TIGHT BINDING BOOK

UNIVERSAL
LIBRARY
OU_160215
UNIVERSAL
LIBRARY

OSMANIA UNIVERSITY LIBRARY

Call No. 530·1 D11R Accession No. 35936

Author D'Abro, R.

Title Rise of the new physics.

This book should be returned on or before the date 1957. last marked below.

the rise of the new physics

the rise of the new physics

volume I

by A. d'Abro

ITS MATHEMATICAL AND PHYSICAL THEORIES

FORMERLY TITLED "DECLINE OF MECHANISM"

IN TWO VOLUMES

dover publications, inc.

*

THE RISE OF THE NEW PHYSICS,
ITS MATHEMATICAL & PHYSICAL THEORIES
was originally published under the title
DECLINE OF MECHANISM.
This second edition contains corrections
by the author and new sections of portraits.
Copyright 1939 by
D. Van Nostrand Co., Inc.
Copyright 1951 by
Dover Publications, Inc.
Manufactured in the U. S. A.

PREFACE

The present century has witnessed the emergence of two grand theories of mathematical physics: the Theory of Relativity and the Quantum Theory. Both theories were conceived for the purpose of coordinating certain bodies of facts which the classical theories were unable to interpret; and neither theory would have seen the day had it not been for the increased refinement of experimental measurements which rendered the disclosure of these facts possible. But although the two theories were born under similar circumstances, they soon branched out in opposite directions. The theory of relativity has developed into a doctrine whose principal field of application is found in the world of large-scale phenomena, whereas the quantum theory has become identified with the microscopic processes of the atomic and subatomic worlds. To this extent the two theories are complementary. In this book we shall be concerned more especially with the quantum theory and with the researches undertaken by physicists in the subatomic world.

Probably the most remarkable discovery that has issued from these researches is that subatomic phenomena do not appear to be mere repetitions on a microscopic scale of the macroscopic processes with which we have become familiar from our daily experience. In the subatomic world mechanical representations and classical concepts are no longer of much avail, except as props to a bewildered imagination which is unable to feel at ease in its new surroundings. But when classical concepts are utilized in the interpretation of subatomic phenomena, we find ourselves confronted with unexpected difficulties: waves and particles seem to dissolve one into the other as though they were the same and yet not the same.

If the subatomic world were neatly separated from the better-known macroscopic one, we might view the new discoveries as peculiar to the subatomic world. There would then be two physics: the classical (or relativistic) physics of the macroscopic world and the quantum physics of the microscopic world. However, the two worlds exhibit no clear-cut separation; and so we must suppose that the strange characteristics of the new world are present also in the world of our common experience. If this be the case classical physics, even in the macroscopic world, must be superseded by quantum physics. No inconsistency with the facts of observation is involved in this conclusion, for the quantum theory shows

that in our ordinary world the strange quantum characteristics should become obscured to such an extent that observations of far greater refinement than we are able to perform in practice would be required to detect them.

The philosophical implications of the new discoveries are still controversial, the main disagreements centering around the status of the principle of rigorous causality. Classical scientists assumed the validity of this principle in connection with the evolution of inorganic systems. Doubts on its validity arose in the earlier years of this century, when the difficulty of accounting for the radiation law led Planck to propose the initial form of the quantum theory. But at that time it was thought that causal connections would be reestablished subsequently as a better understanding of quantum phenomena was obtained, and so no special prominence was given to discussions on the merits of the causal doctrine. Then, in 1927, Heisenberg discovered his celebrated "Uncertainty Relations," as a result of which he contended that a rigorously deterministic scheme could no longer be entertained in physics. Heisenberg's contention occasioned a split between the leading natural philosophers. The quantum theorists, represented by Born, Bohr, and Dirac, supported Heisenberg; Bohr, in particular, suggested his "Principle of Complementarity" as a substitute for the discarded causal doctrine. But the attitude of the quantum theorists was resisted by Planck and by Einstein, who expressed the opinion that the apparent breakdown of the causal doctrine was due to the incompleteness of our present knowledge and that it did not betray any fundamental indeterminism in natural processes.

In the final analysis the quantum theory, like the theory of relativity, is a mere refinement of the classical theories, and to this extent an understanding of the quantum theory requires that we take into account the historical background. Our aim has been to give a semi-popular presentation of the progressive growth of the new ideas, starting from the most elementary notions and giving due consideration to the mathematical development without which the new theories could never have been constructed.

We have divided the book into three parts. Part I deals with the fundamentals of natural philosophy (in relation to physics). Part II contains chapters on mathematics and a review of the more important physical theories of the classical period. In Part III we have discussed the various quantum theories, and have included whatever preliminary mathematical information seemed necessary to their understanding. In the last chapter we have attempted to examine the present status of the doctrine of causality in the light of the new discoveries.

Some of the subject matter might have been omitted: for example, Galois' theory of groups (Chapter XIV), the controversies on the nature of mathematics (Chapter XVI), the Problem of Three Bodies (Chapter XVII), and the theory of contact transformations (Chapter XX). Also Bohr's theory of the atom (Chapters XXVI and XXVII) might have been shortened. But we have preferred to retain these chapters for various reasons. Our excuse for devoting a few pages to Galois' theory of groups and to contact transformations is that these theories are of sufficient interest in themselves to merit mention, and we believe that no elementary presentation of them has yet been given. We have included a brief discussion of the Problem of Three Bodies because the tremendous mathematical difficulties which beset this seemingly simple mechanical problem show how the progress of theoretical physics may be delayed by purely mathematical obstacles. For a somewhat different reason we have given a summary of the controversies on the nature of mathematics. In our opinion these controversies originate from the same psychological differences which appear to be responsible for the current controversy concerning the principle of causality in physics. As such, the controversies in mathematics help to shed light on the nature of the dispute between the quantum theorists and their opponents. Finally we have treated Bohr's theory of the atom at some length because this theory illustrates the difficulty of providing mechanical interpretations for subatomic phenomena. Indeed it was the insufficiency of Bohr's theory that was primarily responsible for the development of the newer quantum theories, and the justification for these newer theories can be appreciated only when the limitations of Bohr's theory are understood.

The non-mathematical reader should experience no serious difficulty in following the presentation, especially so since the more technical chapters (Chapters XXXI to XL) are not essential to a general understanding of the new ideas, and may be omitted in consequence.

New York, N. Y.,
January 1939

CONTENTS

PAGE

PREFACE v

PART I. GENERAL CONSIDERATIONS

I. THE HISTORICAL BACKGROUND OF THE SCIENTIFIC METHOD 3
II. ASSUMPTIONS IN SCIENCE 14
III. THE SIGNIFICANCE OF THEORETICAL PHYSICS 28
IV. TWO KINDS OF SIMPLICITY 32
V. SUCCESSIVE THEORIES OF MATHEMATICAL PHYSICS ARE PROGRESSIVE APPROXIMATIONS 36
VI. CONCEPTUAL CHANGES IN PHYSICAL THEORIES 42
VII. CAUSALITY 45
VIII. CONSERVATION 58
IX. MECHANISTIC THEORIES 61
X. FIELD THEORIES 71
XI. PHENOMENOLOGICAL THEORIES 88
XII. A RETROSPECT ON THE DIFFERENT KINDS OF THEORIES 101
XIII. PSYCHOLOGICAL DIFFERENCES AMONG PHYSICISTS 106

PART II. PHYSICAL THEORIES OF THE CLASSICAL PERIOD

XIV. MATHEMATICAL PRELIMINARIES (Functions and Groups) 117
XV. MATHEMATICAL PRELIMINARIES (Continued) (Differential Equations) 166
XVI. THE CONTROVERSIES ON THE NATURE OF MATHEMATICS 186
XVII. ANALYTICAL MECHANICS (Hamilton's Equations, Three-Body Problem, Figures of Equilibrium for Fluids in Rotation) 214
XVIII. MINIMAL PRINCIPLES AND PRINCIPLES OF ACTION 256
XIX. THE UNDULATORY THEORY OF LIGHT 269
XX. TRANSFORMATIONS, HAMILTON'S WAVE MECHANICS, CONTINUOUS GROUPS 309
XXI. THERMODYNAMICS 331
XXII. THE CLASSICAL KINETIC THEORY OF GASES 374

PORTRAITS

MATHEMATICIANS

following page 268

EUCLID, ARCHIMEDES, GALILEO, SIR ISAAC NEWTON, LEONARD EULER, JOSEPH LOUIS LAGRANGE, PIERRE SIMON LAPLACE, AUGUSTIN CAUCHY, WILLIAM ROWAN HAMILTON, CARL G. J. JACOBI, EVARISTE GALOIS, CHARLES HERMITE, MARIUS SOPHUS LIE, HENRI POINCARÉ, DAVID HILBERT

PRE-QUANTUM PHYSICISTS

following page 426

AUGUSTIN FRESNEL, LORD RAYLEIGH, VIKTOR MEYER, SADI CARNOT, RUDOLF JULIUS EMANUEL CLAUSIUS, JOHN WILLARD GIBBS, WALTER HERMANN NERNST, JAMES CLERK MAXWELL, LUDWIG BOLTZMANN

PART I

GENERAL CONSIDERATIONS

CHAPTER I

THE HISTORICAL BACKGROUND OF THE SCIENTIFIC METHOD

THE craving to understand appears to be one of the characteristic incentives of the human species. With primitive races the form of understanding sought is of a kind that is useful to man in his daily struggle for existence. As such, the motive underlying the urge is a practical one. In later periods sheer curiosity becomes the main driving force, and we find men seeking knowledge for its own sake. But though the urge to acquire knowledge is as old as the human race, the method whereby the urge may be satisfied is a comparatively recent discovery. This method, devised by Galileo and Newton, is known as the scientific method, and the knowledge obtained from its application is called science. Natural philosophy embodies the general outlook that seems best suited to account for the various discoveries. In the case of physics, with which we shall be especially concerned, natural philosophy finds its chief mode of expression in the theories of mathematical physics. To the abstruseness of these theories is due the esoteric character of natural philosophy and the difficulty of weighing its conclusions, unless the critic be well versed in general physics and in mathematics.

Prior to the discovery of the scientific method little progress was made; and though art, literature, and geometry * flourished, man had practically no understanding of natural phenomena. In considering the scientific method, it will be advantageous to stress the successive stages in its application. These may be called:

(a) The observational stage.
(b) The experimental stage.
(c) The theoretical and mathematical stage (in physics).

The order in which these stages have been listed is the order in which they arise in the study of any group of physical phenomena. It is also the chronological order in which they were discovered.

* We are not viewing mathematics as a science but as a form of human reasoning bearing a strong resemblance to logic. Furthermore, throughout this book the stress will be laid on physical science.

(a) *The observational stage.* This first stage consists in observing phenomena with precision. For instance, we may observe that plants die unless supplied with moisture, or that rainbows are associated with storms, or that the stars move around the Pole Star. Observations of this sort, being purely qualitative, are necessarily lacking in definiteness; but we may often attain greater precision by discerning quantitative relationships. Thus the last observation is improved when we note that, in their motions around the Pole Star, the stars describe circles with uniform speed, as though they were rigidly connected. We need scarcely mention that the more precise the information, the better equipped we are to progress; and it is thus apparent that the search for quantitative relationships arose quite naturally from a desire to attain greater accuracy in the observations. Other reasons also favor quantitative observations, reasons which may not have been obvious to the earlier investigators but which become clear when the third, or mathematical, stage comes into consideration.

Quantitative observations and measurements often reveal remarkably simple relationships among the various magnitudes measured. Such relationships are called natural laws and are referred to specifically as "empirical laws." Kepler's laws for the planetary motions are probably the most famous laws of this type. Other kinds of natural laws will be examined as we proceed.

Properly speaking, we should restrict the appelation "scientific method" to quantitative observations, because the mere noting that rainbows are associated with showers or that plants die unless supplied with moisture constitutes an observation of so commonplace a nature that it can scarcely be called scientific. Even with this reservation, the application of the first step in the scientific method must have arisen many centuries ago, for we know that the Chaldaeans and ancient Egyptians had acquired a considerable knowledge of observational astronomy. The Greek astronomers, Hipparchus in particular, established quantitative relationships with a degree of precision that has elicited the admiration of later investigators. To mention but one instance, we may recall Hipparchus's discovery of the precession of the equinoxes.

(b) *The experimental stage.* The second step in the application of the scientific method consists in supplementing the first, or purely observational, step by experiment. Men no longer content themselves with observing phenomena as and when they happen to occur; they go out of their way to produce these phenomena artificially, so as to observe them under different conditions and with the increased accuracy permitted by their repetition. Aside from this important difference, the

same requirements of accurate observation of a quantitative nature (whenever possible) are imposed. The discovery of the experimental method is usually attributed to Galileo, and, for this reason among others, he is regarded as the father of modern science. Some injustice seems to be done to Archimedes, since in all truth it was he who inaugurated the experimental method in his investigations in hydrostatics. Unfortunately for the glory of Greece, Archimedes came too late to found a school. Rome was then rising in power, and the interest of the new civilization lay in conquest and in organization rather than in scientific research. The great lesson taught by Archimedes was soon forgotten, and when in the middle ages the study of Hellenic culture was revived under the scholastics, it was Aristotle and the mystic Plato who were regarded as the most eminent representatives of Greek thought. Galileo lived in a more propitious epoch, and he had the remarkably good fortune of being followed by Newton. Newton, by his discovery of the third (or theoretical) step, to which reference will be made presently, completed the philosophy of the scientific method inaugurated by Galileo. Natural philosophy as pertaining to physics was thus born and has endured without any essential change in its methods to the present day.

Aside from having rendered possible the introduction of the third (or theoretical) step, the second (or experimental) stage has an enormous importance on its own account. The majority of empirical laws in science were obtained by the application of the experimental method. Galileo measured the accelerations of falling bodies and thereby obtained the empirical law of freely falling bodies. Boyle's law for gases and Descartes' law of refraction are other examples of empirical laws derived from experiment. In another department of science we may mention Mendel's observations on the cross-breeding of different varieties of sweet peas; the empirical law obtained is known as Mendel's law.

The rôle played by experiment is less conspicuous in astronomy and in geology, for in these sciences we must be content with observation. In modern astronomy and in astrophysics, however, experiment plays an important, though indirect, rôle. Firstly, it is thanks to physical experiment (supplemented by theory) that telescopes and spectroscopes can be constructed. Secondly, it is experiment that enables us to match in the laboratory the spectral lines observed in the spectra of the stars and thereby to establish their significance. In a broad sense, the successful performance of the first and second steps requires care, precision, and ingenuity in the construction of the experimental apparatus Very little in the nature of speculative assumptions is needed.

(c) *The theoretical stage.* The third step was taken by Newton when he furnished a mathematical theory of the planetary motions, obtaining thereby his law of gravitation. It differs from the experimental step in that it is mathematical and speculative. Furthermore, physics is the only one of the sciences to which the mathematical procedure of the third stage has been applied with outstanding success. Owing to this circumstance physics has developed far more rapidly than the other sciences.

The incentives which prompted the third step are readily accounted for. In the experimental stage facts were discovered, their relations were established, and empirical laws were obtained. But the relations disclosed by experiment were the most obvious ones, and it seemed permissible to suppose that hidden relations also existed. Could these additional relations be revealed, a wider array of facts would be connected than was possible by the experimental method alone, and our understanding of natural processes would be correspondingly increased. To proceed beyond the experimental stage, investigators were compelled, however, to introduce assumptions of a more or less speculative nature. In the majority of cases the implications of these assumptions were too complicated to be deduced by commonplace reasoning and logic. A far subtler method of approach was needed; it was found in the mathematical instrument. By this means the third stage in the scientific method (in physics) was attained.

Thanks to the mathematical instrument, formerly unsuspected relations were revealed, and additional laws were derived. These mathematical laws are no longer called "empirical," for they differ entirely in their mode of derivation from the laws that may have been established in the preceding steps. Newton's law of gravitation is an illustration of a law obtained through the medium of mathematics. The doctrines developed in physics as a result of the application of the third step are called "theories of mathematical physics," or more briefly "theoretical physics." The theory of relativity and the quantum theory are of this type.

