10. STUDIES IN THE LOGIC OF EXPLANATION[1]

1. INTRODUCTION

TO EXPLAIN the phenomena in the world of our experience, to answer the question "why?" rather than only the question "what?" is one of the foremost objectives of empirical science. While there is rather general agreement on this point there exists considerable difference of opinion as to the function and the essential characteristics of scientific explanation. The present essay is an attempt to shed some light on these issues by means of an elementary survey of the basic pattern of scientific explanation and a subsequent more rigorous analysis of the concept of law and the logical structure of explanatory arguments.

1. This essay grew out of discussions with Dr. Paul Oppenheim; it was published in co-authorship with him and is here reprinted with his permission. Our individual contributions cannot be separated in detail; the present author is responsible, however, for the substance of Part IV and for the final formulation of the entire text.

Some of the ideas set forth in Part II originated with our common friend, Dr. Kurt Grelling, who suggested them to us in a discussion carried on by correspondence. Grelling and his wife subsequently became victims of the Nazi terror during the Second World War; by including in this essay at least some of Grelling's contributions, which are explicitly identified, we hope to realize his wish that his ideas on this subject might not entirely fall into oblivion.

Paul Oppenheim and I are much indebted to Professors Rudolf Carnap, Herbert Feigl, Nelson Goodman, and W. V. Quine for stimulating discussions and constructive criticism.

This article was published in *Philosophy of Science*, vol. 15, pp. 135–75.

The elementary survey is presented in Part I; Part II contains an analysis of the concept of emergence; Part III seeks to exhibit and to clarify in a more rigorous manner some of the peculiar and perplexing logical problems to which the familiar elementary analysis of explanation gives rise. Part IV, finally, deals with the idea of explanatory power of a theory; an explicit definition and a formal theory of this concept are developed for the case of a scientific language of simple logical structure.

PART I. ELEMENTARY SURVEY OF SCIENTIFIC EXPLANATION

2. Some Illustrations. A mercury thermometer is rapidly immersed in hot water; there occurs a temporary drop of the mercury column, which is then followed by a swift rise. How is this phenomenon to be explained? The increase in temperature affects at first only the glass tube of the thermometer; it expands and thus provides a larger space for the mercury inside, whose surface therefore drops. As soon as by heat conduction the rise in temperature reaches the mercury, however, the latter expands, and as its coefficient of expansion is considerably larger than that of glass, a rise of the mercury level results. – This account consists of statements of two kinds. Those of the first kind indicate certain conditions which are realized prior to, or at the same time as, the phenomenon to be explained; we shall refer to them briefly as antecedent conditions. In our illustration, the antecedent conditions include, among others, the fact that the thermometer consists of a glass tube which is partly filled with mercury, and that it is immersed into hot water. The statements of the second kind express certain general laws; in our case, these include the laws of the thermic expansion of mercury and of glass, and a statement about the small thermic conductivity of glass. The two sets of statements, if adequately and completely formulated, explain the phenomenon under consideration: they entail the consequence that the mercury will first drop, then rise. Thus, the event under discussion is explained by subsuming it under general laws, i.e., by showing that it occurred in accordance with those laws, in virtue of the realization of certain specified antecedent conditions.

Consider another illustration. To an observer in a rowboat, that part of an oar which is under water appears to be bent upwards. The phenomenon is explained by means of general laws—mainly the law of refraction and the law that water is an optically denser medium than air—and by reference to certain antecedent conditions—especially the facts that part of the oar is in the water, part in the air, and that the oar is practically a straight piece of wood. Thus, here again, the question "*Why* does the phenomenon occur?" is construed as meaning "according to what general laws, and by virtue of what antecedent conditions does the phenomenon occur?"

So far, we have considered only the explanation of particular events occurring at a certain time and place. But the question "Why?" may be raised also in regard to general laws. Thus, in our last illustration, the question might be asked: Why does the propagation of light conform to the law of refraction? Classical physics answers in terms of the undulatory theory of light, i.e. by stating that the propagation of light is a wave phenomenon of a certain general type, and that all wave phenomena of that type satisfy the law of refraction. Thus, the explanation of a general regularity consists in subsuming it under another, more comprehensive regularity, under a more general law. Similarly, the validity of Galileo's law for the free fall of bodies near the earth's surface can be explained by deducing it from a more comprehensive set of laws, namely Newton's laws of motion and his law of gravitation, together with some statements about particular facts, namely, about the mass and the radius of the earth.

3. The Basic Pattern of Scientific Explanation. From the preceding sample cases let us now abstract some general characteristics of scientific explanation. We divide an explanation into two major constituents, the *explanandum* and the *explanans*[2]. By the explanandum, we understand the sentence describing the phenomenon to be explained (not that phenomenon itself); by the explanans, the class of those sentences which are adduced to account for the phenomenon. As was noted before, the explanans falls into two subclasses; one of these contains certain sentences $C_1, C_2, \ldots, C_k$ which state specific antecedent conditions; the other is a set of sentences $L_1, L_2, \ldots, L_r$ which represent general laws.

If a proposed explanation is to be sound, its constituents have to satisfy certain conditions of adequacy, which may be divided into logical and empirical conditions. For the following discussion, it will be sufficient to formulate these requirements in a slightly vague manner; in Part III, a more precise restatement of these criteria will be presented.

I. *Logical conditions of adequacy*

(R1) The explanandum must be a logical consequence of the explanans; in other words, the explanandum must be logically deducible from the information contained in the explanans; for otherwise, the explanans would not constitute adequate grounds for the explanandum.

2. These two expressions, derived from the Latin *explanare*, were adopted in preference to the perhaps more customary terms "explicandum" and "explicans" in order to reserve the latter for use in the context of explication of meaning, or analysis. On explication in this sense, cf. Carnap (1945a), p. 513.

