceases to be conscious: it becomes automatized, 'physiological'. Such a development, far from creating a conscious expectation of law-like succession, or a belief in a law, may on the contrary begin with a conscious belief and destroy it by making it superfluous. In learning to ride a bicycle we may start with the belief that we can avoid falling if we steer in the direction in which we threaten to fall, and this belief may be useful for guiding our movements. After sufficient practice we may forget the rule; in any case, we do not need it any longer. On the other hand, even if it is true that repetition may create unconscious expectations, these become conscious only if something goes wrong (we may not have heard the clock tick, but we may hear that it has stopped).

(b) Habits or customs do not, as a rule, *originate* in repetition. Even the habit of walking, or of speaking, or of feeding at certain hours, begins before repetition can play any part whatever. We may say, if we like, that they deserve to be called 'habits' or 'customs' only after repetition has played its typical part described under (a); but we must not say that the practices in question *originated* as the result of many repetitions.

(c) Belief in a law is not quite the same thing as behaviour which betrays an expectation of a law-like succession of events; but these two are sufficiently closely connected to be treated together. They may, perhaps, in exceptional cases, result from a mere repetition of sense impressions (as in the case of the stopping clock). I was prepared to concede this, but I contended that normally, and in most cases of any interest, they cannot be so explained. As Hume admits, even a single striking observation may be sufficient to create a belief or an expectation—a fact which he tries to explain as due to an inductive habit, formed as the result of a vast number of long repetitive sequences which had been experienced at an earlier period of life.¹¹

¹¹ *Treatise*, section xiii; section xv, rule 4.

But this, I contended, was merely his attempt to explain away unfavourable facts which threatened his theory; an unsuccessful attempt, since these unfavourable facts could be observed in very young animals and babies—as early, indeed, as we like. ‘A lighted cigarette was held near the noses of the young puppies’, reports F. Bäge. ‘They sniffed at it once, turned tail, and nothing would induce them to come back to the source of the smell and to sniff again. A few days later, they reacted to the mere sight of a cigarette or even of a rolled piece of white paper, by bounding away, and sneezing.’¹² If we try to explain cases like this by postulating a vast number of long repetitive sequences at a still earlier age we are not only romancing, but forgetting that in the clever puppies’ short lives there must be room not only for repetition but also for a great deal of novelty, and consequently of non-repetition.

But it is not only that certain empirical facts do not support Hume; there are decisive arguments of a purely logical nature against his psychological theory.

The central idea of Hume’s psychological theory is that of *repetition, based upon similarity* (or ‘resemblance’). This idea is used in a very uncritical way. We are led to think of the water-drop that hollows the stone: of sequences of unquestionably like events slowly forcing themselves upon us, as does the tick of the clock. But we ought to realize that in a psychological theory such as Hume’s, only repetition-for-us, based upon similarity-for-us, can be allowed to have any effect upon us. We must respond to situations as if they were equivalent; *take* them as similar; *interpret* them as repetitions. In this way they become for us *functionally equal*. The clever puppies, we may assume, showed by their response, their way of acting or of reacting, that they recognized or interpreted the second situation as a repetition of the first: that they expected its main element, the objectionable smell, to be present. The situation was a repetition-for-them because they responded to it by *anticipating* its similarity to the previous one.

This apparently psychological criticism has a purely logical basis which may be summed up in the following simple argument. (It

¹² F. Bäge, ‘Zur Entwicklung, etc.’, *Zeitschrift f. Hundeforschung*, 1933; cp. D. Katz, *Animals and Men*, ch. VI, footnote.

happens to be the one from which I originally started my criticism.) The kind of repetition envisaged by Hume can never be perfect; the cases he has in mind cannot be cases of perfect sameness; they can only be cases of similarity. Thus *they are repetitions only from a certain point of view*. (What has the effect upon me of a repetition may not have this effect upon a spider.) But this means that, for logical reasons, there must always be a point of view—such as a system of expectations, anticipations, assumptions, or interests—*before* there can be any repetition; which point of view, consequently, cannot be merely the result of repetition. (See now also appendix *x, (1), to my L.Sc.D.)

We must thus replace, for the purposes of a psychological theory of the origin of our beliefs, the naïve idea of events which are similar by the idea of events to which we react by interpreting them as being similar. But if this is so (and I can see no escape from it) then Hume's psychological theory of induction leads to an infinite regress, precisely analogous to that other infinite regress which was discovered by Hume himself, and used by him to explode the logical theory of induction. For what do we wish to explain? In the example of the puppies we wish to explain behaviour which may be described as *recognizing or interpreting* a situation as a repetition of another. Clearly, we cannot hope to explain this by an appeal to earlier repetitions, once we realize that the earlier repetitions must also have been repetitions-for-them, so that precisely the same problem arises again: that of *recognizing or interpreting* a situation as a repetition of another.