In any historical discussion of the development of the scientific method in physics, we must realize that the theoretical stage could not have arisen before mathematics had attained at least a moderate degree of development. The fact is that all physical theories involve the differential calculus, the calculus of probabilities, or still more advanced parts of analysis. Since none of these mathematical developments were known before Newton's day, it would be unfair to criticize Newton's predecessors for their failure to apply the theoretical method. On the other hand, we may well wonder why the experimental step should have been so tardy in making its appearance, especially when we realize that the transition

from the observational to the experimental step is a natural one. Both steps rely essentially on observation; they differ only in that, in the first step, we observe what happens to occur, whereas in the experimental step we attempt to generate the occurrence. The child who observes the contents of an open box is applying the first step. And when the box is closed and the child deliberately opens it to view its contents, he is unwittingly applying the second step. The experimental method appears so natural today that there has always been something of a paradox, in the failure of the Greeks to have applied it, other than in the most trivial cases. Are we then to believe that the gifted Greek race, that gave to the world such varied geniuses as Homer, Phidias, Sophocles, and Euclid, produced only one thinker, Archimedes, who realized the advisability of opening the box rather than of speculating endlessly on its hidden contents? A possible answer to this question will be considered presently. At all events, the failure of the ancients to rely on experiment was responsible for the insignificance of their contributions to natural philosophy. Indeed, if we except Archimedes and the geometricians, we find that the Greeks confined themselves to wild guesses, supported by metaphysical arguments so transparently irrelevant that even a modern schoolboy could refute them with ease. In these speculations, matter, space, motion, atoms with hooks and eyes, and other notions drawn from commonplace experience played a prominent part. The procedure of the ancients has not even the merit of being consistent, for if we choose to ignore the information revealed by intelligent experimentation, we should in all logic disregard likewise the cruder knowledge we have gained from commonplace observation, and we should refrain from taking this knowledge as the basis of our speculations.

A. H. Compton attributes the strange lack of understanding of the Greeks to the influence of Socrates and of the Persian magicians. In our opinion Compton's explanation cannot contain the whole truth, for we may safely presume that not all the Greeks were mystics. And who aside from a mystic could be swayed by the arguments of the magicians or of Socrates and Plato?

Another explanation, which appears to us more plausible, is that the Greeks were perfectly capable of understanding the value of experiment but that they did not wish to recognize it. Plainly, the application of the experimental method is not a matter of mere intelligence; it involves in a less conspicuous form other qualities; namely, character, sincerity, and modesty. It requires character to seek the truth even when we have reason to fear that it will not be to our liking. It requires sincerity to accept the truth when this truth happens to contradict all

that we have previously professed. Finally, it requires modesty to recognize that man cannot, by his inner vision alone, attain to truth and that he must stoop to experiment. Experiment has always been anathema to the egotist, not necessarily because it involves manual labor, but because it belittles man by placing him in the position of a humble student of Nature instead of revealing him as the Lord of all Creation.

These qualities of courage, sincerity, and modesty, which the ancients seem to have lacked, prevented them from creating a science. Thus we find Pythagoras concealing his discovery of irrationals because it upset his doctrine of numbers. In Galileo's day the same spirit is shown by the metaphysician who refused to look at the heavens through a telescope for fear of being convinced of the errors in Aristotle's teachings. And, if Plato by opening a box could have tested his theory of universals, we may be certain that he would have destroyed the box rather than run the risk of being refuted.

Of course, when we speak of the experimental method as exemplified by the opening of a box to see what it contains, we are taking a trivial illustration, and we may find plenty of trivial instances in which even the Greek metaphysicians resorted to experiment. But they proceeded in a half-hearted, inconsistent way. We shall not dwell on the speculations of the Pythagoraeans and the early Atomists, for at the dawn of any civilization guesses are all that can be expected. Plato may also be disregarded since he was avowedly a mystic seeking truth from his inner soul. A study of Aristotle's writings will be more instructive from our present standpoint.

Aristotle achieved well-deserved fame as the founder of logic; but he was also a student of natural phenomena, and it is with his investigations in this latter field of inquiry that we shall here be concerned. The annual overflow of the Nile excited his curiosity; and to determine its origin he proceeded in a thoroughly commendable way by sending an expedition to the sources of the river. In other words he opened the box and looked inside. Yet the same Aristotle informs us that the increase in weight of a growing plant is due to the matter it absorbs from the soil. Of course Aristotle, like every one else, had observed that plants grow from the ground and do not spring from a bare slab of stone; accordingly, his guess was plausible. But never did he attempt to test it by weighing a pot containing a growing plant at successive periods of its growth. It is significant that, not until the eighteenth century, was this simple test made and the weight of the pot found to increase. The foregoing elementary experiment, which might have convinced Aristotle

of his mistake, would surely have been less costly and more easily accomplished than the sending of an expedition to the sources of the Nile.

Elsewhere Aristotle tells us that women have less teeth than men and that they have fewer ribs than men. We are also told that a body dropped from the mast of a moving ship falls behind the mast. Obviously the experiment was not performed. In all cases we find guesses, some of which, such as the prevalence of earthquakes in the coastal regions, have turned out to be correct. But whether the guesses were correct or incorrect is a minor point; the characteristic fact is that they were not even based on careful observation. As such, they could teach Aristotle nothing of the phenomena in which he was interested. The net result is that Mendel by his simple experiments contributed more to biology than did Aristotle in his massive writings.

A comparative study of the rise of the experimental method in the different civilizations would be of interest in disclosing whether the strange deficiencies of the Greeks were peculiar to them. Unfortunately such a study could be applied only to the civilizations of the white race, for the white race alone has evolved the experimental method on its own initiative. We may note, however, that in the few instances where some kind of a comparison is possible, the application of the experimental method has invariably followed the golden age of literature and art: Archimedes comes after Pericles, Galileo after Dante and Raphael, and Newton after Shakespeare. From these illustrations we might surmise that the failure of the Greeks to develop the experimental method was due to the premature eclipse of their civilization. Other considerations support this view. Homocentric tendencies have always been pernicious to the development of science; and we find that in the early stages of any civilization these tendencies predominate. The history of painting furnishes an illustration. The first subjects to be represented were men, gods fashioned like men, and animals useful to men. Presently landscapes were added, but only as a background; man was still the centre of interest. Only considerably later, with the rise of the Dutch school, were landscapes painted for their own sake. Similarly in literature, the earlier poets sang of their loves and battles but never extolled the beauty of their lands. Even in more modern times Montesquieu in his description of the Alps speaks of them as a mass of rocks which make travelling tedious. Not until the latter part of the eighteenth century, in the writings of J. J. Rousseau, is the beauty of nature considered worthy of more than passing comment. In the scientific field we find the same persistent homocentric tendencies. These tendencies are seen in the heated controversies which followed Copernicus's defense of the heliocentric sys-

tem. The true importance of Copernicus's discovery lies in its having paved the way for Kepler's laws and Newton's investigations. But what caused so much excitement in Copernicus's days was the dethronement of the Earth from its central position in the Universe. Yet who but the egotist cares one whit whether the Earth on which he lives be the centre of Creation or a mere satellite of the sun? In more recent times the antagonism manifested in certain quarters to the theory that man's ancestors were apes is traceable to a similar egotism.* However natural all these homocentric tendencies may be, they have undoubtedly exerted a pernicious influence on the development of disinterested research.

With the application of the experimental method by Galileo, the effects of egotism began to lose their force, at least in scientific matters; facts were sought whether they flattered man or not. The seventeenth century witnessed several attempts to extend to other phenomena the method of Galileo. Thus Descartes measured the angles of incidence and of refraction of a ray of light when it passed from one medium into another, and he obtained from his measurements the empirical law of refraction. Pascal performed experiments in hydrostatics and established the decrease in the atmospheric pressure that accompanies increasing elevation. Boyle studied the change in volume of a gas when compressed at constant temperature. Newton supplemented Galileo's mechanical experiments with others of his own invention and thereby obtained the fundamental laws of mechanics. Experiments were conducted by Newton and by Huyghens in optics. These and a few others constitute the sole contributions of the experimenters of the seventeenth century. The results are meagre, and no more rapid advance is registered in the following century. The slow progress was due to the crudeness of the experimental apparatus and to the limited control the investigators had over the conditions under which their experiments were performed.

The truly significant development that occurred in the seventeenth century was the discovery and application by Newton of the third step in the scientific method. The facts that Newton attempted to coordinate were Kepler's laws of the planetary motions and the mechanical facts expressed by the laws of mechanics. Newton made the assumption that the planetary masses were governed by the same mechanical laws that held for terrestrial bodies. It is the deduction of the necessary mathematical consequences of this simple assumption that constitutes one of

* We are here referring to objections from those who complain that man's dignity is degraded by this theory, and not from those who combat it for religious or social reasons.

Newton's chief titles to fame. As is well known, a direct consequence of Newton's assumption (when consideration is given to Kepler's laws) is the law of gravitation. Several important points are connected with Newton's investigations, but we shall defer fuller consideration of them for the present. Suffice it to say, not only did Newton's theory account for Kepler's laws (and this was its original aim), but it also showed the connection between the fall of bodies and such seemingly independent phenomena as the tides and the precession of the equinoxes. Newton's findings illustrate the ability of a scientific theory to coordinate a greater array of facts than would be possible by means of the experimental method alone.

The importance of Newton's method was so obvious that it could not fail to impress his contemporaries. His success appeared all the more striking when contrasted with Descartes' rival scheme. Descartes was a creative mathematician fully conversant with the limited mathematics of his day; he also realized the importance of ascertaining facts correctly. (To him we owe the empirical laws of reflection and refraction.) Nevertheless, in the later stages of his life, he preferred to pursue the facile rationalizations of the Greek metaphysicians, creating a world to suit his fancy without the slightest regard for the known facts. Thus Descartes sought to account for the planetary motions, and for this purpose he imagined vortices sweeping the planets around the sun. But, in contradistinction to Newton, he failed to take into consideration the precise facts established by Kepler or indeed any of the mechanical facts discovered by Galileo. In Descartes' scheme, the planets might just as well have followed arbitrary curves, open or closed; whereas in the Newtonian scheme, had the courses and the motions of the planets differed from those established by Kepler, Newton's law of the inverse square would have been untenable. The net result of Descartes' faulty procedure was that his theory accounted for nothing, coordinated nothing, and predicted nothing; and his speculations have exerted no greater influence on the subsequent development of natural philosophy in physics than have his similar ones in medicine.

In this condemnation of Descartes, we must make clear that the objectionable feature in his writings is not due to his speculative spirit; indeed, many of the theories of mathematical physics which occupy a high place in the history of science make greater demands on the imagination than anything Descartes ever advanced. What condemns Descartes is his utter disregard for facts. Some excuse for him may be found when we remember that in his time the differential calculus was unknown and the proper treatment of the planetary motions impossible. Less excuse

can be made for Leibnitz. The extravagant writings of Descartes and of Leibnitz did, however, exert some beneficial influence indirectly, by convincing mathematicians of the futility of baseless speculations. Never again do we find creative mathematicians constructing world systems to suit their fancy; instead, they concentrated their efforts on perfecting the mathematical instrument, thereby preparing the way for the theoretical physicist.

A point which cannot be stressed too strongly is that although mathematicians pursue their investigations in a disinterested spirit, their discoveries prove of inestimable value to the theoretical physicist in his attempts at constructing a theory. Our mention of the bearing of the differential calculus on the theory of gravitation is but one illustration among many. Even if all the facts on which the theory of relativity and the quantum theory are based had been known in Newton's time (or even a century later), it would still have been impossible, owing to mathematical limitations, to construct these theories. The general theory of relativity, for instance, could never have arisen prior to Riemann's investigations in pure mathematics on curved spaces.

Having noted the origins of the scientific method, we pass to the principal results of its application in physics. During the seventeenth century Newton effected the synthesis of the facts pertaining to the dynamics of rigid bodies and to the planetary motions; other facts were meagre and incomplete, and there was little left to coordinate. The early part of the eighteenth century did not contribute new facts of importance. Only at its close was the electric current generated by Volta, and the law of electrostatic actions formulated by Coulomb. As against this lack of experimental progress, the eighteenth century witnessed a remarkable mathematical development at the hands of Euler, d'Alembert, Lagrange, Daniel Bernoulli, and Laplace. Most of the advance arose from attempts to further the application of Newtonian dynamics to point-masses and to extended solids. Thus we find Euler obtaining his celebrated equations for solids in rotation; d'Alembert linking statics to Newtonian dynamics by means of his principle; Lagrange condensing Newtonian dynamics into a deductive form in his "Mecanique Analytique"; and Laplace contributing advances of major importance in celestial mechanics. We may also mention Euler's and Lagrange's extension of Newton's dynamics to fluids in motion. Thus statics and hydrostatics, whose laws the genius of Archimedes had established, were supplemented by dynamics and hydrodynamics. Along with this progress in the furthering of Newtonian mechanics, important branches of pure mathematics were developed, such as the calculus of

variations (Lagrange), harmonic analysis (Legendre, Laplace), and partial differential equations (Lagrange).

The eighteenth century, as we have pointed out, did not contribute any physical facts of importance, so that no new physical theories were forthcoming. But during the nineteenth and twentieth centuries a whole mass of experimental data was obtained in every field of physics, thereby providing material for additional physical theories. Nevertheless, only some of the theories of the nineteenth and twentieth centuries would have seen the day, had the mathematical advance failed to keep pace with the experimental discoveries. Fortunately, the nineteenth century was also one of tremendous mathematical development, and thanks to this joint progress, theories of mathematical physics began to appear. The most important of these are thermodynamics (Carnot, Clausius, Maxwell, Gibbs, Nernst); Maxwell's theory of electromagnetism; the kinetic theory of gases (Boltzmann, Maxwell, Gibbs, Einstein); Lorentz's theory of electrons; and, finally, the theory of relativity and the quantum theory.

CHAPTER II

ASSUMPTIONS IN SCIENCE

ACCORDING to Planck, the two fundamental canons of natural philosophy are:

1. There is a real outer world which exists independently of our act of knowing.
2. The real outer world is not directly knowable.*

The first of these statements cannot be proved or disproved either by *a priori* arguments or by experiment; the stand of the solipsist is unassailable. For pragmatic reasons, however, the independent existence of an outer world must be granted.

Planck's second canon states that we have no "direct knowledge" of the real world. But first, what do we mean by direct knowledge? Obviously, knowledge that is unaccompanied by inference must be classed as direct. Thus our immediate sensations of warmth, color, etc., constitute direct knowledge. In a more extended sense direct knowledge applies to our perceptions, though here inference is not entirely lacking. Planck's second canon therefore implies that the real world, which is the cause of our sensations, is not revealed to us by these sensations alone. We might also add that this world cannot be disclosed by mere meditation and introspection. No one conversant with science would dispute Planck's contention; the direct knowledge of the real world claimed by the mystic and the modern neo-realist has no place in a scientific discussion.

In our ordinary life our attention is absorbed by our sensations and perceptions, and the real impersonal world (the reality behind the appearances) is disregarded. But the physicist cannot restrict himself to such subjective statements as "snow is white" or "sugar is sweet"; he is compelled by the nature of his studies to peer beyond and explore the real world. Only thus can he discover the hidden relations that clarify the workings of phenomena. It is the disclosure of these relations that constitutes the aim of physical science. Since direct knowledge is insufficient to reveal the real world, the physicist proceeds in a roundabout

* "Where Is Science Going?" Planck, p. 82. W. W. Norton & Co., 1932.

way, by coordinating direct knowledge (*e.g.*, the readings of his instruments), experiment, elementary inference, and rationalization. The picture he thus obtains represents the real world of physics.

The discovery of atmospheric pressure illustrates these points. The physicist has been led to believe that the atmosphere exerts a pressure of some 14 lbs. to the square inch. What evidence does he advance for this belief? He certainly does not claim that we are directly aware of this pressure on our bodies. Hence, should science be restricted to a cataloguing of immediate sense impressions, the very notion of atmospheric pressure would have no place in physics. The evidence in favor of atmospheric pressure is entirely indirect. Thus we note that a column of mercury stands at a certain height in a Torricelli tube. Or again, we exhaust the air between two hollow hemispheres and find that we are unable to tear them apart. In the second experiment direct knowledge comes into play through the effort we exert to separate the hemispheres, but only by argumentation and reasoning can we ascribe this effort to the existence of atmospheric pressure. Practically the whole of physical science is thus one mass of inference based ultimately, but not immediately, on direct knowledge.

These considerations entail important consequences. In the grouping of phenomena it is the similarities established in the impersonal world that are regarded as of particular significance; the more obvious similarities detected by our unaided senses are held to be of minor importance. For this reason the physicist claims that visible light and invisible ultra-violet light are kindred in nature, for both are electromagnetic vibrations. The accidental circumstance that the human eye detects the first kind of radiation, and not the second, is viewed as unimportant. Indeed, we need not appeal to theoretical physics to find illustrations of the same tendency. Thus, if we were judging from our immediate sensations, we should presumably claim that sugar and saccharin exhibited a striking similarity, for both are white crystals and have a sweet taste. On this basis we should be tempted to class the two substances into one family. Yet, insofar as the chemist is concerned, sugar and saccharin are entirely different: sugar is a carbohydrate containing carbon, hydrogen, and oxygen; saccharin, though sweet to the taste, has the chemical characteristics of an acid, and its molecule contains carbon, hydrogen, oxygen, and also nitrogen and sulphur. Likewise the botanist places in the same family the potato and tomato plants, or the yellow buttercup and the hooded blue aconite. No obvious similarities can justify such a classification. These examples and many others that could be mentioned bring out the important point that it is not the

obvious that is necessarily credited with any deep significance in science. It is the veiled. This attitude on the part of physicists must not be attributed to any particular love for sophistication or to a failure to note the obvious. It is due to a prolonged experience which shows that the truly important similarities, the detection of which leads to the most fruitful results, are not usually those appearing on the surface.