(R2) The explanans must contain general laws, and these must actually be required for the derivation of the explanandum. We shall not make it a necessary condition for a sound explanation, however, that the explanans must contain at least one statement which is not a law; for, to mention just one reason, we would surely want to consider as an explanation the derivation of the general regularities governing the motion of double stars from the laws of celestial mechanics, even though all the statements in the explanans are general laws.

(R3) The explanans must have empirical content; i.e., it must be capable, at least in principle, of test by experiment or observation. This condition is implicit in (R1); for since the explanandum is assumed to describe some empirical phenomenon, it follows from (R1) that the explanans entails at least one consequence of empirical character, and this fact confers upon it testability and empirical content. But the point deserves special mention because, as will be seen in §4, certain arguments which have been offered as explanations in the natural and in the social sciences violate this requirement.

II. *Empirical condition of adequacy*

(R4) The sentences constituting the explanans must be true.

That in a sound explanation, the statements constituting the explanans have to satisfy some condition of factual correctness is obvious. But it might seem more appropriate to stipulate that the explanans has to be highly confirmed by all the relevant evidence available rather than that it should be true. This stipulation, however, leads to awkward consequences. Suppose that a certain phenomenon was explained at an earlier stage of science, by means of an explanans which was well supported by the evidence then at hand, but which has been highly disconfirmed by more recent empirical findings. In such a case, we would have to say that originally the explanatory account was a correct explanation, but that it ceased to be one later, when unfavorable evidence was discovered. This does not appear to accord with sound common usage, which directs us to say that on the basis of the limited initial evidence, the truth of the explanans, and thus the soundness of the explanation, had been quite probable, but that the ampler evidence now available makes it highly probable that the explanans is not true, and hence that the account in question is

not—and never has been—a correct explanation.[3] (A similar point will be made and illustrated, with respect to the requirement of truth for laws, in the beginning of §6.)

Some of the characteristics of an explanation which have been indicated so far may be summarized in the following schema:

	$C_1, C_2, \ldots, C_k$	Statements of antecedent conditions	Explanans
	$L_1, L_2, \ldots, L_r$	General Laws	
Logical deduction →	E	Description of the empirical phenomenon to be explained	Explanandum

Let us note here that the same formal analysis, including the four necessary conditions, applies to scientific prediction as well as to explanation. The difference between the two is of a pragmatic character. If E is given, i.e. if we know that the phenomenon described by E has occurred, and a suitable set of statements $C_1, C_2, \ldots, C_k, L_1, L_2, \ldots, L_r$ is provided afterwards, we speak of an explanation of the phenomenon in question. If the latter statements are given and E is derived prior to the occurrence of the phenomenon it describes, we speak of a prediction. It may be said, therefore, that an explanation of a particular event is not fully adequate unless its explanans, if taken account of in time, could have served as a basis for predicting the event in question. Consequently, whatever will be said in this article concerning the logical characteristics of explanation or prediction will be applicable to either, even if only one of them should be mentioned.[4]

Many explanations which are customarily offered, especially in prescientific discourse, lack this potential predictive force, however. Thus, we may be told that a car turned over on the road "because" one of its tires blew out while the car was traveling at high speed. Clearly, on the basis of just this information, the accident could not have been predicted, for the explanans provides no explicit general laws by means of which the prediction might be effected, nor does it state adequately the antecedent conditions which would be needed for the

3. (Added in 1964.) Requirement (R4) characterizes what might be called a correct or *true explanation*. In an analysis of the logical structure of explanatory arguments, therefore, that requirement may be disregarded. This is, in fact, what is done in section 7, where the concept of *potential explanation* is introduced. On these and related distinctions, see also section 2.1 of the essay "Aspects of Scientific Explanation" in this volume.

4. (Added in 1964.) This claim is examined in much fuller detail, and reasserted with certain qualifications, in sections 2.4 and 3.5 of the essay "Aspects of Scientific Explanation" in this volume.

prediction. The same point may be illustrated by reference to W. S. Jevons's view that every explanation consists in pointing out a resemblance between facts, and that in some cases this process may require no reference to laws at all and "may involve nothing more than a single identity, as when we explain the appearance of shooting stars by showing that they are identical with portions of a comet."[5] But clearly, this identity does not provide an explanation of the phenomenon of shooting stars unless we presuppose the laws governing the development of heat and light as the effect of friction. The observation of similarities has explanatory value only if it involves at least tacit reference to general laws.

In some cases, incomplete explanatory arguments of the kind here illustrated suppress parts of the explanans simply as "obvious"; in other cases, they seem to involve the assumption that while the missing parts are not obvious, the incomplete explanans could at least, with appropriate effort, be so supplemented as to make a strict derivation of the explanandum possible. This assumption may be justifiable in some cases, as when we say that a lump of sugar disappeared "because" it was put into hot tea, but it surely is not satisfied in many other cases. Thus, when certain peculiarities in the work of an artist are explained as outgrowths of a specific type of neurosis, this observation may contain significant clues, but in general it does not afford a sufficient basis for a potential prediction of those peculiarities. In cases of this kind, an incomplete explanation may at best be considered as indicating some positive correlation between the antecedent conditions adduced and the type of phenomenon to be explained, and as pointing out a direction in which further research might be carried on in order to complete the explanatory account.

The type of explanation which has been considered here so far is often referred to as causal explanation.[6] If E describes a particular event, then the antecedent circumstances described in the sentences $C_1, C_2, \ldots, C_k$ may be said jointly to "cause" that event, in the sense that there are certain empirical regularities, expressed by the laws $L_1, L_2, \ldots, L_r$, which imply that whenever conditions of the kind indicated by $C_1, C_2, \ldots, C_k$ occur, an event of the kind described in E will take place. Statements such as $L_1, L_2, \ldots, L_r$, which assert general and unexceptional connections between specified characteristics of events, are customarily called causal, or deterministic, laws. They must be distinguished from the so-called statistical laws which assert that in the long run, an explicitly stated percentage of all cases satisfying a given set of conditions are accompanied by an event of a certain specified kind. Certain cases of scientific explanation

5. (1924) p. 533.