To put it more concisely, similarity-for-us is the product of a response involving interpretations (which may be inadequate) and anticipations or expectations (which may never be fulfilled). It is therefore impossible to explain anticipations, or expectations, as resulting from many repetitions, as suggested by Hume. For even the first repetition-for-us must be based upon similarity-for-us, and therefore upon expectations—precisely the kind of thing we wished to explain. (Expectations must come first, *before* repetitions.)

We see that there is an infinite regress involved in Hume's psychological theory.

Hume, I felt, had never accepted the full force of his own logical analysis. Having refuted the logical idea of induction he was faced with the following problem: how do we actually obtain our knowledge, as a

matter of psychological fact, if induction is a procedure which is logically invalid and rationally unjustifiable? There are two possible answers: (1) We obtain our knowledge by a non-inductive procedure. This answer would have allowed Hume to retain a form of rationalism. (2) We obtain our knowledge by repetition and induction, and therefore by a logically invalid and rationally unjustifiable procedure, so that all apparent knowledge is merely a kind of belief—belief based on habit. This answer would imply that even scientific knowledge is irrational, so that rationalism is absurd, and must be given up. (I shall not discuss here the age-old attempts, now again fashionable, to get out of the difficulty by asserting that though induction is of course logically invalid if we mean by ‘logic’ the same as ‘deductive logic’, it is not irrational by its own standards, and as inductive logic admits; as may be seen from the fact that every reasonable man applies it *as a matter of fact*. As against this, it was Hume’s great achievement to break this uncritical identification of the question of fact—*quid facti?*—and the question of justification or validity—*quid juris?*. (See below, point (13) of the appendix to the present chapter.)

It seems that Hume never seriously considered the first alternative. Having cast out the logical theory of induction by repetition he struck a bargain with common sense, meekly allowing the re-entry of induction by repetition, in the guise of a psychological fact. I proposed to turn the tables upon this theory of Hume’s. Instead of explaining our propensity to expect regularities as the result of repetition, I proposed to explain repetition-for-us as the result of our propensity to expect regularities and to search for them.

Thus I was led by purely logical considerations to replace the psychological theory of induction by the following view. Without waiting, passively, for repetitions to impress or impose regularities upon us, we actively try to impose regularities upon the world. We try to discover similarities in it, and to interpret it in terms of laws invented by us. Without waiting for premises we jump to conclusions. These may have to be discarded later, should observation show that they are wrong.

This was a theory of trial and error—of *conjectures and refutations*. It made it possible to understand why our attempts to force interpretations upon the world were logically prior to the observation of similarities. Since there were logical reasons behind this procedure, I

thought that it would apply in the field of science also; that scientific theories were not the digest of observations, but that they were inventions—conjectures boldly put forward for trial, to be eliminated if they clashed with observations; with observations which were rarely accidental but as a rule undertaken with the definite intention of testing a theory by obtaining, if possible, a decisive refutation.

V

The belief that science proceeds from observation to theory is still so widely and so firmly held that my denial of it is often met with incredulity. I have even been suspected of being insincere—of denying what nobody in his senses can doubt.

But in fact the belief that we can start with pure observations alone, without anything in the nature of a theory, is absurd; as may be illustrated by the story of the man who dedicated his life to natural science, wrote down everything he could observe, and bequeathed his priceless collection of observations to the Royal Society to be used as inductive evidence. This story should show us that though beetles may profitably be collected, observations may not.

Twenty-five years ago I tried to bring home the same point to a group of physics students in Vienna by beginning a lecture with the following instructions: 'Take pencil and paper; carefully observe, and write down what you have observed!' They asked, of course, what I wanted them to observe. Clearly the instruction, 'Observe!' is absurd.¹³ (It is not even idiomatic, unless the object of the transitive verb can be taken as understood.) Observation is always selective. It needs a chosen object, a definite task, an interest, a point of view, a problem. And its description presupposes a descriptive language, with property words; it presupposes similarity and classification, which in their turn presuppose interests, points of view, and problems. 'A hungry animal', writes Katz,¹⁴ 'divides the environment into edible and inedible things. An animal in flight sees roads to escape and hiding places . . . Generally speaking, objects change . . . according to the needs of the animal.' We

¹³ See section 30 of L.Sc.D.

¹⁴ Katz, *loc. cit.*

may add that objects can be classified, and can become similar or dissimilar, *only* in this way—by being related to needs and interests. This rule applies not only to animals but also to scientists. For the animal a point of view is provided by its needs, the task of the moment, and its expectations; for the scientist by his theoretical interests, the special problem under investigation, his conjectures and anticipations, and the theories which he accepts as a kind of background: his frame of reference, his ‘horizon of expectations’.

The problem ‘Which comes first, the hypothesis (H) or the observation (O)?’ is soluble; as is the problem, ‘Which comes first, the hen (H) or the egg (O)?’. The reply to the latter is, ‘An earlier kind of egg’; to the former, ‘An earlier kind of hypothesis’. It is quite true that any particular hypothesis we choose will have been preceded by observations—the observations, for example, which it is designed to explain. But these observations, in their turn, presupposed the adoption of a frame of reference: a frame of expectations: a frame of theories. If they were significant, if they created a need for explanation and thus gave rise to the invention of a hypothesis, it was because they could not be explained within the old theoretical framework, the old horizon of expectations. There is no danger here of an infinite regress. Going back to more and more primitive theories and myths we shall in the end find unconscious, inborn expectations.