We may now examine more fully some of the points we touched upon in the previous chapter. We mentioned that in the second, or experimental, stage of the scientific method, no *a priori* assumptions of a speculative character are introduced. More precisely, we meant that the assumptions made are of a type which any man would naturally accept in the pursuit of his ordinary activities and are not of the highly speculative kind that we shall meet with in the theories of mathematical physics. Obviously, even in the simplest experiment assumptions cannot be avoided; ordinary generalization, without which science could not proceed, rests on an assumption. Thus an experiment repeated one thousand times cannot give us any positive assurance that the result will be the same when the experiment is performed once again. On the other hand, if we are ultra cautious and refuse to generalize, experiment becomes useless since it cannot be used as a source of more general knowledge. Faced with the alternative of choosing safety with sterility or doubt with possible progress, the physicist adopts the same course as does the layman in his daily activities; he relinquishes safety and he generalizes. By so doing he is undoubtedly introducing an assumption; namely, the postulate that generalization is permissible. In this tenet consists the principle, or postulate, of scientific induction.

Since the postulate of induction constitutes a mere assumption and is therefore speculative, all empirical laws are necessarily themselves speculative. However, the postulate of induction cannot be regarded as speculative in the same sophisticated sense as are the postulates introduced in the construction of theories, for even in our daily activities we apply induction constantly. Presumably its instinctive appeal is due to a certain inertia of our minds, which find it easier to follow a pattern than to contemplate a new possibility at each step.

Though constantly utilized in science, the principle of induction must be applied with caution. For example, there is no more conclusively established fact of observation than the passage of heat by conduction from a hotter to a colder body. It takes place under the most varied conditions. Small wonder then that Clausius regarded this unilateral passage of heat as exhibiting a fundamental law of Nature (the Law of Entropy). Yet today, for theoretical reasons connected with the kinetic

theory of gases, we believe that this law is only statistically valid, and we have reasons to suppose that if we could experiment long enough, heat would occasionally be found to leave the colder body of its own accord and increase the temperature of the warmer one. We have here an illustration in which the original attempt to apply the postulate of induction led to error.

There is also another difficulty which besets the application of the principle of induction. The principle states that the same experiment repeated under the same conditions will always yield the same result. When stated in this way, the principle is seen to be an expression of the law of causality. Unfortunately we cannot repeat an experiment under exactly the same conditions. The time of the day, the season of the year, the past history of the bodies operated on are subject to incessant change. No two experiments can truthfully be said to occur under exactly the same conditions.

In practice, we overcome this difficulty by disregarding some of the varying conditions, on the plea that their influence is negligible or strictly nil. Since we cannot test all conceivable influences, it is only in virtue of an assumption that we may rule out some of them. The assumptions accepted in this process of elimination are usually of the most plausible kind, but they are more or less speculative nevertheless, and further investigation may prove them to be unwarranted. A simple hypothetical illustration is the following: The first man who placed an egg in boiling water found that the egg became hard-boiled, and in view of the postulate of induction he might have been tempted to express this result as an empirical law. It would have required a mind of uncanny foresight to have felt the necessity of varying the conditions of the experiment by repeating it on a very high mountain; and yet, had this been done, the influence of elevation would have been found to be far from negligible, since the egg would no longer have become hard-boiled.

In the foregoing example our hypothetical experimenter erred in discarding as irrelevant the influence of elevation, but in the majority of cases the experimenter's selection of supposedly irrelevant influences has not been challenged at later times. As an illustration let us consider Galileo's experiments on balls rolling down inclined planes. Galileo certainly did not operate with balls of every hue and color. Hence when he concluded from his experiments that the acceleration of any ball rolling down an inclined plane would be constant, he was implicitly assuming that the color of the ball was irrelevant. Galileo's assumption, in contradistinction to that of the experimenter in the earlier example, is recognized today as fully justified. The point we wish to stress here,

however, is that both assumptions were plausible. We should find quite generally that the assumptions introduced by the experimenters when they disregard certain influences as irrelevant, are always plausible assumptions, even though they may turn out to be incorrect.

Let us examine whether, in addition to the type of assumptions previously enumerated, others may not also be found in the performance of experiments, *e.g.*, in Galileo's. Galileo's measurements of the acceleration of the balls necessitated measurements of spatial distances and of intervals of time. He utilized rigid rods or marks on his inclined planes for the measurement of distances, and sand clocks or others of a similar kind for the measurement of time. We might therefore claim that Galileo was introducing assumptions on the manner in which space and time should be measured. We do not believe, however, that the choice of a system of measurement should be classed as an assumption. Rather, the use of rods and of clocks constituted a definition of the way in which Galileo proposed to measure space and time. As is well known, some kinds of measurements lead to simpler results than others, and such measurements are then claimed to be the correct ones. But so long as the measurements we propose to apply are fully stipulated, a definition and not an assumption is at stake.

In most experiments many of the assumptions merely express the results of earlier investigations. Modern experiments, in particular, rely to a considerable extent on previous experimental results. Thus today when a physicist performs experiments on cathode rays, he assumes from the start that the rays are formed of swiftly moving electrons, an assumption which is merely a conclusion derived from earlier experiments.

Our analysis of the assumptions made in the course of an experiment shows that these assumptions are either of the trivial kind, which any sane man would naturally accept, or that they are assumptions inherited as conclusions from former experiments. But if we pursue a similar analysis for theories of mathematical physics, we shall find that the assumptions introduced are of an entirely different kind. In any theory of mathematical physics one or more speculative assumptions are accepted as a basis. Some restrictions are imposed by common agreement, however. In the first place, the assumptions must not conflict with the facts we are attempting to co-ordinate. Secondly, they must not be inevitable consequences of these facts, for, if they were, they would not constitute independent assumptions. Thirdly, our assumptions must be such as to give rise to observable consequences, preferably of a quantitative nature, so that they may be checked by accurate measurement. If the assumptions

do not lead to consequences which can be observed, we shall never be in a position to know whether they are justified or not, and our theory will degenerate into a set of irrelevant statements.

Newton's theory of the planetary motions is one of the first theories of mathematical physics to have been evolved,* and it will pay us to examine it more closely. As usually happens in the development of a scientific theory, the strictly logical course is not always followed; the investigator may guess at the final result before verifying it. But since our present purpose is to uncover some of Newton's more conspicuous assumptions, we shall adhere to the logical rather than to the historical order of his discovery.

Newton's aim was to construct a theory which could account for Kepler's empirical laws. Kepler's laws, derived from Tycho Brahé's astronomical observations, are as follows:

1. The radius joining the sun to a planet sweeps over equal areas in equal times.
2. The orbit described by a planet is an ellipse with the sun at one of its foci.
3. The squares of the periods of revolution of the various planets are proportional to the cubes of the major axes of the respective elliptical orbits.

Newton first assumed that the planets were massive bodies whose motions were controlled by the same laws of mechanics that were established by Galileo and by Newton himself for terrestrial bodies. In this consists Newton's fundamental assumption. We have here an example of an assumption which differs entirely from those we mentioned when discussing the assumptions of the experimenters. Though plausible, Newton's assumption is speculative and is not one which any man would inevitably make. As a case in point, Aristotle and Descartes, both of whom were concerned with the planetary motions, failed to suspect any connection between the mechanics of terrestrial bodies and that of the planets; they did not therefore make the assumption in question. In more recent times Hegel, the metaphysician, decided that the planetary motions were not governed by the mechanics of earthly bodies. Though there is no excuse for Hegel's aberration, we mention it to stress that Newton's assumption has not appeared inevitable to all men.

If we accept Newton's assumption and the laws of mechanics, Kepler's second law shows that the planets must necessarily be submitted to forces.

* The mathematical theory of mechanics had previously been furnished by Newton.

If there were no forces the planets would pursue straight courses according to the law of inertia. But the existence of forces acting on the planets is about all the information that can be obtained when we rely solely on commonplace logical reasoning; anything else that we might say would be a mere guess. It is noteworthy in this connection that Hooke, who made the same assumption as Newton, surmised correctly that the force acting on a planet was directed towards the sun and was inversely proportional to the square of the planet's distance. Hooke's guess was correct, but precisely because it was a mere guess, it was worthless; and Hooke is not given credit for the discovery of the law of gravitation. To solve the problem rigorously, recourse must be had to that more powerful method of investigation offered by mathematics. Newton's solution of this mathematical problem is considered his major achievement.

Let us mention briefly the results derived from the mathematical analysis. Kepler's first law requires that the planets be acted upon by forces directed towards the sun. Kepler's second law shows that the force must be inversely proportional to the square of the distance from planet to sun. Finally, Kepler's third law requires that the same permanent field of force be acting on all the planets.* These mathematical deductions may be summarized in the statement: Each planet must be attracted towards the sun by a force which is proportional to the mass of the planet, and the magnitude of this force must vary inversely to the square of the distance. Newton next assumed that the sun itself was responsible for the force. The assumption appeared to be reasonable since it amounted to supposing that if the sun were absent, the planets would not move as they did. On the other hand it created a difficulty, for it implied that Kepler's third law could not be correct. The fact is that if the sun were responsible for the forces exerted on the planets, Newton's mechanical law of action and reaction would require that each planet should exert an equal and opposite force on the sun. These forces of reaction would then communicate an accelerated motion to the sun, and this solar motion in turn would preclude the perfect accuracy of Kepler's third law. A conflict between empirical laws and a physical theory usually entails the rejection of the theory, but, for reasons which will be mentioned presently, the degree of inaccuracy which Newton's theory conferred on Kepler's laws was so slight that the discrepancy

* In the absence of the third law, it would be permissible to assume that if the various planets happened to be situated at the same distance from the sun, the intensity of the force acting on them might differ from one planet to another.

was attributed to the lack of refinement of Tycho Brahé's astronomical observations.

In the logical development of the theory the next step introduced the assumption that just as the force exerted by the sun on a planet was proportional to the mass of the planet, so also the equal and opposite force exerted by the planet on the sun was proportional to the mass of the sun. The assumption was plausible, for according to terrestrial mechanics the force of attraction pulling a body towards the earth was known to be proportional to the mass of the body. But we must remember that the forces holding between the planets and sun had not yet been identified with the force of gravitation, and it was therefore hazardous to generalize from one situation to the other. At all events, if this last assumption of Newton's is accepted, the force of attraction of the sun on a planet (and vice versa) is given by $k\frac{Mm}{r^2}$, where M and m are the masses of the sun and of the planet, r is the distance between the two bodies, and k a fixed constant (the gravitational constant). According to Newton's formula, if the mass of the sun were twice as great, the force of attraction on the planets would be doubled, and in particular the earthly year would be shortened. This conclusion in itself gives an indication of the speculative nature of the assumption, for Newton had only one solar system at his disposal and hence could not compare the actions of suns of different masses. Newton, on the basis of his formula, observed that in view of the enormous mass which it seemed permissible to ascribe to the gigantic sun, the discrepancy between his theory and Kepler's third law became so small as to be negligible in practice. For this reason the conflict between Newton's theory and Kepler's empirical laws was regarded as unimportant.

The Newtonian theory indicates that the mass of the sun and the masses of the planets generate fields of force the intensities of which are proportional to the masses. Newton therefore supposed that the mutual attractions between the sun on the one hand, and the planets on the other, should hold among the planets also. Let us observe that Newton was here guided solely by the inner consistency of his theory and not by Kepler's laws. Indeed his new assumption went counter to Kepler's laws, for it indicated that the planets should exert mutual perturbations on one another and hence that Kepler's laws should be even more inaccurate than they were already supposed to be. Calculation showed, however, that, in view of the relatively small masses of the planets, the mutual perturbations should be very small and would certainly pass unobserved unless the planetary positions were charted with great accuracy

over protracted periods of time. For these reasons, the assumption of the mutual attraction of all the bodies in the solar system was seen to be compatible with the facts of observation. Strong support for the theory was found in a somewhat different direction. Thus the moon describes a quasi-circular orbit around the earth, and on the strength of the laws of mechanics this motion is possible only if the moon is attracted towards the earth. If then the earth attracts the moon, there is no obvious reason why it should not also exert attractions on the other planets. Furthermore, the weight of a body on the earth's surface indicates the existence of an attraction, and it would therefore seem plausible to suppose that the motion of the moon and the weight of a body are related phenomena. Newton made this assumption. Fortunately he was able to submit it to a quantitative test. Thus, if the assumption is correct, we must infer that the acceleration of the moon as it falls towards the earth and the acceleration of a falling stone near the earth's surface are manifestations of the same field of gravitational force generated by our planet. This conclusion may be tested provided we know the magnitude of the earth's radius and the distance between the moon and the earth. Newton, on the basis of the information available in his time, uncovered a serious discrepancy between his theoretical anticipations and the observed results. With the true spirit of scientific caution, he accordingly refused to publish his theory. Several years later more accurate measurements of the earth's radius and of the distance between the earth and moon were performed by Picard, and Newton availing himself of Picard's results found that the earlier discrepancies vanished. He then decided, in 1682, to publish his theory of gravitation. By induction he supposed that what was true for the masses of the solar system would also be true for all the masses of the universe. The theory of universal gravitation was the result.

Several important points are revealed by this analysis. We have listed a number of assumptions introduced in the theory, but of these the first alone, which assumes the planetary motions to be controlled by the laws of terrestrial mechanics, is truly speculative; the others are more or less inevitable. This first assumption, owing to its speculative nature, is typical of the assumptions encountered in the theories of physics.

Another feature brought to light by the foregoing analysis is the extreme fragility of a theory of mathematical physics. The slightest discrepancy between the anticipations of the theory and the facts observed may invalidate the theory. An instance was mentioned when we said that Newton did not publish his theory until a numerical discrepancy

which he believed to exist was removed. Of course only when the discrepancies cannot be attributed to errors of observation is a theory in danger; the example of Newton's theory being incompatible with Kepler's laws illustrates a case where the discrepancies are attributed to faulty observations and do not demand the rejection of the theory. Many are the theories, however, which have had to be abandoned because of discrepancies that could not be ignored. The layman usually hears little of such theories, for they are soon forgotten and text-books do not mention them.

The fragility of theories of mathematical physics arises because these theories profess to anticipate accurate quantitative results, which is the feature that enables them to be tested. A theory which anticipates nothing definite or, at best, vague qualitative results is always more difficult to demolish, but this advantage is cancelled by its uselessness. These remarks are well illustrated in Descartes' vortex theory for the planetary motions. In the main all theories of mathematical physics present the aspects we have discussed, though minor differences distinguish one theory from another. In the kinetic theory of gases, the speculative assumption is that a gas is formed of perfectly elastic molecules colliding and rebounding according to the laws of mechanics. A second assumption is the so-called "ergodic hypothesis"; we shall explain it in the chapter on the kinetic theory. The first of these two assumptions is obviously speculative and is by no means one which every man would naturally make. In point of fact an alternative assumption was at one time contemplated; it was that gases were formed of molecules which repelled one another and thereby generated the expansive force. This latter theory was soon abandoned, for it did not appear to be compatible with the gas laws.

In Maxwell's electromagnetic theory the speculative assumption is that all electric currents are closed. Thus we know that if we join two conductors at different potentials by means of a wire, a current flows from one conductor to the other until the potential becomes the same on either conductor. Such a current is called an open one because the path it describes does not close round on itself. The originality of Maxwell's theory consists in the supposition that the circuit, though open in appearance, is in reality closed by the generation of another kind of current flowing from the second conductor to the first through the surrounding ether. This ethereal current, which Maxwell calls a displacement current, is assumed by him to produce the same magnetic effects as an ordinary current. Maxwell's assumption ensures the law of con-

servation for electricity. From the mathematical standpoint it entails the addition of a new term to the equations which, prior to Maxwell's time, were supposed to represent electromagnetic changes. But in whatever way Maxwell's assumption is expressed, it remains extremely speculative; and Maxwell's theory received serious attention only when Hertz, by direct experiment, corroborated one of its major anticipations (electromagnetic waves).

As a last illustration of assumptions at the basis of a theory, we may mention Lesage's assumption that gravitation is due to the bombardment of "ultramundane" corpuscles impinging on material bodies. Lesage himself was not a mathematician, and his theory did no more than state his assumption. But when the theory was subsequently worked out mathematically by G. H. Darwin and by Lorentz, it was found to entail consequences which were in disagreement with facts and was therefore rejected. Here we have one of the many examples of a theory which has been proved untenable.