6. (Added in 1964.) Or rather, causal explanation is one variety of the deductive type of explanation here under discussion; see section 2.2 of "Aspects of Scientific Explanation."

involve "subsumption" of the explanandum under a set of laws of which at least some are statistical in character. Analysis of the peculiar logical structure of that type of subsumption involves difficult special problems. The present essay will be restricted to an examination of the deductive type of explanation, which has retained its significance in large segments of contemporary science, and even in some areas where a more adequate account calls for reference to statistical laws.[7]

4. Explanation in the Nonphysical Sciences. Motivational and Teleological Approaches. Our characterization of scientific explanation is so far based on a study of cases taken from the physical sciences. But the general principles thus obtained apply also outside this area.[8] Thus, various types of behavior in laboratory animals and in human subjects are explained in psychology by subsumption under laws or even general theories of learning or conditioning; and while frequently the regularities invoked cannot be stated with the same generality and precision as in physics or chemistry, it is clear at least that the general character of those explanations conforms to our earlier characterization.

Let us now consider an illustration involving sociological and economic factors. In the fall of 1946, there occurred at the cotton exchanges of the United States a price drop which was so severe that the exchanges in New York, New

7. The account given above of the general characteristics of explanation and prediction in science is by no means novel; it merely summarizes and states explicitly some fundamental points which have been recognized by many scientists and methodologists.

Thus, e.g., Mill says: "An individual fact is said to be explained, by pointing out its cause, that is, by stating the law or laws of causation, of which its production is an instance", and "a law or uniformity in nature is said to be explained, when another law or laws are pointed out, of which that law itself is but a case, and from which it could be deduced." (1858, Book III, Chapter XII, section 1). Similarly, Jevons, whose general characterization of explanation was critically discussed above, stresses that "the most important process of explanation consists in showing that an observed fact is one case of a general law or tendency." (1924, p. 533). Ducasse states the same point as follows: "Explanation essentially consists in the offering of a hypothesis of fact, standing to the fact to be explained as case of antecedent to case of consequent of some already known law of connection." (1925, pp. 150-51). A lucid analysis of the fundamental structure of explanation and prediction was given by Popper in (1935), section 12, and, in an improved version, in his work (1945), especially in Chapter 25 and in note 7 for that chapter.— For a recent characterization of explanation as subsumption under general theories, cf., for example, Hull's concise discussion in (1943a), chapter I. A clear elementary examination of certain aspects of explanation is given in Hospers (1946), and a concise survey of many of the essentials of scientific explanation which are considered in the first two parts of the present study may be found in Feigl (1945), pp. 284 ff.

8. On the subject of explanation in the social sciences, especially in history, cf. also the following publications, which may serve to supplement and amplify the brief discussion to be presented here: Hempel (1942); Popper (1945); White (1943); and the articles *Cause* and *Understanding* in Beard and Hook (1946).

Orleans, and Chicago had to suspend their activities temporarily. In an attempt to explain this occurrence, newspapers traced it back to a large-scale speculator in New Orleans who had feared his holdings were too large and had therefore begun to liquidate his stocks; smaller speculators had then followed his example in a panic and had thus touched off the critical decline. Without attempting to assess the merits of the argument, let us note that the explanation here suggested again involves statements about antecedent conditions and the assumption of general regularities. The former include the facts that the first speculator had large stocks of cotton, that there were smaller speculators with considerable holdings, that there existed the institution of the cotton exchanges with their specific mode of operation, etc. The general regularities referred to are—as often in semi-popular explanations—not explicitly mentioned; but there is obviously implied some form of the law of supply and demand to account for the drop in cotton prices in terms of the greatly increased supply under conditions of practically unchanged demand; besides, reliance is necessary on certain regularities in the behavior of individuals who are trying to preserve or improve their economic position. Such laws cannot be formulated at present with satisfactory precision and generality, and therefore, the suggested explanation is surely incomplete, but its intention is unmistakably to account for the phenomenon by integrating it into a general pattern of economic and socio-psychological regularities.

We turn to an explanatory argument taken from the field of linguistics.[9] In Northern France, there are in use a large variety of words synonymous with the English 'bee', whereas in Southern France, essentially only one such word is in existence. For this discrepancy, the explanation has been suggested that in the Latin epoch, the South of France used the word 'apicula', the North the word 'apis'. The latter, because of a process of phonologic decay in Northern France, became the monosyllabic word 'é'; and monosyllables tend to be eliminated, especially if they contain few consonantic elements, for they are apt to give rise to misunderstandings. Thus, to avoid confusion, other words were selected. But 'apicula', which was reduced to 'abelho', remained clear enough and was retained, and finally it even entered into the standard language, in the form 'abeille'. While the explanation here described is incomplete in the sense characterized in the previous section, it clearly exhibits reference to specific antecedent conditions as well as to general laws.[10]

9. The illustration is taken from Bonfante (1946), section 3.

10. While in each of the last two illustrations, certain regularities are unquestionably relied upon in the explanatory argument, it is not possible to argue convincingly that the intended laws, which at present cannot all be stated explicitly, are of a causal rather than a statistical character. It is quite possible that most or all of the regularities which will be discovered as sociology develops will be of a statistical type. Cf., on this point, the suggestive observations in Zilsel (1941), section 8, and (1941a). This issue does not affect, however, the main point we wish to make here, namely that in the social no less than in the physical sciences, subsumption under general regularities is indispensable for the explanation and the theoretical understanding of any phenomenon.

While illustrations of this kind tend to support the view that explanation in biology, psychology, and the social sciences has the same structure as in the physical sciences, the opinion is rather widely held that in many instances, the causal type of explanation is essentially inadequate in fields other than physics and chemistry, and especially in the study of purposive behavior. Let us examine briefly some of the reasons which have been adduced in support of this view.