The theory of inborn *ideas* is absurd, I think; but every organism has inborn *reactions* or *responses*; and among them, responses adapted to impending events. These responses we may describe as ‘expectations’ without implying that these ‘expectations’ are conscious. The newborn baby ‘expects’, in this sense, to be fed (and, one could even argue, to be protected and loved). In view of the close relation between expectation and knowledge we may even speak in quite a reasonable sense of ‘inborn knowledge’. This ‘knowledge’, however, is not *valid a priori*; an inborn expectation, no matter how strong and specific, may be mistaken. (The newborn child may be abandoned, and starve.)

Thus we are born with expectations; with ‘knowledge’ which, although not *valid a priori*, is *psychologically or genetically a priori*, i.e. prior to all observational experience. One of the most important of these expectations is the expectation of finding a regularity. It is connected with an inborn propensity to look out for regularities, or with a *need* to

find regularities, as we may see from the **pleasure of the child who satisfies this need.**

This 'instinctive' expectation of finding regularities, which is psychologically *a priori*, corresponds very closely to the 'law of causality' which Kant believed to be part of our mental outfit and to be *a priori* valid. One might thus be inclined to say that Kant failed to distinguish between psychologically *a priori* ways of thinking or responding and *a priori* valid beliefs. But I do not think that his mistake was quite as crude as that. For the expectation of finding regularities is not only psychologically *a priori*, but also logically *a priori*: it is logically prior to all observational experience, for it is prior to any recognition of similarities, as we have seen; and all observation involves the recognition of similarities (or dissimilarities). But in spite of being logically *a priori* in this sense the expectation is not valid *a priori*. For it may fail: we can easily construct an environment (it would be a lethal one) which, compared with our ordinary environment, is so chaotic that we completely fail to find regularities. (All natural laws could remain valid: environments of this kind have been used in the animal experiments mentioned in the next section.)

Thus Kant's reply to Hume came near to being right; for the distinction between an *a priori* valid expectation and one which is both genetically and logically prior to observation, but not *a priori* valid, is really somewhat subtle. But Kant proved too much. In trying to show how knowledge is possible, he proposed a theory which had the unavoidable consequence that our quest for knowledge must necessarily succeed, which is clearly mistaken. When Kant said, 'Our intellect does not draw its laws from nature but imposes its laws upon nature', he was right. But in thinking that these laws are necessarily true, or that we necessarily succeed in imposing them upon nature, he was wrong.¹⁵

¹⁵ Kant believed that Newton's dynamics was *a priori* valid. (See his *Metaphysical Foundations of Natural Science*, published between the first and the second editions of the *Critique of Pure Reason*.) But if, as he thought, we can explain the validity of Newton's theory by the fact that our intellect imposes its laws upon nature, it follows, I think, that our intellect must succeed in this; which makes it hard to understand why *a priori* knowledge such as Newton's should be so hard to come by. A somewhat fuller statement of this criticism can be found in ch. 2, especially section x, and chs. 7 and 8 of the present volume.

Nature very often resists quite successfully, forcing us to discard our laws as refuted; but if we live we may try again.

To sum up this logical criticism of Hume's psychology of induction **we may consider the idea of building an induction machine.** Placed in a simplified 'world' (for example, one of sequences of coloured counters) such a machine may through repetition 'learn', or even 'formulate', laws of succession which hold in its 'world'. If such a machine can be constructed (and I have no doubt that it can) then, it might be argued, my theory must be wrong; for if a machine is capable of performing inductions on the basis of repetition, there can be no logical reasons preventing us from doing the same.

The argument sounds convincing, but it is mistaken. In constructing an induction machine we, the architects of the machine, must decide *a priori* what constitutes its 'world'; what things are to be taken as similar or equal; and what kind of 'laws' we wish the machine to be able to 'discover' in its 'world'. In other words we must build into the machine a framework determining what is relevant or interesting in its world: the machine will have its 'inborn' selection principles. The problems of similarity will have been solved for it by its makers who thus have interpreted the **'world' for the machine.**

VI

Our propensity to look out for regularities, and to impose laws upon nature, leads to the psychological phenomenon of *dogmatic thinking* or, more generally, dogmatic behaviour: we expect regularities everywhere and attempt to find them even where there are none; events which do not yield to these attempts we are inclined to treat as a kind of 'background noise'; and we stick to our expectations even when they are inadequate and we ought to accept defeat. This dogmatism is to some extent necessary. It is demanded by a situation which can only be dealt with by forcing our conjectures upon the world. Moreover, this dogmatism allows us to approach a good theory in stages, by way of approximations: if we accept defeat too easily, we may prevent ourselves from finding that we were very nearly right.

It is clear that this *dogmatic attitude*, which makes us stick to our first impressions, is indicative of a strong belief; while a *critical attitude*,