In addition to the speculative assumptions that vary from one physical theory to another, there is one fundamental assumption which is implicitly contained in all these theories. It consists in the supposition that mathematical analysis is applicable to the physical world.* We shall refer to this supposition as "the mathematical assumption." The validity of the mathematical assumption hinges on the realization of several conditions in Nature, the most important of which is that Nature should manifest a deterministic scheme and hence be controlled by laws. But this first condition by itself would be insufficient; it is also necessary that Nature should display some measure of simplicity, uniformity, and unity. The necessity for this second set of conditions is well illustrated in the game of chess. As Poincaré observes, mathematics cannot apprise us of the strongest moves in chess because there are too many pieces whose moves are entirely different; simplicity, uniformity and unity are lacking.

It would be futile to decide by *a priori* arguments whether or not Nature exhibits the various conditions conducive to the mathematical assumption; all that we can do is be guided by experiment. The mathematical assumption is by no means a dogma which is accepted blindly; it is a mere tentative assumption, the validity of which is subsequently tested when the consequences of a theory are compared with the results

* We are referring here to the applicability of mathematics in its broadest sense, and are not concerned with the purely technical difficulties which the solution of the mathematical problems may present.

of experiment and observation. As such, it is an assumption similar to the one introduced by Newton when he assumed that the planets were controlled by the same mechanical laws as were ordinary terrestrial bodies. It may be that Newton himself started out with a firm conviction that Nature was "mathematical," but, whether this be so or not, the fact remains that the success of Newton's theory, not his personal beliefs, vindicated the mathematical assumption, at least in the restricted case of the planetary motions. In recent times discoveries in the quantum theory have led some of the foremost physicists to suppose that there are no mathematically exact laws in the physical world, and that the best we can attain are probabilities. If this be so, the mathematical assumption cannot have absolute validity, but even then it retains its usefulness to a high degree of approximation in many cases.

Thus far we have been concerned with the general requirements that must be satisfied if mathematics is to be applicable to physical processes. But we must realize that even if all these requirements are satisfied, we cannot be certain that any given physical problem can be solved in practice, even when the laws controlling the system are known. The fact is that the causes affecting the evolution of the phenomenon may be so numerous and varied that their transcription into mathematical form would yield a problem of insuperable difficulty. For example it would be idle for the mathematician to attempt long-range weather forecasting.

Let us then inquire into the further conditions that must be realized for mathematical solutions of physical problems to be obtainable in practice. A first condition is that, in the physical system of interest, the interacting parts which must be taken into consideration should not be too numerous. Now when we realize that our universe forms a unit in which each physical event influences every other physical event, we may well wonder how a sufficiently simple system can ever present itself for mathematical study. For example according to Newton's theory of gravitation, every mass in the universe attracts every other mass; we cannot therefore view the motion of a planet round the sun as due solely to the sun's attraction. A rigorous treatment would require that we take into consideration the actions of all the stars. Fortunately the attraction of the sun is so much greater in its effect than are any of the other influences that we can afford to disregard the other influences. Thanks to this fortunate circumstance the motion of the planet is furnished by the solution of a relatively simple mathematical problem. We conclude that one of the major reasons why mathematics can be applied successfully in practice is that, in many cases, only a very limited number

of influences are predominant; the others may then be treated as non-existent. However, even when a simplification of this sort is permissible, the mathematical solution of a physical problem may still present insuperable difficulties. The relatively simple problem of N gravitating masses in celestial mechanics is an example.

The purely mathematical difficulties which may confront the theoretical physicist in his investigations are responsible for his dependence on the mathematical knowledge of the day and therefore on the discoveries of the pure mathematicians. In a general way the task of the theoretical physicist is to translate into mathematical language the physical problem of interest and the assumptions he has made. He then seeks to deduce the implications of his assumptions and hence of his theory. But to do so, he must solve a purely mathematical problem, and unless he can solve the problem, he is helpless.

To many, the mathematicization of physical phenomena is abhorrent. The contention is that mathematical analysis requires that we single out the quantitative features of a phenomenon, while the qualitative features are disregarded or else artificially reduced to quantitative differences. According to these objectors an undue stress is thereby placed on quantity, with the result that the outlook of the natural philosopher is biased.

The shallowness of such criticisms resides in a confusion between contemplation and investigation. Contemplation makes no demands on the powers of thought; it only affects our emotions, which by their very nature are imprecise and fleeting. In scientific investigation on the other hand, we seek to learn more than the immediately obvious, and since we are not demi-gods, we have to test our theories and expectations. But tests, to be of any value, must be precise. Vagueness necessarily covers a multitude of different possibilities. It is the desire to obtain accuracy which compels men to attach particular importance to quantity and to measurements. We may speak of an object as being light or heavy, but only when we have measured its precise weight in grams or in tons can we attribute any definite value to its lightness or its heaviness. If, for reasons of personal preference, we insist on describing Nature in terms of such vague qualitative attributes as heavy, soft, rough, green, or if we refuse to measure time under the plea that, by so doing, we are mutilating it, how shall we be able to establish any relations or laws that are sufficiently precise to predict anything definite? Our emancipation from quantity will thus have been purchased at the price of ignorance. Possibly in the dim future, the human species will have evolved to such an extent that man will gauge qualitative differences as accurately as

he comprehends quantitative ones today. But until that time comes, we must be satisfied with the forms of knowledge at our disposal, and for this reason the stress must necessarily be placed on quantity in theoretical physics. The entire dispute, as we see it, arises from a conflict between emotion and reason. As a matter of fact, there is no justification for such a conflict, since both emotion and reason have their place in Nature and are not mutually exclusive; the important point, however, is not to confuse them.

CHAPTER III

THE SIGNIFICANCE OF THEORETICAL PHYSICS

PRIMARILY, the aim of a theory of mathematical physics is to coordinate and establish relations among observed phenomena. General laws are sought to unify isolated, disconnected facts. From the standpoint of the natural philosopher this aim in itself would be worthy of effort. But theories of mathematical physics, quite aside from their aesthetic and philosophic significance, have also a considerable practical interest. They often lead to the discovery of unsuspected phenomena, and on many occasions they throw light on matters which observation and experiment alone would be unable to elucidate.

Let us first give a few illustrations of new phenomena which were anticipated by various theories.

(1) Experiment shows that when a narrow beam of light is made to fall on a doubly refracting crystal, the beam is split into two separate beams which proceed in different directions. Hamilton, submitting this phenomenon of double refraction to mathematical investigation, proved that if the incident beam of light were to fall under a certain incidence, it should spread out into a cone on entering the crystal. This phenomenon, called "conical refraction," was unknown prior to Hamilton's calculations but was subsequently verified by experiment.

(2) Poisson deduced from Fresnel's undulatory theory of light that the shadow of a small disk might be represented by a bright patch surrounded by dark rings. The prediction was surprising but was soon verified.

(3) Maxwell predicted, as a result of his kinetic theory of gases, that the viscosity of a given gas at fixed temperature should be independent of the pressure. This fact, though unexpected, was subsequently proved correct.

(4) Lorentz predicted, on the basis of his electronic theory, that the light emitted by a monochromatic source and subjected to the action of a magnetic field should be split into three differently colored radiations. This phenomenon was verified by Zeeman; it is called the "Zeeman effect."

(5) De Broglie's theory suggested the possibility that a shower of electrons would be diffracted on passing through an aperture, much in the same way as waves of light. Diffraction phenomena of this type were detected later by Davisson and Germer.

These examples, though interesting, are less characteristic of the practical importance of theories of mathematical physics than are others we shall mention presently. It is true that they afford illustrations of cases where unsuspected phenomena are predicted and where experiments are suggested to test the correctness of the predictions. To this extent the theories concerned facilitate discovery. On the other hand, we may presume that the experiments suggested would eventually have been performed even in the absence of the theories. Someone, at some time or other, would accidently have made light fall on a crystal in the way requisite to yield the phenomenon of conical refraction predicted by Hamilton, so that this curious phenomenon would have been discovered in any case. Similarly for the Zeeman effect: Even the most casual observer could not fail to notice it by means of a modern spectroscope of high resolving power. Finally, de Broglie's theory had so little to do with the discovery of Davisson and Germer that these investigators made their important discovery by accident. We may add that the greater number of important experimental discoveries have been made by accident, or at least without the assistance of any mathematical theory. Radioactivity, X-rays, the diffraction of X-rays by crystals are phenomena which were discovered accidentally.

However, we may mention many illustrations in which the usefulness of physical theories is more apparent. For instance, before Newton constructed his theory, astronomers could have foretold the future position of a planet by applying Kepler's laws. But they could not have predicted the path and motion a stray comet would pursue on entering the solar system. Thanks to Newton's theory this information became available. Furthermore, as we mentioned on a previous page, the theory of gravitation leads to a correction of Kepler's laws and to a recognition of their approximate nature. Theory in this case remedies the imperfections of observation. It is also by relying on the theory of gravitation that we can compute the masses of the various bodies of the solar system. In the absence of the theory of gravitation, we should be helpless. Much of our knowledge in modern physics (for instance the mass of the electron) is obtained indirectly by taking into consideration some mathematical theory. Other illustrations follow:

Michelson, in one of his experiments, sought to measure the velocity of light passing through carbon disulphide. His experiment consisted in directing a beam of light on an aperture that could be opened and closed by a shutter. As the shutter was rapidly opened and closed, a jet, or "packet," of light passed through the hole and was transmitted

through the carbon disulphide. Michelson assumed that the velocity with which this packet was propagated was the velocity of the waves of light through the medium. Great was his surprise, therefore, when he found that the velocity of the packet of light fell considerably below the velocity that light was expected to have in carbon disulphide. Michelson's result was so unexpected that physicists were at a loss to account for it.

The solution of the puzzle was given by a theoretical physicist, the late Lord Rayleigh. Rayleigh made a mathematical study of the transmission of wave motions and found that in some cases a packet of waves might advance with a velocity that differed considerably from that of the wave crests. Two different velocities were thus uncovered: the velocity of the packet, or the "group velocity," and the velocity of the individual wave crests, or the "wave velocity." In the case of waves of light transmitted through a highly dispersive medium, like carbon disulphide, Rayleigh's investigations showed that the velocity of a packet of light (the group velocity) should fall considerably below the velocity of advance of the individual waves (the wave velocity). Calculation then proved that the group velocity coincided precisely with the velocity measured by Michelson. The error was not in Michelson's measurements, but in his interpretation of his experiment. Unwittingly, he had measured the wrong velocity, namely, the group velocity, instead of the wave velocity (commonly called the velocity of light), which it was his intention to measure. At the same time, Rayleigh's investigations showed that the earlier measurements of the velocity of light, performed by Römer, Foucault, and Fizeau, suffered from the same confusion. In each case the group velocity, not the wave velocity, was measured. Thus, thanks to Rayleigh's theoretical investigations a new interpretation was placed on the significance of certain experiments, and the misconceptions of the experimenters were corrected. A more detailed analysis of Rayleigh's wave packets will be given in Chapter XIX.

A second example of the practical utility of physical theories concerns the principle of entropy. Common experience corroborates this principle. For instance, we have no reason to suppose that heat can leave a piece of ice and warm a neighboring fire. Yet, when Boltzmann developed the kinetic theory of gases, he showed that even if this theory were only approximately correct, the principle of entropy could not possibly have absolute validity. As may well be gathered, this discovery has considerable importance, and we note that it was arrived at by

means of a mathematical theory and not by direct experiment. As a third example, we may mention the microscope. The construction of the microscope was revolutionized by Abbe, not by tentative modifications here and there, but as a result of an application of the undulatory theory of light.

CHAPTER IV

TWO KINDS OF SIMPLICITY

WHEN we say that one theory of mathematical physics is simpler than another, the epithet "simple" may be interpreted in two ways.

(a) *Simplicity in the Assumptions.* Initial assumptions are inevitable in any physical theory, for some assumptions must always be made when the physical facts are transcribed into mathematical language. But though initial assumptions are necessary, they must not be too numerous; and above all we must not introduce a new assumption to account for each new experimental fact. In more technical language, this last reservation is expressed by the statement that hypotheses *ad hoc* must be avoided. A physical theory which satisfies the foregoing restrictions is said to be "simple in its assumptions."

There are two reasons for requiring this form of simplicity. First let us recall that the aim of a physical theory is to deduce mathematically the implications of a given body of assumptions, or hypotheses. If, then, to the fundamental hypotheses others are added in the course of our deductions, we shall be unable to extrapolate the theory to any extent, for we shall never be assured that the introduction of new hypotheses is at an end. A second reason why theories that contain too large a number of hypotheses *ad hoc* must be rejected is that we can account for anything we wish if we allow ourselves the privilege of introducing corrective assumptions as and when needed. A theory that accounts for anything we choose is useless.

These considerations are illustrated in a theory tentatively advanced by Jeans at the Solvay Congress of 1911. The radiation which experiment proves to be present in a heated enclosure had been accounted for by Planck by means of his revolutionary quantum theory. To obviate the strange quantum theory, Jeans, supported by Lord Rayleigh, claimed that the radiation could be interpreted along classical lines if we assumed that the enclosure were to leak in an appropriate way. He pictured the situation by imagining an enclosure connected to pipes, and he assumed that these pipes leaked in such a manner as to account for the facts observed.

Commenting on Jeans's suggestion, Poincaré made the following observation:

"It is obvious that, by giving suitable dimensions to the connecting pipes between the enclosures and suitable magnitudes to the leaks, Jeans can account for any experimental result. But this is not the rôle of physical theories. Physical theories must not introduce as many arbitrary constants as there are phenomena to be accounted for; they must establish connections among the various experimental facts and, above all, must lead to predictions." *

Jeans's suggestion was not considered further, and Planck's theory was preferred in spite of its unfamiliar concepts.

Many are the physical theories that have been abandoned because of defects of the foregoing type. Usually, however, such defects do not appear in the early stages of a theory; only as the theory develops and is confronted with new facts does the necessity for introducing additional hypotheses *ad hoc* arise. The theory then becomes cumbersome and must be remodelled from its foundations. Bohr's theory of the atom is a case in point. We shall see nevertheless that the scrapping of a theory does not mean that the theory has been useless. Some of its more important results will remain, and usually it serves as a guide to the more accurate theory that supersedes it. The urge to attain simplicity in the assumptions is not characteristic solely of mathematical physics. In our commonplace attempts to coordinate facts, we also follow the same urge, and we usually prefer a simple, direct explanation to an involved one.

(b) *Mathematical Simplicity.* The kind of simplicity we have just discussed is an essential requirement in any theory of mathematical physics. But we must now mention another important kind of simplicity. It may be understood when we say that the theoretical physicist (and for that matter also the experimenter and the observer) operates as though he believed in the mathematical simplicity of Nature.

For example, Kepler concluded from Tycho Brahé's observations that the planets described those mathematically simple curves called ellipses. Yet Kepler must have realized that Tycho Brahé's observations could boast only of a limited measure of accuracy, and that orbits differing slightly from perfect ellipses would also be consistent with Brahé's findings. Nevertheless Kepler chose the mathematically simple curves. It is true that he was a firm believer in the simplicity of Nature, and that for this reason he would presumably have dismissed complicated orbits as impossible. But this is not the point at issue. Today we fully realize

* Solvay Congress of 1911.

that Nature is not simple, and yet any modern astronomer would follow the same course as Kepler. In any case if, in view of the inaccuracies of observation, the evolution of a phenomenon may be represented equally well by many different graph lines, one of which is mathematically simpler than the others, the simpler graph is preferred.

This same rule is observed in theoretical physics: The simplest mathematical solution is the one that is accepted. Newton started from Kepler's laws and found that, in accordance with his own mechanical assumptions, the sun must attract the planets with a force varying inversely to the square of the distance. There is no doubt that this is the only law that can be accepted if we assume Kepler's laws to be rigorous. But Newton on developing his theory was compelled subsequently to recognize that Kepler's laws could not be rigorous. As a result the necessity for retaining the law of the inverse square was no longer present, and Newton might just as well have selected any one of a vast number of other laws, differing slightly from the classical one and yet satisfying the requirements of the astronomical observations of the day. All other laws, however, would have been more complicated from the standpoint of the mathematician, and, besides, it would have been gratuitous to select one of them rather than the other. Newton accordingly retained the law of the inverse square—the simplest of them all. Only two centuries later, after more accurate observations of the planetary motions over long periods of time had been performed, was the classical law suspected of being incorrect. The more refined astronomical observations, in conjunction with the discoveries in electromagnetics, then gave rise to the general theory of relativity. Here, also, Einstein pursued the same method and chose the simplest law among the ones compatible with his assumptions and with the facts of observation. These considerations show that the modern scientist, who no longer believes in the inherent simplicity of Nature, can claim no theoretical justification for singling out the simplest mathematical law in any particular case. His urge is purely pragmatic and arises from the fact that the mathematical implications of a simple law are more easily obtained than are those of a complicated one.