One of the most familiar among them is the idea that events involving the activities of humans singly or in groups have a peculiar uniqueness and irrepeatability which makes them inaccessible to causal explanation because the latter, with its reliance upon uniformities, presupposes repeatability of the phenomena under consideration. This argument which, incidentally, has also been used in support of the contention that the experimental method is inapplicable in psychology and the social sciences, involves a misunderstanding of the logical character of causal explanation. Every individual event, in the physical sciences no less than in psychology or the social sciences, is unique in the sense that it, with all its peculiar characteristics, does not repeat itself. Nevertheless, individual events may conform to, and thus be explainable by means of, general laws of the causal type. For all that a causal law asserts is that any event of a specified kind, i.e. any event having certain specified characteristics, is accompanied by another event which in turn has certain specified characteristics; for example, that in any event involving friction, heat is developed. And all that is needed for the testability and applicability of such laws is the recurrence of events with the antecedent characteristics, i.e. the repetition of those characteristics, but not of their individual instances. Thus, the argument is inconclusive. It gives occasion, however, to emphasize an important point concerning our earlier analysis: When we spoke of the explanation of a single event, the term "event" referred to the occurrence of some more or less complex characteristic in a specific spatio-temporal location or in a certain individual object, and not to *all* the characteristics of that object, or to all that goes on in that space-time region.

A second argument that should be mentioned here[11] contends that the establishment of scientific generalizations—and thus of explanatory principles—for

11. Cf., for example, F. H. Knight's presentation of this argument in (1924), pp. 251-52.

human behavior is impossible because the reactions of an individual in a given situation depend not only upon that situation, but also upon the previous history of the individual. But surely, there is no *a priori* reason why generalizations should not be attainable which take into account this dependence of behavior on the past history of the agent. That indeed the given argument "proves" too much, and is therefore a *non sequitur*, is made evident by the existence of certain physical phenomena, such as magnetic hysteresis and elastic fatigue, in which the magnitude of a specific physical effect depends upon the past history of the system involved, and for which nevertheless certain general regularities have been established.

A third argument insists that the explanation of any phenomenon involving purposive behavior calls for reference to motivations and thus for teleological rather than causal analysis. For example, a fuller statement of the suggested explanation for the break in the cotton prices would have to indicate the large-scale speculator's motivations as one of the factors determining the event in question. Thus, we have to refer to goals sought; and this, so the argument runs, introduces a type of explanation alien to the physical sciences. Unquestionably, many of the—frequently incomplete—explanations which are offered for human actions involve reference to goals and motives; but does this make them essentially different from the causal explanations of physics and chemistry? One difference which suggests itself lies in the circumstance that in motivated behavior, the future appears to affect the present in a manner which is not found in the causal explanations of the physical sciences. But clearly, when the action of a person is motivated, say, by the desire to reach a certain objective, then it is not the as yet unrealized future event of attaining that goal which can be said to determine his present behavior, for indeed the goal may never be actually reached; rather—to put it in crude terms—it is (a) his desire, present before the action, to attain that particular objective, and (b) his belief, likewise present before the action, that such and such a course of action is most likely to have the desired effect. The determining motives and beliefs, therefore, have to be classified among the antecedent conditions of a motivational explanation, and there is no formal difference on this account between motivational and causal explanation.

Neither does the fact that motives are not accessible to direct observation by an outside observer constitute an essential difference between the two kinds of explanation; for the determining factors adduced in physical explanations also are very frequently inaccessible to direct observation. This is the case, for instance, when opposite electric charges are adduced in explanation of the mutual attraction of two metal spheres. The presence of those charges, while eluding direct observation, can be ascertained by various kinds of indirect test,

and that is sufficient to guarantee the empirical character of the explanatory statement. Similarly, the presence of certain motivations may be ascertainable only by indirect methods, which may include reference to linguistic utterances of the subject in question, slips of pen or tongue, etc.; but as long as these methods are "operationally determined" with reasonable clarity and precision, there is no essential difference in this respect between motivational explanation and causal explanation in physics.

A potential danger of explanation by motives lies in the fact that the method lends itself to the facile construction of *ex post facto* accounts without predictive force. An action is often explained by attributing it to motives conjectured only after the action has taken place. While this procedure is not in itself objectionable, its soundness, requires that (1) the motivational assumptions in question be capable of test, and (2) that suitable general laws be available to lend explanatory power to the assumed motives. Disregard of these requirements frequently deprives alleged motivational explanations of their cognitive significance.

The explanation of an action in terms of the agent's motives is sometimes considered as a special kind of teleological explanation. As was pointed out above, motivational explanation, if adequately formulated, conforms to the conditions for causal explanation, so that the term "teleological" is a misnomer if it is meant to imply either a non-causal character of the explanation or a peculiar determination of the present by the future. If this is borne in mind, however, the term "teleological" may be viewed, in this context, as referring to causal explanations in which some of the antecedent conditions are motives of the agent whose actions are to be explained.[12]

Teleological explanations of this kind have to be distinguished from a much more sweeping type, which has been claimed by certain schools of thought to be indispensable especially in biology. It consists in explaining characteristics of an organism by reference to certain ends or purposes which the characteristics are said to serve. In contradistinction to the cases examined before, the ends are not assumed here to be consciously or subconsciously pursued by the organism in question. Thus, for the phenomenon of mimicry, the explanation is sometimes offered that it serves the purpose of protecting the animals endowed with it from detection by its pursuers and thus tends to preserve the species. Before teleological hypotheses of this kind can be appraised as to their

12. For a detailed logical analysis of the concept of motivation in psychological theory, see Koch (1941). A stimulating discussion of teleological behavior from the standpoint of contemporary physics and biology is contained in the article (1943) by Rosenblueth, Wiener, and Bigelow. The logic of explanation by motivating reasons is examined more fully in section 10 of the essay "Aspects of Scientific Explanation" in the present volume.