The two types of simplicity we have discussed under the headings (a) and (b) have often proved confusing to the layman. There is some justification for this confusion, because we often find a competent author stating that Newton's law of gravitation is simpler than Einstein's and then a few pages later making the opposite statement. Both statements are correct, but they do not refer to the same type of simplicity. For instance, suppose we are called upon to determine, in accordance with

Newton's law, the motions of the two stars of a double-star system (Two-Body Problem). The solution is exceedingly simple; it was given by Newton himself in spite of the limited mathematical knowledge of his time. But suppose we wish to solve this same problem by utilizing Einstein's law of gravitation. The problem is so difficult that it has never been solved. The reason is that Newton's law is mathematically simpler than Einstein's; and hence from the standpoint (b) (of mathematical simplicity) Newton's law is the simpler.

On the other hand, when account is taken of certain astronomical phenomena unknown to Newton, we are justified in saying that from the standpoint (a) (of simplicity in the assumptions) Einstein's law is the simpler. Thus Einstein's law accounts for the peculiar motion of the planet Mercury, for the red shift in the lines of the solar spectrum, and for the precise bending of a ray of star-light that grazes the sun's limb. Newton's law accounts for none of these phenomena unless we supplement it with suitable assumptions devised for the occasion. But the complications we thereby introduce imply that the classical Newtonian theory, when confronted with the more recently discovered astronomical facts, is lacking in that form of simplicity which we called "simplicity in the assumptions." Obviously in this respect, Einstein's law of gravitation is the simpler.

CHAPTER V

SUCCESSIVE THEORIES OF MATHEMATICAL PHYSICS ARE PROGRESSIVE APPROXIMATIONS

CLOSELY allied with the ideas of simplicity developed in the foregoing chapter, is the fact that the successive theories of physics pertaining to the same field are progressive approximations. As we shall see, the implications of this situation have a heuristic value in the theory of relativity and in the quantum theory. Earlier experimenters gathered their information with crude instruments, and sometimes with no instruments at all. In addition, all the experiments they were able to make were performed under conditions that deviated only slightly from those with which we come in contact in our commonplace existence. Temperatures ranged from that of ice to that of fire; pressures were of the order of that of the atmosphere; velocities were low, far less than that of light; masses were of average magnitude, neither physically infinitesimal nor enormous. We may designate this range of experience as the commonplace level. Then, as knowledge increased and the technique of experimentation became more perfect, other levels were explored. Such levels refer, for instance, to temperatures in the neighborhood of the absolute zero, to exceedingly high vacua, to enormous pressures, to velocities nearing that of light, to masses that are physically infinitesimal. We shall call these levels the remoter levels. Even today there are levels which have never been explored. As yet we cannot experiment under the enormous pressures and high temperatures that are thought to reign in the interior of the stars.

The earlier experimenters were conversant solely with facts gathered from the commonplace level, and the aim of the earlier physical theories was to coordinate these commonplace facts. But a physical theory does not restrict itself to a coordination of what is already known; it also predicts other phenomena, many of which may pertain to the remoter levels of experience. Until such time therefore as these remoter levels are explored, we may be unable to submit the earlier theories to any decisive test. In practically every case the exploration of the remoter levels has necessitated a revision of the earlier theories, and the result has been that the history of theoretical physics is a story of successively discarded theories.

What we now propose to show is that although the successive theories of a given class may differ considerably in the concepts they introduce, yet from the mathematical point of view each theory is merely a refinement of its predecessor. The successive theories thus form a chain of progressive approximations.

Since the evolution of the law of gravitation is one of the simplest to follow, we shall devote some space to its study. Galileo's observations were restricted to regions in close proximity to the earth's surface; hence the experiments of his day showed no trace of any decrease in the weight of a body when its elevation was increased. The natural inference was that the gravitational field developed by the earth was the same at all elevations.

Newton shattered this belief when he proved that the intensity of the field decreases as the elevation is increased. But what is important for our present purpose is to examine the method that led Newton to his discovery. Newton could not have corrected Galileo had he merely confined his attention to the experiments of his great predecessor. Even today no experiment could detect a difference in the weight of a body when situated on the earth's surface or at a few feet in elevation. Only by seeking to coordinate a larger number of facts drawn from a remoter level, notably, by extending his observations to the moon and planets, was Newton able to progress.

Another point to be noted is that Newton's law of the inverse square passes over into a law of constant force when we restrict our attention to points in the immediate neighborhood of the earth's surface. The reason is that, whether a body be resting on the surface of the ground or be set at the top of a high tower, the difference in its distance to the earth's centre is negligible in comparison with the radius of our planet. Newton's law thus requires that the weight of a body should be practically the same in the two situations. Hence we are justified in saying that Galileo's belief expressed a first approximation to truth, an approximation which tended to become rigorous for small elevations. In short, Newton's law of the inverse square appears as a refinement of Galileo's idea.

For more than two centuries Newton's correction was deemed adequate. Then came the general theory of relativity. Newton's law was found in turn to be but an approximation, a far better one than that of Galileo, but an approximation nevertheless. The correction of Newton by Einstein came about in exactly the same way as that of Galileo by Newton. It was not by performing more refined measurements on the planetary motions or on the fall of bodies that Einstein corrected New-

ton's theory; it was by exploring new levels and extending thereby the range of the facts to be coordinated. The general theory of relativity is indeed but a plausible outgrowth of the special theory, and this latter theory is based on facts unknown to Newton; namely, on the results of highly refined electromagnetic experiments and on the behavior of beta-particles moving with speeds comparable to that of light. We are here in the remote level of extremely high velocities. As is well known to students of the mathematical theory of relativity, Einstein's law of gravitation differs less and less from Newton's as the intensity of the gravitational field and the velocities of the bodies concerned become smaller and smaller. In other words Newton's law tends to become correct under such limiting conditions. For this reason Einstein's law of gravitation may be regarded as a refinement of Newton's law, just as the latter is a refinement of Galileo's belief.

We have therefore to inquire whether Einstein's theory is the final revision or whether it is itself only a more refined approximation. Quantum mechanics proves conclusively that the theory of relativity cannot be the last word. Hence we might suppose that further investigation would confront us with an unending series of more and more accurate theories embracing an ever widening array of facts. Absolute truth and a final picture would be beyond our reach. A slight modification of this statement must be made if we accept the conclusions of the leading quantum theorists. According to Heisenberg, Bohr, Born, and Dirac, the scale of successive approximations will not extend indefinitely, for, owing to the Uncertainty Principle, the limit of accurate knowledge has already been reached in the quantum theory. Beyond, lies a fog of uncertainty due to the peculiarities of Nature herself. All that we can know are probabilities. It is unnecessary at this stage to venture an opinion on the merits of the new quantum philosophy, for in any case one conclusion remains unchanged: absolute truth is beyond our reach.

Let us now examine the implications that result from the progressive approximations of scientific theories. We have said that, when we pass from the remoter levels of the high velocities and enormous gravitational fields (in which the discrepancies between Newton and Einstein become perceptible) to the more commonplace level considered by Newton, the theory of relativity gradually merges into Newtonian science. Also, if we concentrate on the still more commonplace level near the earth's surface, investigated by Galileo, we find that both the relativity and the Newtonian theories merge into Galileo's beliefs.

Thus is illustrated an elementary rule of great importance in theoretical science. It may be condensed into the following statement: A more

accurate theory passes over successively into the less accurate ones that have preceded it, as we step down from the remoter levels to the more commonplace ones. This statement is general, and its justification may be understood when we recall that the more accurate theories are mere refinements of the preceding ones and are arrived at by considering wider and wider arrays of facts. Numerous illustrations of this general rule may be found in the history of science. Let us mention some of them.

In the theory of relativity, gravitational forces, when compounded, do not obey the elementary rule of the parallelogram, as Newton believed. This is so because Einstein's gravitational equations are not linear. Yet, if we assume the forces to become weaker and weaker, the rule of the parallelogram holds in the limiting case. Similarly if, instead of considering a fast-moving planet like Mercury, we are dealing with one moving more leisurely with respect to the sun, *e.g.*, Neptune, Einstein's law of gravitation approximates more and more to Newton's law of the inverse square.

The curious rule for the addition of velocities in the special theory of relativity passes over into the classical rule when the relative velocities involved become smaller and smaller. Simultaneity tends to become absolute, and 4-dimensional space-time passes over into separate space and time.

In the new statistics of Bose-Einstein and of Fermi-Dirac we also find that the formulae pass over into the classical formulae of Boltzmann's kinetic theory when the temperature and pressure are assumed to approximate to their more habitual commonplace values.

These examples also show why textbooks continue to utilize Newtonian mechanics even though their authors know that relativity mechanics is a better approximation. The reason is that, for the small velocities with which we are usually concerned, the two mechanics yield results so nearly alike that in practice no experiment would be sufficiently refined to detect their difference. And since Newtonian mechanics is mathematically the simpler, as explained in the previous chapter, there is every advantage in retaining it.

Heuristic Considerations—Of invaluable help in discovering the laws of the newer theories is the realization that the successive theories must become indistinguishable when the commonplace level is neared. In the special theory of relativity for instance, one of the fundamental assumptions is that all natural laws should satisfy a certain condition of invariance.* This requirement rules out the classical mechanical laws

* Invariance under the Lorentz transformation.

of Galileo and Newton. Unfortunately it leaves open a wide choice of conceivable mechanical laws and does not restrict us to any specific law. We might, of course, avail ourselves of the experimental measurements performed on swiftly-moving beta-particles, and from these experiments derive the amended mechanical laws. But the experiments are so difficult and so scant that their results could not furnish a sufficient basis for the formulation of exact laws. It is here that the heuristic considerations come into play. Thus we know that the mechanical laws of the special theory of relativity must tend to merge into the classical ones when we are dealing with situations on the commonplace level (low velocities). This clue, taken in conjunction with the general requirements of invariance of the theory of relativity, suffices to establish the new mechanical laws.

Precisely the same situation occurs in the general theory of relativity. Here the requirements of invariance show that Newton's law of gravitation cannot be correct, and, as before, these requirements are not sufficiently stringent to impose any specific amended law. But when we remember that for low velocities and weak gravitational fields the new law must pass over into the classical law of the inverse square, we are furnished with a clue that enables us to obtain the amended law of gravitation.

More so than the theory of relativity, does the quantum theory offer examples of the application of the method just outlined. In the quantum theory we are no longer guided by the general requirements of invariance, which play so large a part in the theory of relativity; and so more groping in the dark must precede our formulation of the quantum laws. Before showing how the quantum theorists proceeded, we must mention that the refinements of the quantum theory become noticeable only when we pass from the commonplace level to the level of very small masses, the level of the electron for example. More precisely the level we refer to is the one in which the "action" * of the mechanical system of interest is exceedingly small. If, then, we follow the inverse course, passing from the remoter level to the level of common experience, we must expect the quantum laws to pass over into the classical ones. Thus, the same heuristic clue, which we found so effective in the theory of relativity, is also available in the quantum theory. But whereas in Einstein's theory no particular name was given to this clue, in the quantum theory a name has been coined. The clue is called "Bohr's Correspondence Principle,"

* Action is a technical concept that it has been found advantageous to introduce into dynamics.

and it is of considerable importance in the formulation of the quantum laws.

A Comparison of the Levels of the Theory of Relativity and of the Quantum Theory—Both the relativity and the quantum theories have issued from the necessity of amending the classical laws when remoter levels were considered. This prompts us to inquire: Which of the two levels, that of relativity or of the quantum theory, is the more remote from the commonplace level? If we can answer this question, we shall have determined which of the two theories is the more accurate. Unfortunately no straightforward answer can be given because the levels involved are complementary rather than successive. The level of high velocities and powerful gravitational fields (for which the refinements of the theory of relativity are needed) is totally different from the level of microscopic masses and microscopic quantities of action characteristic of the quantum theory. And so, if we have average masses moving at enormous speeds, we must apply the relativity theory but may disregard the quantum theory. Conversely, if our interest lies with microscopic masses moving at low velocities, we must have recourse to the quantum theory and may disregard the relativity theory. It is obvious, however, that a theory, more rigorous than either of the two theories individually, would result from their fusion. Such a theory would correspond to the level relating simultaneously to high velocities, to microscopic masses, and to microscopic quantities of action.

The first attempt to combine the two theories was made by Sommerfeld when he applied relativity mechanics to Bohr's atom. But Sommerfeld's results have lost much of their interest, because Bohr's atom has since been shown to be only an approximation to the more general concepts developed in the quantum mechanics of Heisenberg, Dirac, and Schrödinger. An amalgamation of the more advanced quantum theories with the theory of relativity was effected subsequently by Dirac in his celebrated paper on the nature of the electron. The two theories here impose mutual refinements one on the other. Dirac's theory has since inspired other investigators.

CHAPTER VI

CONCEPTUAL CHANGES IN PHYSICAL THEORIES

IN THE previous chapter we have seen that the successive physical theories proceed by progressive mathematical approximations. But we must be careful to understand that this statement is intended to apply more particularly to the mathematical equations which form, so to speak, the scaffolding of the theory. The revision of concepts, when we pass from one theory and level to another, is usually much more drastic. For instance, according to Newtonian science, the planets are subjected to forces pulling them towards the sun—an example of action at a distance. The general theory of relativity, though but a mathematical refinement of the Newtonian theory, rules out completely any such action at a distance. We must now say that the planets follow the geodesics of 4-dimensional space-time which is warped in the neighborhood of the sun. It is this warped structure of space-time, in the immediate vicinity of the point occupied by the planet, that guides the planet along a curve from one instant to another; and Einstein's law of gravitation expresses the general nature of the space-time curvature, or warping. As may be gathered, the conceptual departure from the Newtonian scheme is very great.

In a number of cases concepts which appear natural on the commonplace level lose all meaning on the remoter levels. One of the simplest examples is afforded by the concept of temperature. On the commonplace level, warmth and cold appeal directly to our senses, and the concept of temperature is most useful in a science like thermodynamics, which deals largely with commonplace phenomena. But when we pass to the level of molecular dimensions, the entire concept of temperature becomes meaningless. There is no sense in speaking of the temperature of a single molecule, for a large number of molecules in motion is necessary to give meaning to temperature. And just as some concepts may become meaningless when we pass to remoter levels, so also new concepts may arise. We have only to mention the 4-dimensional space-time of relativity. Here we have a blending of space and time into a 4-dimensional continuum, whereas on the commonplace level space and time appear to be entirely separate. We cannot usually give expression in words to new concepts that differ widely from those of the commonplace level. Only by perse-

verance and prolonged meditation can an idea of what is implied force itself upon us. To impediments of this sort most of the leading theoretical physicists ascribe the present conceptual difficulties in the quantum theory. As Bohr writes:

"In the adaptation of the relativity requirement to the quantum postulate, we must therefore be prepared to meet with a renunciation as to visualization in the ordinary sense going still further than in the formulation of the quantum laws considered here. Indeed, we find ourselves here on the very path taken by Einstein of adapting our modes of perception borrowed from the sensations to the gradually deepening knowledge of the laws of Nature. The hindrances met with on this path originate above all in the fact that, so to say, every word in the language refers to our ordinary perception." *

Even mathematics, which had long been thought immune from conceptual changes of this sort, appears today to be entering the same troubled region. According to Brouwer, the logical principle of excluded middle, which appears so inevitable in our everyday life, must be rejected where infinite aggregates are concerned. The exalted laws of logic, long assumed to be the laws of the human mind, thus appear to be mere empirical laws, subject to revision as the field of our mathematical experience is widened.

Modern physics owes a large part of its difficulty to the apparent inappropriateness of many of our former concepts; and to enlarge and modify these concepts much of the labor of modern theoretical physicists is directed. The necessity for enlarging our ideas in natural philosophy is not surprising. To take a simple illustration, we have but to imagine a man who, after having lived for years in his home town out of contact with the world at large, decided to travel. He would discover strange people, climates, trees, and animals, of which he had never dreamed, and he would return with a less dogmatic viewpoint. His views would conflict with those of his fellow citizens who had remained at home, believing the entire world to be like the little world of their limited experience—a belief they would doubtless back up with arguments they would call "logical." It is the same in natural philosophy. Cooped as we are in the world of common experience, we must endeavor to break away and explore the remoter levels, and construct theories befitting not solely the commonplace levels, but also all those we have explored. Only thus can knowledge (limited at that) be acquired. Those who have failed to explore will naturally construct theories compatible with their restricted

* Atomic Theory and the Description of Nature; Cambridge, 1934; p. 90.

experience but entirely out of touch with the wider fields which they ignore.

The Greeks have given an illustration of the stagnation of knowledge that results from a restricted viewpoint, and it will always be to the immortal glory of Galileo and Newton that they founded the methods of investigation which initiated intellectual emancipation. At the same time, the restricted viewpoint accounts for the feeling of diffidence which the layman experiences when informed of many of the conclusions of the modern theories, *e.g.*, the paradoxes of the theory of relativity. He feels that these conclusions conflict too strongly with the notions he has formed from his long association with the commonplace level, and, like the man who has never left his home town, he believes that the whole unexplored universe should conform to the ideas he happens to cherish. His arguments are vain, for we have seen that there is no real conflict between the disclosures of the more advanced theories and those drawn from our daily activities: the former tend to merge into the latter when we pass from the remoter levels to the level of common experience. An apparent conflict arises only when we pursue the unjustifiable course of assuming that all the levels are mere repetitions of the one we know best. The greatest revolutions in modern thought have issued from a study of the remoter levels, and they would never have arisen had these levels been ignored.

CHAPTER VII

CAUSALITY

COMMON observation suggests the presence of causal connections in inorganic Nature; even the savage or child acts on the assumption that the same situations produce the same effects. In Greek antiquity there arose the doctrine of fate and destiny. We find the doctrine expressed in the tragedy of "Oedipus Rex," and it survives to the present day in such expressions as "being born under a lucky star." But this form of predetermination is vague and fails to specify exactly when and how the fortunate or unfortunate events will occur. Furthermore no unbroken causal chain extending from the initial event to its consequences is indicated.

The first rigorous formulation of the doctrine of causality and its mathematical expression by means of a differential equation are due to Newton. Until recent times Newton's presentation has satisfied the requirements of natural philosophy. Whether the doctrine can still be entertained or whether it must be rejected entirely is a moot point, the discussion of which will be taken up in later chapters.

The popular understanding of the doctrine of causality may be condensed into the statement: "The same causes generate the same effects." So vague a statement can be of little use in a rigorous scientific discussion. For instance, a stone is released and falls to the ground. What constitutes the cause of the fall? Is it the gravitational pull of the earth, or are we to regard the gravitational pull as given, and identify the cause with the act of releasing the stone? We might also suppose that the release of the stone and the gravitational pull are to be regarded jointly as representing the cause. Further definitions are obviously needed before we can agree on what constitutes the cause in any particular case.

Nevertheless the popular idea of causality may serve as an introduction to the scientific formulation of the doctrine. The doctrine necessarily embodies a belief in the presence of order and regularity in Nature, for without these characteristics the same causes would not always give rise to the same effects. In the scientific approach, regularity in Nature is connected with the existence of natural laws, which govern the different classes of phenomena (*e.g.*, the law of gravitation). The natural laws,

however, owing to their generality, impose only a limitation on possibilities and do not map out any definite course for future events. For this reason, even if we are acquainted with the law which controls the workings of a system, we cannot predict the evolution of the system. To determine the latter, we must supplement the general law with additional information. We shall illustrate these features by discussing a mechanical system.

Suppose our system is formed of point-masses moving freely under their mutual gravitational attractions. The general laws controlling the evolution of the system are known: they are the laws of mechanics and of gravitation. We wish now to determine how the configuration of the system will change in the course of time. The laws alone cannot give this information, for the future configurations of the system are necessarily affected by the initial configuration from which we start. Additional information is thus required. The additional information which suffices to make the problem determinate may be expressed in various ways, but there is one mode of expression which harmonizes with the requirements of the doctrine of causality. It consists in the specification of the positions and of the velocities (in magnitude and in direction, as referred to a Galilean frame) of the several point-masses at some initial instant. The additional information, when expressed in this way, is usually said to constitute the *initial conditions* of the system. When the initial conditions are specified, the application of the general laws permits us (in theory) to predict the configuration of the system at any future instant, and also to determine its past configurations.

The contingency of the initial conditions is in marked contrast with the permanency of the laws. Thus we may assign any initial positions and velocities we choose to the point-masses at the initial instant, and, according to the initial conditions selected, the evolution of the system will proceed along one line or another. But the laws governing the evolution will always be the same. The instant of time which we have referred to as the initial instant may be chosen arbitrarily, and we may assume it to be the present instant or some past instant or even some instant in the future. This last choice might, however, lead to confusion, and we shall not consider it for the present. It is often convenient to say that the initial conditions define the initial *state* of the system. The initial state of a mechanical system of point-masses is therefore represented by the positions and velocities of all the point-masses at the initial instant. More generally the state of the mechanical system at any instant—past, present, or future—is defined by the positions and velocities at the instant

of interest. Generalizing from mechanics to all physical systems, we may formulate the doctrine of causality as follows:

The evolution of every physical system is controlled by rigorous laws. These, taken in conjunction with the initial state of the system (assumed to be isolated), determine without ambiguity all future states and also all past ones. The entire history of the system throughout time is thus determined by the laws and by the initial state.

The scientific presentation of the doctrine of causality removes much of the vagueness which surrounds the commonplace concept of "cause." Let us revert to the example of the falling stone. The cause of the fall might popularly be identified with the act of releasing the stone, or with the law of gravitation, or with both. Thus, the cause might be the initial state, or the law controlling the evolution of the system, or both. Usually, however, in the ordinary vernacular, a "cause" refers to an instantaneous event; and if we accept this restriction, the law, which acts over a protracted period of time, cannot be identified with the cause. The sole candidate for this title is then the initial state of the system. We shall therefore identify the cause with the initial state. As a result of this identification, the law controlling the evolution of a phenomenon is seen to act as the connecting link between the cause and its effects: it represents the causal chain.

Several points in our statement of the scientific doctrine of causality require elucidation.

1. Let us assume that the initial instant is the present one. The scientific doctrine of causality then implies that the present state determines not only the future states, but also the past ones. However, we are violating the customary use of words when we speak of the present as determining the past; for this mode of expression would imply that a present condition is the cause of past events, whereas in the commonplace understanding of causality, the passage of time plays an important part, and it is customary to view the cause as preceding the effect. We may, however, attenuate the conflict between the scientific doctrine and the more familiar conception by noting that the statement "the present determines the past" is equivalently expressed by saying: "If the present state is such and such, the past state must necessarily have been so and so."

It is easy enough to give any number of illustrations which seem to refute the scientific law of causality in its reference to the past. For instance a small piece of ice is thrown into hot water. We know that the ice will melt and disappear and the water attain a lower temperature. There is no difficulty in drawing conclusions in this case because the conclusions concern the future. But suppose now we wish to reverse the

order of our deductions and discover the past from the present state of the water. This appears impossible, for how can we possibly tell whether there ever was any ice in the water? The same present state of the water might have been attained in many different ways. These conclusions, however, are too hasty. To investigate the situation properly, we must take into consideration the microscopic constitution of the water. Thus, if we view the liquid as a mechanical system formed of myriads of water molecules, its present state is defined not by the uniform temperature of the mass of water, but by the positions and velocities of all the individual molecules; and we may perfectly well suppose that the present state, when defined in this way, determines the past history of the water. At all events, none of the commonplace objections can refute this view.

2. In ordinary conversation we say that the present is the cause of the future. But the present to which we refer is not necessarily assumed to consist in a mathematical instant of time. Usually, by the word *present* we wish to imply a more or less protracted period. For instance we receive a telegram and act accordingly. The reception of the telegram is the cause of our subsequent actions. Yet it is not the actual reception of the telegram that dictates our future course; it is the reading and the comprehending of the message—and this reading is not instantaneous. In the scientific formulation of the law of causality, the situation is different, for the cause (such as we have defined it) refers to the initial state and is thus an instantaneous event. We might, if we so chose, extend the scientific doctrine so as to cover cases where the initial conditions would no longer be instantaneous but would be spread over some interval of time. This course has not been followed, however.

3. We have said that the initial state determines all other states. But the fact that the initial state determines all others does not necessarily mean that a knowledge of the initial state automatically confers a knowledge of the evolution of the system. No information can be derived unless we are acquainted with the general laws which control the workings of the system. In practice, even a knowledge of the laws might prove useless, for, in order to make use of this knowledge in any given situation, we should have to solve a mathematical problem whose difficulty might defeat all efforts.

4. Our allusion to an isolated system in the statement of the law of causality is readily explained. If our system is not isolated, it is subjected to the influences of the outside world, and we are no longer justified in saying that the history of the system is determined solely by the initial state and the internal laws. To overcome this difficulty, we must enlarge the system so that it will comprise the external influences alluded to.

Proceeding in this way, we should eventually have to consider the entire universe, for no limited portion of the universe can be regarded as strictly isolated. We conclude that the causal doctrine (if it be valid at all) can apply only to the entire universe. Laplace, in his presentation of the doctrine, stressed this last feature. In practice, however, some limited systems, such as the solar system, may be viewed as very approximately isolated, and it is with such systems that we shall be concerned.

5. We have illustrated the principle of causality in connection with mechanical systems formed of point-masses, and have seen that in such systems a state is defined by the positions and velocities of all the masses at the instant of interest. This definition of a state is necessarily peculiar to mechanical systems, and we must now give a more general definition which will apply to physical systems of all kinds. The general definition of a state is the following: The state of a physical system is specified by the aggregate of those items of information which pertain to the system at any given instant and which, when taken in conjunction with the laws controlling the system, suffice to determine its history.

Let us now examine the procedure which will furnish the items of information necessary to determine a state in any specific physical system. The first step will be to establish the laws which control the evolution of the system and to express these laws in the form of differential equations. Now a characteristic of differential equations is to define restrictions which are not sufficiently comprehensive to determine any precise result. (This feature is indeed one of the reasons which make differential equations so suitable for the expression of natural laws.) To render determinate the implications of a differential equation, we must furnish complementary specifications. A study of the differential equation shows what forms these specifications must assume. Suppose, then, that we have translated the laws controlling the system of interest into the mathematical form of differential equations. If we can establish the specifications which render these equations determinate and which refer to conditions holding at an instant of time, we shall automatically have determined the general items of information which define a "state" of our system. The initial state of the system will then be expressed, as usual, by the aggregate of these items of information at an initial instant. It was by following this method that the conditions defining the initial state of a mechanical system were discovered by the founders of mechanics.

The definition of the initial conditions, or initial state, varies with the physical system considered. For example, let us suppose we are dealing with the flow of heat by conduction through an infinitely extended solid. Fourier established the law of heat conduction and ex-

pressed it by means of a differential equation (Fourier's equation). A study of this equation then shows that the initial state of our system is defined by the temperature distribution throughout the solid at the initial instant. In the case of electromagnetic phenomena occurring *in vacuo,* the laws, or equations, were discovered by Maxwell. From the form of these equations it can be proved that the initial distribution of the electric and of the magnetic vectors throughout infinite space defines the initial conditions, or initial state.

But the situation is not always so simple. If electrons are present, Lorentz's electromagnetic equations show that for the future to be determined, not only must we specify the initial values of the electric and magnetic vectors, but that we must also specify the motions of the electrons throughout a certain interval of time. The possibility of defining an initial state thus eludes us, and we may no longer say that the evolution of the system is determined by any initial conditions. This situation seems strange, for it implies that the causal doctrine is no longer valid. More attentive consideration shows, however, that the causal doctrine is not endangered. The fact is that, when electrons are present, Lorentz's electromagnetic equations are incomplete because there are less equations than variables. In more familiar language, the laws are incompletely stated. If these laws were rendered complete, an initial state could be defined and the evolution of the system could be determined without our having to specify the motions of the electrons throughout a certain time. Thus, the requirements of the causal doctrine would be satisfied. Incidentally, we may mention that various attempts were made to complete Lorentz's laws. Such attempts have been abandoned, however, on account of their speculative nature.

A somewhat similar illustration is afforded by the theory of elasticity. In some cases the equations, or laws, require that the entire past history of the elastic body be specified before the future evolution can be gauged. A situation of this sort seems to conflict with the simple conception of causality that we have been discussing, since one of the characteristics of this conception is to assume that the initial state which determines the evolution is an instantaneous one. The difficulty is overcome when we find that the past history of the elastic body may equivalently be expressed by a number of separate items of information at an initial instant.

Tests of the Doctrine of Causality—Our analysis shows that even if our measurements were perfect, an accurate test of the doctrine of causality would require that we experiment on a truly isolated system. Since, however, the only truly isolated system of which we can conceive

is the entire universe and since it would be impossible to experiment on that system, we conclude that in any case an accurate test of the doctrine of causality cannot be performed.

We may, however, be satisfied with experiments on a system of manageable proportions which is approximately isolated. For instance, experiments performed in the laboratory appear unaffected by the changing conditions of the universe at large. But here another difficulty confronts us. In any observation we necessarily introduce extraneous factors into the system (*e.g.*, our measuring instruments), and these disturb the system and destroy its isolation. Thus we may determine the temperature of a liquid by placing a thermometer in it. But unless the thermometer happens to be at the same temperature as the liquid, it will heat or cool the latter, with the result that the temperature finally read will not be the original temperature of the liquid. Furthermore, we know that human measurements are necessarily imperfect, so that even if they did not disturb the magnitudes to be measured, they would still yield only approximate results.

In view of the practical impossibility of testing the doctrine of strict causality by experiment, we must regard this doctrine as a dogma which may be accepted by some and rejected by others. Classical science was well aware of these elementary difficulties, but the contention was that although rigorous causality could not be tested, approximate causality had proved itself to be valid in experiments conducted on approximately isolated systems. The objection that all observations necessarily disturb the magnitudes to be measured and thereby render the readings uncertain was dismissed on the plea that, by exerting sufficient care, we could minimize the effects of these disturbances and, in theory at least, reduce them to a vanishing limit. Thus in the illustration we gave of the liquid and the thermometer, we may minimize the disturbance by taking a large mass of liquid and a thermometer of small heat capacity. In short, in the opinion of classical scientists the impossibility of testing the law of causality was merely of a practical nature; it was due to casual limitations, and no theoretical impossibility was involved. This distinction between a practical and a theoretical impossibility is fundamental in the recent controversies to which the law of causality has given rise.

During the classical period many arguments in support of, and in opposition to, the doctrine of causality were advanced by metaphysicians. The majority of these arguments do not deal with the precise scientific doctrine but with the vague understanding of causality entertained in metaphysics and in our daily activities. According to some, causal connections are the mere expression of regular repetitions. Thus, if B al-

ways follows A, then A is to be regarded as the cause of B. This contention was countered by inquiring whether "day" should be regarded as the cause of "night." But all such arguments and counterarguments have no bearing on the scientific doctrine of causality, for in them the word "cause" is used too loosely. The mere fact that B follows A does not entitle us to view A as the cause of B. To be justified in saying that the same cause generates the same effects, we must include under the appellation "cause" all the items of information that are required to render the future sequence of events determinate. The sophisticated notion of "initial state" was devised for this express purpose; and we have seen that the definition of a state cannot be obtained until the laws controlling the system have been established. It is idle therefore to speak of cause and effect in any given situation until we have agreed on the laws. In the illustration mentioned, "day" and "night" do not define states and for this reason cannot be treated as cause and effect. If we wish to introduce the notion of state, and thereby of cause, we may do so, but we must understand that the state of the earth in relation to the sun is defined by the specification of that half of the earth which faces the sun (this corresponds to the specification of day or night), and also by the specification of the rate of rotation of our planet. This last item of information is essential in the definition of the state. Its importance is readily understood when we note that, if the earth were to rotate more slowly and to turn always the same half towards the sun, day or night would be eternal and would not follow one another.

Newton's Mathematical Expression of Causal Connections—One of Newton's greatest achievements is his discovery of the means of representing a physical law, and hence a causal chain, by a mathematical scheme. A more detailed explanation of Newton's method is given in Chapter XV. For the present we shall be concerned only with the less technical aspects.

In our presentation of the doctrine of causality we have mentioned that the initial conditions, or state, taken in conjunction with the law controlling a system, determine the history of the system. For the purpose of discussion we may suppose that the continuous evolution of the system is represented by a continuous curve. The various points of the curve represent the various states of the system at successive instants of time; the point of departure of the curve denotes the initial state. According to this picture, if we move continuously along the curve, the continuous succession of states is described. Let us select a succession of points A, B, C, D along the curve; the point A being situated at the

point of departure, represents the initial state. We assume these points packed very closely, so that the states they represent follow one another in rapid succession during the evolution of the system.

To this picture we apply the doctrine of causality, which asserts that the initial point, or state, A determines all the others. But, as Newton observed, we may also say that the initial state A determines the contiguous state B; and that this state B, viewed as a new initial state, determines the next state C; and so on. The fact that the point, or state, A determines the state B implies a connection, or a relation, between the states A and B; the same holds for the other couples of states, B and C, and C and D, and so on. Now the relation retains exactly the same mathematical form whether it refer to the first two states, A and B, or to any other of the couples of consecutive states. In other words, the permanent relation does not change from point to point or from instant to instant during the evolution of the system. It is this permanent relation which represents the law controlling the system in Newton's form of representation.*

Suppose we change the initial conditions while retaining the same system. In our simile, the path of the curve which represents the evolution of the system will be changed, but the permanent relation between consecutive points, or states, will not change. We have here the expression of the permanence of the law as contrasted with the contingency of the initial conditions. The law controlling the evolution of a system is thus seen to express a permanent relation between any two successive states that follow each other at an infinitesimal separation in time; it represents a continuous causal chain forming a link between each ante-

* A more precise presentation would require us to add that the relation which constitutes the law is arrived at by a limiting process, *i.e.*, by supposing that the interval of time between the occurrence of the two successive states contemplated becomes infinitesimal.