potential explanatory power, their meaning has to be clarified. If they are intended somehow to express the idea that the purposes they refer to are inherent in the design of the universe, then clearly they are not capable of empirical test and thus violate the requirement (R3) stated in §3. In certain cases, however, assertions about the purposes of biological characteristics may be translatable into statements in non-teleological terminology which assert that those characteristics function in a specific manner which is essential to keeping the organism alive or to preserving the species.[13] An attempt to state precisely what is meant by this latter assertion—or by the similar one that without those characteristics, and other things being equal, the organism or the species would not survive—encounters considerable difficulties. But these need not be discussed here. For even if we assume that biological statements in teleological form can be adequately translated into descriptive statements about the life-preserving function of certain biological characteristics, it is clear that (1) the use of the concept of purpose is not essential in these contexts, since the term "purpose" can be completely eliminated from the statements in question, and (2) teleological assumptions, while now endowed with empirical content, cannot serve as explanatory principles in the customary contexts. Thus, e.g., the fact that a given species of butterfly displays a particular kind of coloring cannot be inferred from—and therefore cannot be explained by means of—the statement that this type of coloring has the effect of protecting the butterflies from detection by pursuing birds, nor can the presence of red corpuscles in the human blood be inferred from the statement that those corpuscles have a specific function in assimilating oxygen and that this function is essential for the maintenance of life.

One of the reasons for the perseverance of teleological considerations in biology probably lies in the fruitfulness of the teleological approach as a heuristic device: Biological research which was psychologically motivated by a teleological orientation, by an interest in purposes in nature, has frequently led to important results which can be stated in nonteleological terminology and which increase our knowledge of the causal connections between biological phenomena.

Another aspect that lends appeal to teleological considerations is their anthropomorphic character. A teleological explanation tends to make us feel that we really "understand" the phenomenon in question, because it is accounted for in terms of purposes, with which we are familiar from our own experience of purposive behavior. But it is important to distinguish here understanding

13. An analysis of teleological statements in biology along these lines may be found in Woodger (1929), especially pp. 432 ff; essentially the same interpretation is advocated by Kaufmann in (1944), Chapter 8.

in the psychological sense of a feeling of empathic familiarity from understanding in the theoretical, or cognitive, sense of exhibiting the phenomenon to be explained as a special case of some general regularity. The frequent insistence that explanation means the reduction of something unfamiliar to ideas or experiences already familiar to us is indeed misleading. For while some scientific explanations do have this psychological effect, it is by no means universal: The free fall of a physical body may well be said to be a more familiar phenomenon than the law of gravitation, by means of which it can be explained; and surely the basic ideas of the theory of relativity will appear to many to be far less familiar than the phenomena for which the theory accounts.

"Familiarity" of the explanans is not only not necessary for a sound explanation, as has just been noted; it is not sufficient either. This is shown by the many cases in which a proposed explanans sounds suggestively familiar, but upon closer inspection proves to be a mere metaphor, or to lack testability, or to include no general laws and therefore to lack explanatory power. A case in point is the neovitalistic attempt to explain biological phenomena by reference to an entelechy or vital force. The crucial point here is not—as is sometimes said—that entelechies cannot be seen or otherwise directly observed; for that is true also of gravitational fields, and yet, reference to such fields is essential in the explanation of various physical phenomena. The decisive difference between the two cases is that the physical explanation provides (1) methods of testing, albeit indirectly, assertions about gravitational fields, and (2) general laws concerning the strength of gravitational fields, and the behavior of objects moving in them. Explanations by entelechies satisfy the analogue of neither of these two conditions. Failure to satisfy the first condition represents a violation of (R3); it renders all statements about entelechies inaccessible to empirical test and thus devoid of empirical meaning. Failure to comply with the second condition involves a violation of (R2). It deprives the concept of entelechy of all explanatory import; for explanatory power never resides in a concept, but always in the general laws in which it functions. Therefore, notwithstanding the feeling of familiarity it may evoke, the neovitalistic account cannot provide theoretical understanding.

The preceding observations about familiarity and understanding can be applied, in a similar manner, to the view held by some scholars that the explanation, or the understanding, of human actions requires an empathic understanding of the personalities of the agents[14]. This understanding of another person in terms of one's own psychological functioning may prove a useful heuristic device in the search for general psychological principles which might provide

14. For a more detailed discussion of this view on the basis of the general principles outlined above, cf. Zilsel (1941), sections 7 and 8, and Hempel (1942), section 6.

a theoretical explanation; but the existence of empathy on the part of the scientist is neither a necessary nor a sufficient condition for the explanation, or the scientific understanding, of any human action. It is not necessary, for the behavior of psychotics or of people belonging to a culture very different from that of the scientist may sometimes be explainable and predictable in terms of general principles even though the scientist who establishes or applies those principles may not be able to understand his subjects empathically. And empathy is not sufficient to guarantee a sound explanation, for a strong feeling of empathy may exist even in cases where we completely misjudge a given personality. Moroever, as Zilsel has pointed out, empathy leads with ease to incompatible results; thus, when the population of a town has long been subjected to heavy bombing attacks, we can understand, in the empathic sense, that its morale should have broken down completely, but we can understand with the same ease also that it should have developed a defiant spirit of resistance. Arguments of this kind often appear quite convincing; but they are of an *ex post facto* character and lack cognitive significance unless they are supplemented by testable explanatory principles in the form of laws or theories.

Familiarity of the explanans, therefore, no matter whether it is achieved through the use of teleological terminology, through neovitalistic metaphors, or through other means, is no indication of the cognitive import and the predictive force of a proposed explanation. Besides, the extent to which an idea will be considered as familiar varies from person to person and from time to time, and a psychological factor of this kind certainly cannot serve as a standard in assessing the worth of a proposed explanation. The decisive requirement for every sound explanation remains that it subsume the explanandum under general laws.

PART II. ON THE IDEA OF EMERGENCE

5. Levels of Explanation. Analysis of Emergence. As has been shown above, a phenomenon may be explainable by sets of laws of different degrees of generality. The changing positions of a planet, for example, may be explained by subsumption under Kepler's laws, or by derivation from the far more comprehensive general law of gravitation in combination with the laws of motion, or finally by deduction from the general theory of relativity, which explains—and slightly modifies—the preceding set of laws. Similarly, the expansion of a gas with rising temperature at constant pressure may be explained by means of the Gas Law or by the more comprehensive kinetic theory of heat. The latter explains the Gas Law, and thus indirectly the phenomenon just mentioned, by means of (1) certain assumptions concerning the micro-

behavior of gases (more specifically, the distributions of locations and speeds of the gas molecules) and (2) certain macro-micro principles, which connect such macro-characteristics of a gas as its temperature, pressure and volume with the micro-characteristics just mentioned.

In the sense of these illustrations, a distinction is frequently made between various *levels of explanation*.[15] Subsumption of a phenomenon under general laws directly connecting observable characteristics represents the first level; higher levels require the use of more or less abstract theoretical constructs which function in the context of some comprehensive theory. As the preceding illustrations show, the concept of higher-level explanation covers procedures of rather different character; one of the most important among them consists in explaining a class of phenomena by means of a theory concerning their micro-structure. The kinetic theory of heat, the atomic theory of matter, the electromagnetic as well as the quantum theory of light, and the gene theory of heredity are examples of this method. It is often felt that only the discovery of a micro-theory affords real scientific understanding of any type of phenomenon, because only it gives us insight into the inner mechanism of the phenomenon, so to speak. Consequently, classes of events for which no micro-theory was available have frequently been viewed as not actually understood; and concern with the theoretical status of phenomena which are unexplained in this sense may be considered as one of the roots of the doctrine of emergence.

Generally speaking, the concept of *emergence* has been used to characterize certain phenomena as "novel," and this not merely in the psychological sense of being unexpected,[16] but in the theoretical sense of being unexplainable, or unpredictable, on the basis of information concerning the spatial parts or other constituents of the systems in which the phenomena occur, and which in this context are often referred to as "wholes." Thus, e.g., such characteristics of water as its transparence and liquidity at room temperature and atmospheric pressure, or its ability to quench thirst have been considered as emergent on the ground that they could not possibly have been predicted from a knowledge of the properties of its chemical constituents, hydrogen and oxygen. The weight of the compound, on the contrary, has been said not to be emergent because it is a mere "resultant" of its components and could have been predicted by simple addition even before the compound had been formed. The conceptions of explanation and prediction which underly this idea of emergence call for various critical observations, and for corresponding changes in the concept of emergence.

15. For a lucid brief exposition of this idea, see Feigl (1945), pp. 284–88.

16. Concerning the concept of novelty in its logical and psychological meanings, see also Stace (1939).

PHIL 101: The following section (with red lines in the margin) is optional reading

(1) First, the question whether a given characteristic of a "whole," w, is emergent or not cannot be significantly raised until it has been stated what is to be understood by the parts or constituents of w. The volume of a brick wall, for example, may be inferable by addition from the volumes of its parts if the latter are understood to be the component bricks, but it is not so inferable from the volumes of the molecular components of the wall. Before we can significantly ask whether a characteristic W of an object w is emergent, we shall therefore have to state the intended meaning of the term "part of." This can be done by defining a specific relation Pt and stipulating that those and only those objects which stand in Pt to w count as parts of constituents of w. 'Pt' might be defined as meaning "constituent brick of" (with respect to buildings), or "molecule contained in" (for any physical object), or "chemical element contained in" (with respect to chemical compounds, or with respect to any material object), or "cell of" (with respect to organisms), etc. The term "whole" will be used here without any of its various connotations, merely as referring to any object w to which others stand in the specified relation Pt. In order to emphasize the dependence of the concept of part upon the definition of the relation Pt in each case, we shall sometimes speak of Pt-parts, to refer to parts as determined by the particular relation Pt under consideration.

(2) We turn to a second point of criticism. If a characteristic of a whole is counted as emergent simply if its occurrence cannot be inferred from a knowledge of all the properties of its parts, then, as Grelling has pointed out, no whole can have any emergent characteristics. Thus, to illustrate by reference to our earlier example, the properties of hydrogen include that of forming, if suitably combined with oxygen, a compound which is liquid, transparent, etc. Hence the liquidity, transparence, etc. of water *can* be inferred from certain properties of its chemical constituents. If the concept of emergence is not to be vacuous, therefore, it will be necessary to specify in every case a class G of attributes and to call a characteristic W of an object w emergent relatively to G and Pt if the occurrence of W in w cannot be inferred from a complete characterization of all the Pt-parts with respect to the attributes contained in G, i.e. from a statement which indicates, for every attribute in G, to which of the parts of w it applies. Evidently, the occurrence of a characteristic may be emergent with respect to one class of attributes and not emergent with respect to another. The classes of attributes which the emergentists have in mind, and which are usually not explicitly indicated, will have to be construed as nontrivial, i.e. as not logically entailing the property of each constituent of forming, together with the other constituents, a whole with the characteristics under investigation. Some fairly simple cases of emergence in the sense so far specified arise when the class G is restricted to certain simple properties of the parts, to the exclusion of spatial

or other relations among them. Thus, the electromotive force of a system of several electric batteries cannot be inferred from the electromotive forces of its constituents alone without a description, in terms of relational concepts, of the way in which the batteries are connected with each other.[17]