For example, the states of a mechanical system, at consecutive instants of time $t_0, t_1, t_2 \ldots t_n \ldots$, are defined by the positions and velocities at these instants. Suppose, then, we are dealing with a particle of mass m acted on by a permanent force along a straight line. If we set x and v to measure the position and velocity of the particle at any instant t, the consecutive states are defined by (x_0, v_0), $(x_1, v_1) \ldots (x_n, v_n) \ldots$. The permanent relation to which we have referred in the text is then of the general form

$$\lim_{t_{n+1} \to t_n} m \frac{v_{n+1} - v_n}{t_{n+1} - t_n} = f(x_n),$$

where the function $f(\ )$, which defines the force, is the same for all values of n, and hence is the same at all instants of time.

cedent state and its consequent. The mathematical expression of the relation, and hence of the law, is given by a differential equation. Thus, natural laws (in the form given by Newton) are expressed by differential equations. For this reason, differential equations are the mathematical instruments which express Newton's presentation of the causal doctrine.

Einstein, commenting on Newton's formulation of the causal principle in mechanics, writes:

"In order to give his system mathematical form, Newton had first to discover the concept of the differential coefficient, and to enunciate the Laws of Motion in the form of differential equations—perhaps the greatest intellectual stride that it has ever been granted to any man to make." *

When the law governing the evolution of a system is known and its mathematical transcription, the corresponding differential equation, is obtained, we are in a position to derive a knowledge of the evolution of a system from any given initial state. Thus, if A represents the initial state, the differential equation, by expressing the relationship between A and the next state B, enables us to obtain a knowledge of B. Since we know the state B, a second application of the differential equation yields the state C, and so on. We must remember, however, that these states occur at instants of time that are separated only by infinitesimal intervals. Consequently. if we wish to obtain a knowledge of the states that the system will assume after some finite interval of time, we shall have to repeat the foregoing procedure an infinite number of times in succession. Of course this would be impracticable. The difficulty is overcome by Newton's discovery of the method of integration. Thanks to this powerful method it is always possible, in theory at least, to follow the causal chain over finite intervals. In practice, however, rigorous integration may present insuperable mathematical difficulties, and in such cases methods of approximation must be resorted to.

The Indeterminism of Quantum Mechanics—Today, less than three centuries after Newton made his momentous discovery, doubts are being cast on the validity of the rigorous causal connections of classical science. The attack is due not to the difficulties previously noted (namely, to the impossibility of testing the doctrine in practice), but to totally new discoveries in the subatomic world, where the mysterious quantum phenomena become noticeable.

* James Clerk Maxwell; Cambridge, 1931; p. 69.

We mentioned that the practical impossibility of testing the doctrine of rigorous causality arises from the impossibility of operating on perfectly isolated systems; first of all, because no such systems exist; and secondly, because, even if we grant the existence of such systems, we cannot observe their internal workings without disturbing them and thereby destroying their isolation. In addition, human measurements are necessarily imprecise. These difficulties were not regarded as fatal to the doctrine, because approximately isolated systems could be found, and our measurements could be so refined that they would not perceptibly disturb the magnitudes to be measured. The situation was expressed by the statement that, in practice, rigorous causality could not be tested, but that no theoretical obstacle barred the way to a rigorous test.

The novelty resulting from the discoveries of the quantum theory is that we now have reasons to suspect a definite theoretical impossibility which would render illusory any attempt to test rigorous causal connections. The quantum theorists, under the lead of Born, Heisenberg, Bohr, and Dirac, agree with the classical scientists in recognizing that the practical difficulties of testing rigorous causal connections may be disregarded; but they are adamant in their claim that the recently discovered theoretical impossibility cannot be dismissed so lightly. As may be gathered from this synopsis, the philosophy upheld by the quantum theorists has not arisen from a revised interpretation of classical facts but from the necessity of coordinating a new body of facts unsuspected in former days.

The various aspects of the new philosophy will be considered in greater detail in the latter part of this book. For the present we shall restrict ourselves to the more salient points. The developments of the quantum theory indicate that the "uncertainty relations," discovered by Heisenberg, prohibit us *in principle* from effecting simultaneous accurate measurements of so-called conjugate magnitudes. This very general rule applies in particular to position and momentum, or velocity, to time and energy, to the electric and the magnetic intensities at the same point in an electromagnetic field. Thus, if the position of the centre of a particle is measured with accuracy, the unpredictable disturbance, entailed by the measurement itself, causes the particle's momentum to become vague. The classical contention that, by exercising sufficient care, we might reduce the disturbance indefinitely is here no longer valid, for the essence of the uncertainty relations is that the limit we might hope to attain is not vanishing but is finite. Obviously, if this principle is accepted, the state of a mechanical system, involving as it does a simultaneous knowledge of the positions and momenta of the various masses, cannot be known with accuracy. Consequently, a test of rigorous causality

is impossible in mechanics. The same conclusion may be extended to all departments of physics.

It is important to understand that the limitations imposed by the uncertainty relation are theoretical, and not practical. Thus the uncertainty relations do not interfere with our measuring, as accurately as we choose, the position alone or the momentum alone of a particle; they only prevent us from executing simultaneous measurements with accuracy. Practical difficulties, on the other hand, would render illusory the perfect accuracy of either kind of measurements. According to the quantum theorists, however, practical difficulties may be waived aside exactly as they were in the classical critique of the causal doctrine. We must therefore conclude that accurate simultaneous measurements are ruled out by the uncertainty relations, not because these measurements are *practically* impossible, but because they are *impossible in principle.* The source of this theoretical impossibility, unsuspected by classical science, must be sought in the very nature of things. It is intimately connected with Planck's constant h. The theoretical impossibility would vanish if this constant, the value of which is finite, were to be infinitesimal, as classical science had implicitly assumed.

These summary remarks may help to clarify the points at issue. The problem is quite general and does not affect solely the doctrine of causality. We are called upon to decide whether a concept which cannot be tested or a magnitude which cannot be measured, *in principle* (in contradistinction to *in practice*), should forthwith be classed as meaningless and cease to play any part in a theoretical discussion.* The dilemma is not entirely new, for it arose in the theory of relativity. Thus according to the special theory of relativity, we must assume that, for reasons which have nothing in common with practical experimental difficulties, a velocity through the stagnant ether cannot be detected. Einstein accordingly claimed a velocity of this type to be meaningless, and all theoretical physicists today endorse his stand. Absolute velocity is thereby dismissed because it cannot be observed in principle. On the other hand, only a practical difficulty prevents us from detecting the possible existence of small planets beyond the orbit of Pluto. For this reason we should not class as meaningless the assertion that such planets may exist; at best, we should claim it to be unproved.

In view of Einstein's attitude towards absolute velocity, we might expect him to adopt a similar one with respect to rigorous causality.

* This brief statement does not do full justice to the complexity of the question. The problem will be examined in greater detail in the last chapter.

Yet this he does not do. Einstein and Planck both retain a belief in rigorous causality, extending it even to living matter. But those who have contributed most to the development of the new quantum mechanics resist Einstein's views and insist that rigorous causality is a myth.

Despite appearances to the contrary, the contentions of the new school are not in obvious conflict with the very high measure of approximation with which rigorous causal connections seem to be realized on the commonplace level of experience, *e.g.*, in ordinary mechanics. Analysis shows that the theoretical uncertainty, which prohibits a simultaneous accurate measurement of position and of velocity, is noticeable only when we are dealing with the very minute masses of the subatomic world. With ordinary masses the theoretical uncertainty, though still existing, falls below the practical uncertainties, which are due to the imperfection of human observations, and is completely submerged by the latter. This gradual obliteration of the quantum uncertainties, as we pass to the commonplace level of average masses, is associated with the gradual passage of the quantum theory into classical science. We have here an illustration of Bohr's correspondence principle, to which reference has already been made.

CHAPTER VIII

CONSERVATION

COMMONPLACE observation teaches us that, despite the presence of change in Nature, some properties seem to remain unaffected. For instance, when water is poured from a jar into a glass, the space filled by the water changes in shape but the volume of the water is unchanged. Thus, the volume of the water appears to exhibit a conservative property. From homely examples of this sort may have arisen the belief in the existence of a fundamental space-filling substance which could neither be created nor destroyed, and which would endure unchanged throughout time. In more technical language a substance of this sort is said to be conserved, or to satisfy a law of conservation.

Modern science does not lend support to this primitive belief in conservative substances. In the first place the existence of rigorous laws, and of laws of conservation in particular, is possible only if Nature exhibits a deterministic scheme. But the trend of modern science is to deny strict determinism, so that on this score alone rigorous conservation can have no place in natural philosophy. The suspicion that Nature betrays no deterministic scheme is, however, recent; and for argument's sake we may suppose that the modern view is erroneous and that rigorous laws are valid. If rigorous laws are assumed, it will always be possible to construct by mathematical means physical magnitudes having conservative properties, but these magnitudes may bear little or no resemblance to the intuitive notion of a substance. Thus, in any case there is no justification for asserting in a dogmatic way that a fundamental conservative substance is a necessity. As an illustration of the procedure whereby conservative magnitudes can be constructed when a rigorous law is known, we select the following:

The law of perfect gases for any given mass of gas is

$$pv = AT,$$

where p, v, and T denote the pressure, volume, and absolute temperature respectively, and where A is a constant characteristic of the gas and of the mass of gas considered. By dividing this equation through by T, we see that the product of the pressure and volume, divided by the absolute temperature, always retains the same value. The artificial magnitude

thus built up has the conservative property; and we note that this magnitude has nothing in common with the naive conception of a substance. Our example brings out a further point. The magnitude we have constructed is rigorously conservative only insofar as the law of perfect gases is itself rigorously accurate. And since we can never boast of the perfect accuracy of any empirical law, we can never be certain that a magnitude is strictly conservative.

Usually, no particular names are coined for the conservative magnitudes that can be constructed more or less artificially. In other cases names are invented. The "total energy" of a so-called conservative mechanical system is an illustration of a conservative magnitude which is built up artificially and which has received a name. The total energy is the sum of the kinetic and potential energies, which individually are not conserved. "Magnetic induction" is another illustration of an artificially constructed conservative entity. A general class of manufactured magnitudes which are conserved is encountered in classical mechanics. Such quantities are known as the "first integrals" of the motion. The total mechanical energy of a conservative dynamical system is precisely one of these first integrals; and so are the total momentum and total angular momentum of an isolated mechanical system.

The illustrations we have given are of a sophisticated kind in that they assume the prior discovery of natural laws. But we may also examine the concept of conservation more directly. As an example, let us imagine a hypothetical investigator bent on discovering by direct measurement some magnitude which is conserved. His first idea might be to view the volume of a liquid as unalterable. He would soon find, however, that a liquid was not rigorously incompressible and that its volume could be decreased by pressure. "Volume," therefore, could not be regarded as a conservative property. Subsequently he might notice that, insofar as his measurements could detect, the weight of the liquid remained unchanged whether the liquid were compressed or not. He would then claim that a conservative magnitude was manifested in "weight." But Newton's theory of gravitation would dispel this belief, for it would require that the weight should decrease considerably if the volume of liquid were transported to the moon's surface. According to Newtonian mechanics, the correct conservative magnitude is "mass." Mass is a more abstract concept than weight since it is the quotient of weight by acceleration. Then comes the special theory of relativity according to which mass cannot be conserved, for the total mass of an electron and proton, when these two corpuscles are packed closely, is less than the sum of the two separate masses. The theory of relativity furnishes, however, a conserva-

tive entity—energy. Our hypothetical investigator would thus conclude that energy is the entity which is rigorously conserved. Incidentally, in spite of its wide use in everyday discussion, the word energy in science refers to a concept still more abstract than mass; and if we except its conservative properties, it bears no resemblance to a space-filling substance.

We may now inquire whether the conservation of energy is rigorous or whether energy, like mass, weight, and volume, may not have to be rejected in our search for a truly conservative magnitude. As things stand today, it is difficult to answer the question with assurance. Bohr at one time rejected the rigorous law of the conservation of energy, viewing conservation as a statistical manifestation. The experiments of Bothe in connection with the Compton effect led Bohr subsequently to revise his views. His new attitude, which is widely accepted, is embodied in his "Principle of Complementarity," a principle that is intimately connected with Heisenberg's "uncertainty relations." Bohr's principle does not deny the conservation of energy (and of momentum) in certain cases, but in such cases it deprives conservation of its intuitive appeal, for it renders impossible the space-localization of the energy which is conserved. On the other hand, if the energy and momentum are located with any degree of accuracy, the principles of conservation cease to be rigorous. Besides, as we mentioned at the beginning of the chapter, if rigorous causal relations do not exist in Nature (and this is the claim made by the quantum theorists), the strict conservation of any magnitude, be it energy or anything else, appears impossible.

CHAPTER IX

MECHANISTIC THEORIES

MECHANICAL phenomena were the first physical phenomena in which supposedly rigorous laws were detected. Mechanics thus appeared as the stronghold of rigorous determinism, and it stood in marked contrast to teleological doctrines and to those in which free will seemed to play an important part. Owing to these peculiarities a mechanistic interpretation of a phenomenon was construed to mean one in which rigorous determinism was assured. Indeed, with some writers, the words "mechanism" and "determinism" were often treated as practically synonymous.

But in view of the subsequent development of science, this confusion of words is apt to be misleading. In the first place, rigorous determinism is no more characteristic of classical mechanics than of many other departments of physical science, *e.g.*, field theories or thermodynamics. Furthermore, the laws of mechanics may, if we choose, be expressed in a form which smacks of teleology.* Finally, recent discoveries in quantum mechanics have prompted some of the leading physicists to question whether the true mechanics of Nature is in any sense deterministic. Our first concern will therefore be to define the meanings of the words we propose to use.

By a mechanistic theory, we shall mean one in which the only physical magnitudes considered are those which are characteristic of mechanics, *e.g.*, force, momentum, mass, strains, and stresses. A theory which is not restricted exclusively to magnitudes of this sort will be called non-mechanistic. The non-mechanistic theories comprise the field theories and phenomenological theories, such as thermodynamics. These other kinds of theories will be discussed in Chapters X and XI. For the present we shall be concerned with mechanistic theories.

Mechanistic theories may be classed into two main groups according to whether they deal with discrete or with continuous media. In the first group we have the atomistic theories (*e.g.*, the kinetic theory of gases). In the second group we may mention the theory of elasticity, and Fresnel's theory of the continuous and elastic ether. A noteworthy fact is

* See Chapter XVIII.

that mechanistic theories preceded field theories and that the mechanistic theories of the corpuscular type preceded the continuous theories. As we shall see, this historical order may be attributed to a number of causes, some of them physical and others mathematical.

Greek Atomism—Even in the remotest civilizations men were presumably impressed by the difference between substances like water, which appear continuous, and others such as sand, which are discrete. Presumably also, some early thinker may have wondered whether, in spite of its continuous appearance, water was not in reality formed of microscopic corpuscles. Fragments from the earlier Greek writers indicate that problems of this sort formed the starting point of the doctrine of atomism. Democritus is usually credited with the invention of this doctrine; more probably, however, the atomic hypothesis arose in remoter days among the Chaldeans and the Egyptians; it is not sufficiently original to be credited to any one man in particular. Inasmuch as Democritus's writings have been lost, we must turn to the writings of Lucretius for the most exhaustive presentation of the doctrine.

The major aim of atomism is to interpret the qualitative differences, which impress our senses, by means of hidden occurrences of a more or less uniform nature. Atomism thus presupposes that the real world is not the world of our sensations, but is a hidden world which we perceive only indirectly through our senses. In this respect atomism is in agreement with modern science, for science also assumes that a sensation such as "green" is due to a certain number of vibrations taking place every second in the electromagnetic field. In Greek atomism the hidden occurrences postulated are the motions of atoms through the void. The atoms are impenetrable, unchangeable, and eternal; they have different shapes and sizes, the tiniest atoms of all being the atoms of fire. The motions of the atoms produce the subjective sensations of warmth, light, smell, and they are the cause of all the manifestations of change in the universe. Solid and semi-permanent bodies, such as stones, are mere agglomerations of atoms. Beyond the atoms there is no further mystery, for by hypothesis the atoms are the ultimate building blocks of the universe; and it is therefore useless to inquire into the secrets of their inner parts. Thus, all further mystery is banished *a priori*.

But this simple scheme affords no means of explaining the existence of semi-permanent bodies. For, if the atoms be moving about at random, how can we ever expect them to hold together? The difficulty was removed by assuming the atoms to have hooks and eyes, thanks to which they could interlock. Thus, the coherence of matter was accounted for.

Like all primitive people, the Greeks experienced an uncontrollable desire to compress all knowledge into a single formula. Whereas today a man who would claim omniscience would be regarded as a charlatan, the wise men among the ancients boasted of their ability to answer all questions. Small wonder, then, that the simplified world picture presented in the hypothesis of atomism should have appealed so strongly to many of the Greeks. In our opinion the sole merit of atomism is that it interprets quality by means of quantity, and by so doing enables the human mind to reason with greater precision.