(3) Finally, the predictability of a given characteristic of an object on the basis of specified information concerning its parts will obviously depend on what general laws or theories are available.[18] Thus, the flow of an electric current in a wire connecting a piece of copper and a piece of zinc which are partly immersed in sulfuric acid is unexplainable, on the basis of information concerning any nontrivial set of attributes of copper, zinc and sulphuric acid, and the particular structure of the system under consideration, unless the theory available contains certain general laws concerning the functioning of batteries, or even more comprehensive principles of physical chemistry. If the theory includes such laws, on the other hand, then the occurrence of the current is predictable. Another illustration, which at the same time provides a good example for the point made under (2) above, is afforded by the optical activity of certain substances. The optical activity of sarco-lactic acid, for example, i.e. the fact that in solution it rotates the plane of polarization of plane-polarized light, cannot be predicted on the basis of the chemical characteristics of its constituent elements; rather, certain facts about the relations of the atoms constituting a molecule of sarco-lactic acid have to be known. The essential point is that the molecule in question contains an asymmetric carbon atom, i.e. one that holds four different atoms or groups, and if this piece of relational information is provided, the optical activity of the solution can be predicted provided that furthermore the theory available for the purpose embodies

17. This observation connects the present discussion with a basic issue in Gestalt theory. Thus, e.g., the insistence that "a whole is more than the sum of its parts" may be construed as referring to characteristics of wholes whose prediction requires knowledge of certain structural relations among the parts. For a further examination of this point, see Grelling and Oppenheim (1937-38) and (1939).

18. Logical analyses of emergence which make reference to the theories available have been propounded by Grelling and recently by Henle (1942). In effect, Henle's definition characterizes a phenomenon as emergent if it cannot be predicted, by means of the theories accepted at the time, on the basis of the data available before its occurrence. In this interpretation of emergence, no reference is made to characteristics of parts or constituents. Henle's concept of predictability differs from the one implicit in our discussion (and made explicit in Part III of this article) in that it implies derivability from the "simplest" hypothesis which can be formed on the basis of the data and theories available at the time. A number of suggestive observations on the idea of emergence and on Henle's analysis of it are presented in Bergmann's article (1944). The idea that the concept of emergence, at least in some of its applications, is meant to refer to unpredictability by means of "simple" laws was advanced also by Grelling in the correspondence mentioned in note (1). Reliance on the notion of simplicity of hypotheses, however, involves considerable difficulties; in fact, no satisfactory definition of that concept is available at present.

the law that the presence of one asymmetric carbon atom in a molecule implies optical activity of the solution; if the theory does not include this micro-macro law, then the phenomenon is emergent with respect to that theory.

An argument is sometimes advanced to the effect that phenomena such as the flow of the current, or the optical activity, in our last examples, are absolutely emergent at least in the sense that they could not possibly have been predicted before they had been observed for the first time; in other words, that the laws requisite for their prediction could not have been arrived at on the basis of information available before their first occurrence.[19] This view is untenable, however. On the strength of data available at a given time, science often establishes generalizations by means of which it can forecast the occurrence of events the like of which have never before been encountered. Thus, generalizations based upon periodicities exhibited by the characteristics of chemical elements then known enabled Mendeleev in 1871 to predict the existence of a certain new element and to state correctly various properties of that element as well as of several of its compounds; the element in question, germanium, was not discovered until 1886. A more recent illustration of the same point is provided by the development of the atomic bomb and the prediction, based on theoretical principles established prior to the event, of its explosion under specified conditions, and of its devastating release of energy.

As Grelling has stressed, the observation that the predictability of the occurrence of any characteristic depends upon the theoretical knowledge available, applies even to those cases in which, in the language of some emergentists, the characteristic of the whole is a mere resultant of the corresponding characteristics of the parts and can be obtained from the latter by addition. Thus, even the weight of a water molecule cannot be derived from the weights of its atomic constituents without the aid of a law which expresses the former as some specific mathematical function of the latter. That this function should be the sum is by no means self-evident; it is an empirical generalization, and at that not a strictly correct one, as relativistic physics has shown.

19. C. D. Broad, who in chapter 2 of his book (1925) gives a clear account and critical discussion of the essentials of emergentism, emphasizes the importance of "laws" of composition in predicting the characteristics of a whole on the basis of those of its parts (*op. cit.*, pp. 61ff.); but he subscribes to the view characterized above and illustrates it specifically by the assertion that "if we want to know the chemical (and many of the physical) properties of a chemical compound, such as silver-chloride, it is absolutely necessary to study samples *of that particular compound*. . . . The essential point is that it would also be useless to study chemical compounds in general and to compare their properties with those of their elements in the hope of discovering a *general* law of composition by which the properties of *any* chemical compound could be foretold when the properties of its separate elements were known." (p. 64) That an achievement of precisely this sort has been possible on the basis of the periodic system of the elements is noted above.

Failure to realize that the question of the predictability of a phenomenon cannot be significantly raised unless the theories available for the prediction have been specified has encouraged the misconception that certain phenomena have a mysterious quality of absolute unexplainability, and that their emergent status has to be accepted with "natural piety," as C. L. Morgan put it. The observations presented in the preceding discussion strip the idea of emergence of these unfounded connotations: emergence of a characteristic is not an ontological trait inherent in some phenomena; rather it is indicative of the scope of our knowledge at a given time; thus it has no absolute, but a relative character; and what is emergent with respect to the theories available today may lose its emergent status tomorrow.

The preceding considerations suggest the following *redefinition of emergence:* The occurrence of a characteristic W in an object w is emergent relative to a theory T, a part relation Pt, and a class G of attributes if that occurrence cannot be deduced by means of T from a characterization of the Pt-parts of w with respect to all the attributes in G.

This formulation explicates the meaning of emergence with respect to *events* of a certain kind, namely the occurrence of some characteristic W in an object w. Frequently, emergence is attributed to *characteristics* rather than to events; this use of the concept of emergence may be interpreted as follows: A characteristic W is emergent relatively to T, Pt, and G if its occurrence in *any* object is emergent in the sense just indicated.

As far as its cognitive content is concerned, the emergentist assertion that the phenomena of life are emergent may now be construed, roughly, as an elliptic formulation of the following statement: Certain specifiable biological phenomena cannot be explained, by means of contemporary physico-chemical theories, on the basis of data concerning the physical and chemical characteristics of the atomic and molecular constituents of organisms. Similarly, the thesis of an emergent status of mind might be taken to assert that present-day physical, chemical, and biological theories do not suffice to explain all psychological phenomena on the basis of data concerning the physical, chemical, and biological characteristics of the cells or of the molecules or atoms constituting the organisms in question. But in this interpretation, the emergent character of biological and psychological phenomena becomes trivial; for the description of various biological phenomena requires terms which are not contained in the vocabulary of present-day physics and chemistry; hence we cannot expect that all specifically biological phenomena are explainable, i.e. deductively inferable, by means of present-day physico-chemical theories on the basis of initial conditions which themselves are described in exclusively physico-chemical terms. In order to obtain a less trivial interpretation of the assertion that the

PHIL 101: resume assigned reading

phenomena of life are emergent, we have therefore to include in the explanatory theory of those presumptive laws presently accepted which connect the physico-chemical with the biological "level", i.e., which contain, on the one hand, certain physical and chemical terms, including those required for the description of molecular structures, and on the other hand, certain concepts of biology. An analogous observation applies to the case of psychology. If the assertion that life and mind have an emergent status is interpreted in this sense, then its import can be summarized approximately by the statement that no explanation, in terms of micro-structure theories, is available at present for large classes of phenomena studied in biology and psychology.[20]

Assertions of this type, then, appear to represent the rational core of the doctrine of emergence. In its revised form, the idea of emergence no longer carries with it the connotation of absolute unpredictability—a notion which is objectionable not only because it involves and perpetuates certain logical misunderstandings, but also because, not unlike the ideas of neovitalism, it encourages an attitude of resignation which is stifling for scientific research. No doubt it is this characteristic, together with its theoretical sterility, which accounts for the rejection, by the majority of contemporary scientists, of the classical absolutistic doctrine of emergence.[21]

PART III. LOGICAL ANALYSIS OF LAW AND EXPLANATION

6. Problems of the Concept of General Law. From our general survey of the characteristics of scientific explanation, we now turn to a closer examination of its logical structure. The explanation of a phenomenon, we noted, consists in its subsumption under laws or under a theory. But what is a law, what is a theory? While the meaning of these concepts seems intuitively clear, an attempt to construct adequate explicit definitions for them encounters considerable difficulties. In the present section, some basic problems of the concept of law will be described and analyzed; in the next section, we intend to propose, on the basis of the suggestions thus obtained, definitions of law and of explanation for a formalized model language of a simple logical structue.

20. The following passage from Tolman (1932) may serve to support this interpretation: ". . . 'behavior-acts,' though no doubt in complete one-to-one correspondence with the underlying molecular facts of physics and physiology, have, as 'molar' wholes, certain emergent properties of their own. . . . Further, these molar properties of behavior-acts cannot in the present state of our knowledge, i.e., prior to the working-out of many empirical correlations between behavior and its physiological correlates, be known even inferentially from a mere knowledge of the underlying, molecular, facts of physics and physiology" (*op. cit.*, pp. 7-8). In a similar manner, Hull uses the distinction between molar and molecular theories and points out that theories of the latter type are not at present available in psychology. Cf. (1943a), pp. 19ff.; (1943), p. 275.

21. This attitude of the scientist is voiced, for example, by Hull in (1943a), pp. 24-28.

The concept of law will be construed here so as to apply to true statements only. The apparently plausible alternative procedure of requiring high confirmation rather than truth of a law seems to be inadequate: It would lead to a relativized concept of law, which would be expressed by the phrase "sentence S is a law relative to the evidence E." This does not accord with the meaning customarily assigned to the concept of law in science and in methodological inquiry. Thus, for example, we would not say that Bode's general formula for the distance of the planets from the sun was a law relative to the astronomical evidence available in the 1770s, when Bode propounded it, and that it ceased to be a law after the discovery of Neptune and the determination of its distance from the sun; rather, we would say that the limited original evidence had given a high probability to the assumption that the formula was a law, whereas more recent additional information reduced that probability so much as to make it it practically certain that Bode's formula is not generally true, and hence not a law.[22]

Apart from being true, a law will have to satisfy a number of additional conditions. These can be studied independently of the factual requirement of truth, for they refer, as it were, to all logically possible laws, no matter whether factually true or false. Adopting a term proposed by Goodman[23], we will say that a sentence is *lawlike* if it has all the characteristics of a general law, with the possible exception of truth. Hence, every law is a lawlike sentence, but not conversely.

Our problem of analyzing the notion of law thus reduces to that of explicating the concept of lawlike sentence. We shall construe the class of lawlike sentences as including analytic general statements, such as 'A rose is a rose', as well as the lawlike sentences of empirical science, which have empirical content.[24] It will not be necessary to require that each lawlike sentence permissible in explanatory contexts be of the second kind; rather, our definition of explanation will be so constructed as to guarantee the factual character of the totality of the laws—though not of every single one of them—which function in an explanation of an empirical fact.

22. The requirement of truth for laws has the consequence that a given empirical statement S can never be definitely known to be a law; for the sentence affirming the truth of S is tantamount to S and is therefore capable only of acquiring a more or less high probability, or degree of confirmation, relative to the experimental evidence available at any given time. On this point, cf. Carnap (1946). For an excellent nontechnical exposition of the semantical concept of truth, which is here invoked, the reader is referred to Tarski (1944).

23. (1947), p. 125.

24. This procedure was suggested by Goodman's approach in (1947). Reichenbach, in a detailed examination of the concept of law, similarly construes his concept of nomological statement as including both analytic and synthetic sentences: cf. (1947). Chapter VIII.