As usually occurs when a doctrine becomes popular, the inevitable metaphysician appeared upon the scene. He proceeded to explain that the entire doctrine was a necessary consequence of some of his tenets. Atomism in particular was, according to him, a necessary consequence of the presence of change and of permanency in Nature. Permanency was illustrated by the indestructible eternal atoms, and change by the varying configurations of the atomic aggregate. We need scarcely point out that any conservative property postulated for any changing configuration would be just as suitable in accounting for permanency and change as is the atomic hypothesis; but perhaps one should not be too critical of the naivete of these early speculations, for they were advanced practically at the dawn of thought. There is less excuse, however, for those who defend them today.

If we purge Greek atomism of its unnecessary metaphysical trappings, it has more to commend it; but even so, it exhibits only a superficial similarity with the corpuscular theories of science. In the first place, scientific atomism does not profess to furnish a universal interpretation of all things. Originally, it was put forward as a hypothesis so as to account for certain restricted empirical facts, notably for some of the properties of matter. Other existents were not regarded as atomistically constituted. Electricity and heat, for instance, were at one time supposed to be continuous fluids. At a later date the continuous electromagnetic field was introduced. At no time was the force of gravitation interpreted in terms of chains of hooked atoms extending from the sun to a planet. Instead, the force of gravitation was postulated as an independent hypothesis.

A further difference between ancient and modern atomism becomes apparent when we recall that the unalterable, eternal existence of the atoms, which the Greeks assumed essential to their ideas, is precisely one of the beliefs that is rejected by modern science. The closest approach to the Greek atom is illustrated by the electron or the proton. Yet these two existents are so far from being deemed unalterable and eternal that

physicists have assumed they may coalesce, give rise to radiation and hence to something totally different. With the discovery of wave mechanics, it is doubtful whether even matter can be interpreted solely in terms of corpuscles and of electromagnetic fields; wave magnitudes of a mysterious nature also appear to be required.

Having noted the differences between the two varieties of atomism, we may examine the reasons which gave rise to atomism in science. Matter was the first existent to suggest itself as being formed of corpuscles. Thus, if we mix a pint of water with a pint of alcohol, we obtain less than two pints of the mixture. The phenomenon is easily accounted for if we assume the water and the alcohol to be granular; the finer water corpuscles may then squeeze in between the coarser corpuscles of the alcohol. Similarly, when subjected to sufficient pressure, water can be made to ooze through a sheet of gold; the porousness thereby revealed by the gold suggests a granular structure. Finally, the compressibility of gases and the basic laws of chemistry furnished additional evidence of the atomicity of matter.

The first physical theory of an atomistic nature was not, however, a theory of matter, but a theory of light. Newton believed that the phenomenon of double refraction could more easily be accounted for by supposing that light was formed of corpuscles instead of being a wave manifestation, as was suggested by Huyghens. But Newton's theory was more in the nature of a sketch than a well-developed mathematical theory; and when the theory of light was subsequently constructed in the nineteenth century, Huyghens's assumption and not Newton's was accepted.

Mechanistic theories of a corpuscular type attained a considerable degree of development when phenomena dealing with ordinary matter were studied. We have mentioned the empirical evidence that suggested the corpuscular structure of matter, but, as we shall now see, other reasons relating to Newton's theory of the solar system conspired to give priority to corpuscular theories.

The study of moving solids was one of the most immediate which forced itself on the attention of men, and for this reason the laws of dynamics were formulated before those of electromagnetics and thermodynamics, for example. Newtonian dynamics was thus the first physical doctrine to be established on a secure mathematical basis. When Newton's attention was directed to the planetary motions, he assumed the planets and sun to be so many material bodies moving in accordance with the same mechanical laws that hold on the earth's surface. Even if we grant that this assumption was not inevitable, it was at all events plausi-

ble, and it has never been challenged by subsequent scientific developments. If this assumption is made, the interpretation of the planetary motions necessarily becomes a mechanistic one. Furthermore, the planets and sun are spherical bodies which we may suppose to be approximately homogeneous; and Newton proved that, from the standpoint of gravitational action, a spherical homogeneous mass is equivalent to a point-mass having the same total mass. Hence, in Newton's theoretical treatment the sun and the planets may be replaced by so many point-masses. Thus, the first grand theory of mathematical physics was, to all intents and purposes, a mechanistic theory involving the motions of point-masses. Let us observe that the mechanistic nature of Newton's theory was not prompted by any particular preference for mechanistic interpretations. As for its association with point-masses, it was obviously imposed by the very configuration of the solar system and was quite foreign to any atomic hypothesis.

The following century (the eighteenth) witnessed an intensive development of pure mathematics, thanks to which a systematic attack was made on many of the problems of celestial mechanics left unsolved by Newton. Little was achieved in the other departments of theoretical physics, chiefly for the reason that very few new experimental facts were established. Certain properties of matter, such as adhesion, cohesion, and the phenomenon of capillarity, were, however, investigated by the various experimenters, and, on the basis of their findings, attempts were made to construct mathematical theories. Owing to the empirical evidence, a corpuscular structure was credited to matter in these theories. Newton's treatment of the planetary motions served as a model for the new theories, the molecules of matter being likened to point-masses controlled by the accepted laws of mechanics. It is true that conditions were considerably more complicated than they were in the solar system. In particular, ordinary gravitational forces between molecules proved insufficient to account for the cohesion of matter, so that other laws of attraction differing from that of the inverse square were introduced. But from our present standpoint such changes are of minor importance, because in any case the new theories were necessarily mechanistic and corpuscular.

Among the most celebrated of these mechanistic theories of the eighteenth century is the one elaborated by Laplace in connection with the phenomenon of capillarity. Also worthy of mention is Bernoulli's suggestion of the kinetic theory of gases—a theory according to which a gas is formed of perfectly elastic molecules in rapid motion, colliding and rebounding in conformity with the laws of mechanics. The kinetic

theory was developed in the following century by Maxwell and Boltzmann; it was also extended to liquids, but with less success. In connection with these kinetic theories we may note that the corpuscles of matter, which constitute the ultimate elements on which the theories are based, are not assumed to be ultimate in any absolute sense. The molecules of the gas are known to be divisible into atoms, and these atoms in turn may be disrupted into smaller components. Nevertheless, since the molecules behave as though they were permanent units in the phenomena contemplated by the kinetic theory, they may be treated as such in the elaboration of the theory. Atomism, from this standpoint, implies merely a high degree of stability; nothing further is intended.

The next theory of corpuscular physics to appear was formulated so as to coordinate certain electrical phenomena. In the early years of the nineteenth century, the electric current that flows in a wire was thought to be a continuous fluid in motion, and not until Faraday established the laws of electrolysis was the corpuscular nature of electricity suspected. This suspicion received added support when cathode rays were discovered (these are produced by causing an electric current to pass through a highly rarefied gas). Hertz, in his last experiments, proposed to determine whether the cathode rays were streams of swiftly moving particles or wave manifestations. Subsequently the corpuscular view triumphed, the cathode particles being called beta-particles, or electrons. Since then, numerous experiments have established the corpuscular nature of electricity, though we may add that in view of the latest developments in the quantum theory, the concept of corpuscles is far from being as clear cut as was formerly believed.

The first mathematical theory of electric phenomena viewed from a corpuscular standpoint is Lorentz's theory of electrons. When we say that this theory is of the corpuscular variety, our statement is not quite correct, for Lorentz retains the continuous electromagnetic field side by side with his corpuscular electrons. The motives which led Lorentz to introduce electrons arose from the failure of Maxwell's continuous theory to account for several well-known physical phenomena, notably for dispersion. These considerations are clarified by the following short digression on Maxwell's theory.

Maxwell's equations determine the possible distributions and variations of the electromagnetic field. If the region of space in which the field is situated is a vacuum, Maxwell's equations which control the field are called "the equations of electromagnetics *in vacuo*." Maxwell also considered the case where the field was present in the interior of matter.

He supposed that, in a transparent medium such as glass (assumed homogeneous and isotropic), the field would be modified, and that the modification could be expressed by supposing that the space was characterized by an appropriate numerical constant K the value of which depended on the nature of the matter. Space empty of matter was characterized by the value unity attributed to the constant K, but in the interior of matter K exceeded unity. Maxwell's theory leads to the identification of waves of light with electromagnetic waves. According to his theory, the velocity of propagation of electromagnetic waves (and hence of light waves) in a homogeneous medium has the value $\frac{c}{\sqrt{K}}$ where c is the velocity of light *in vacuo* and K characterizes the medium. Since the value of K exceeds unity inside matter, we conclude that waves of light are propagated with reduced speed in the interior of matter. The velocity of light will consequently suffer a drop when the light passes from a vacuum (or air) into glass. Now the wave theory shows that this decrease in velocity will cause a ray of light to be bent when it penetrates from air into glass under a slanting incidence. In more technical terms, the ray of light will be refracted. Furthermore, Maxwell's theory suggests various experimental means of measuring the value of the constant K for glass, and thereby enables us to predict the precise degree of bending a ray of light will sustain. Suffice it to say, the predictions drawn from Maxwell's theory are in general agreement with the well-known phenomenon of refraction.

Though Maxwell's theory accounts for refraction, it is unable to predict the important phenomenon of dispersion, according to which rays of light of different color are not bent in the same degree (violet light being refracted more strongly than red light). It is to dispersion that the appearance of the rainbow is due. The reason Maxwell's theory fails to account for dispersion is easily understood. Thus we have seen that the bending of a ray of light falling on a slab of glass under a given incidence is governed by the value of the constant K for glass. Inasmuch as, in Maxwell's theory, this constant is determined solely by the glass and not by the color of the light, all rays of light should be bent in exactly the same degree, and dispersion should not arise. To account for dispersion, we must assume that K is characteristic not only of the matter, *e.g.*, the glass, but that it must also vary according to the color of the light. In short, Maxwell's theory in the presence of matter is seen to be but an approximation. This realization prompted Lorentz's refinement.

At the time Lorentz was initiating his theory, the empirical evidence suggested the existence of corpuscles of electricity, and so Lorentz availed himself of this discovery in order to remedy the difficulty connected with the constant K. He sought to give a physical interpretation of the dependence of K on the color of the incident light by supposing that, in the interior of transparent matter, electrons were distributed and were free to vibrate about fixed positions. Under the exciting influence of the incident light, these electrons would be set into vibration with the same frequency as that of the incident light. In this way the required dependence was established mathematically between the characteristics of the glass and the frequency of the light. Dispersion was thus explained by refining Maxwell's oversimplified treatment of matter. Dispersion, however, was not the only phenomenon that Maxwell's theory had failed to interpret; magnetic rotary polarization was another. But here, also, the electron theory cleared up the major difficulties.

Lorentz's theory readily accounts for the difference between conductors and non-conductors. Thus in dielectrics (*i.e.,* non-conductors), such as glass, the electrons cannot move appreciably from their fixed positions, but in conductors, such as metals, the electrons are free to move throughout the substance. If, then, we create a difference in potential in a metal, all the electrons will swarm in the same direction, and an electric current will result.

Mechanistic Theories Based on Continuity—We have seen why mechanistic theories were the first to be considered, and we have also seen that, in view of the corpuscular structure credited to matter, these theories were of the atomistic type. In the course of the nineteenth century, mechanistic theories of a different kind were constructed. In these, matter was treated as though it were continuous. The reasons that led to the formulation of the continuous theories may be understood from the following considerations.

If we accept the corpuscular theory of matter, we must assume that the water in a glass is formed by the agglomeration of myriads of tiny molecules acting on one another in accordance with more or less complicated laws. But in practice, whether the corpuscular theory be valid or not, water in bulk behaves like a continuous fluid and we may treat it accordingly. In the early theories of mathematical physics, where the corpuscular theory was adhered to, the rigorous treatment of problems dealing with matter required that we consider the motion of each individual molecule. As may well be imagined, the rigorous procedure was

impossible in practice, and recourse was had to average effects and to approximations. But then, since the corpuscular theory did not hold its promise of leading to rigorous predictions and since it had the disadvantage of involving complicated mathematics, it seemed simpler to disregard the corpuscular theory entirely and to proceed as we should in ordinary life by viewing matter as continuous. Considerations of this sort led to the development of mechanistic theories of the continuous type.

A noteworthy fact is that the mechanistic theories of the continuous type began to appear only in the first half of the nineteenth century. In view of the plausible reasons that prompted their construction, we might be surprised at the tardiness of their arrival. The reason must be sought in the mathematical limitations of the day. The mathematical treatment of corpuscular media involves the application of differential equations of the simpler type studied by Newton and by the mathematicians of the eighteenth century. On the other hand, the treatment of continuous deformable media requires the application of a more complicated kind of differential equation, the so-called "partial differential equation." The mathematicians of the eighteenth century, notably Euler and Lagrange, had made some investigations on partial differential equations, but only in the following century, thanks to Cauchy, Poisson, Jacobi, and others, were the more important properties of these equations established. It then became possible to study mathematically the statics and dynamics of continuous deformable media, and the mathematical theory of elasticity was founded. In this way the first mechanistic theory of the continuous type arose.

Contemporaneously with the mathematical study of elastic continuous media, the early nineteenth century witnessed the rise of the undulatory theory of light. Experiments performed by Young and by Fresnel suggested that light consisted of waves propagated through space, much as elastic waves are propagated through an elastic medium. In many respects the analogy appeared so striking that Fresnel postulated a space-filling elastic medium, the ether, the deformations of which were controlled by mechanical laws and in which waves could be propagated. Fresnel's assumption constitutes the hypothesis of the elastic ether. Physicists realized that certain properties of Fresnel's elastic ether would have to differ considerably from those of material bodies, but since it was unnecessary to suppose that the ether was matter in the ordinary sense of the word, the features of dissimilarity did not appear alarming. It seemed gratuitous and unnecessary to make the assumption that this postulated ether was corpuscular in structure; the ether was therefore

regarded as a continuum. The mathematics required by Fresnel's theory was at the disposal of investigators thanks to the prior study of partial differential equations and of elastic media. Fresnel's theory was thus mechanistic but based on an underlying continuous medium.

Maxwell's theory, to which reference has already been made, seemed at first to be of the same general kind as Fresnel's. The electromagnetic phenomena studied by Maxwell were regarded as manifestations of mechanical changes occurring in the continuous ether. But, for reasons which will be understood in the next chapter, mechanical attributes were eventually removed from the ether, and mechanical interpretations of electromagnetic phenomena abandoned. It was the failure of mechanical interpretations to account for electromagnetic phenomena that led to the development of the field theories. With these we shall be concerned in the next chapter.

CHAPTER X

FIELD THEORIES

IN PHYSICS the world *field* is given to the continuous distribution of some "condition" prevailing throughout a continuum. We are using the vague word "condition" purposely, because the precise nature of the magnitude, the distribution of which constitutes the field, varies according to the problem at issue.

When the condition is adequately described at each point of space by a simple number (*i.e.*, a scalar), we have what is known as a scalar field. Temperature is such a condition, and so the temperature distribution throughout a volume is a physical illustration of a scalar field. In many cases the condition at each point of space has a direction as well as a magnitude; in this case an arrow, or vector, is required to represent it. The field is then called a vector field. A field of force or the field defined by the instantaneous velocities of the various points of a fluid in motion are illustrations of vector fields. Finally, if the condition is described by a tensor at each point, we have a tensor field. The distribution of the stresses in an elastic medium is of this type; and so also is the field of gravitation in the 4-dimensional space-time of the general theory of relativity. Whatever its nature, the field is called "permanent" or "variable" according to whether the distribution of the condition remains unchanged or varies with time. We might also consider a discrete field, for instance the field defined by the velocity vectors of the various particles of sand in a sandstorm. In a field of this sort each point at which a particle of sand is situated, at the instant considered, is associated with a velocity vector defining the velocity of the particle; the intervening points at which no sand particles are present have no vectors attached to them. It must be emphasized, however, that discrete fields are seldom considered. Usually, when we speak of a field, a continuous one is understood; and the fields we shall consider in the present chapter will always be continuous.

The physical condition, whose continuous distribution constitutes the field, differs from one situation to another. In some cases it may be interpreted by means of mechanical categories, such as force, velocity, acceleration, displacement, strain or stress. The field is then of a mechan-

istic nature. Sometimes, however, mechanical categories cannot account for the condition. In this event, the field is non-mechanical.

We must now examine what is meant by a field theory. A field theory is a theory whose aim is to study the peculiar condition, or field, which is thought to pervade the ether of space in certain cases. Formerly, the ether was viewed as an elastic medium, having many of the properties of matter, and so a field theory did not appear to differ essentially from a mechanistic theory of the continuous type, such as the theory of elasticity. But according to modern views, the two kinds of theory are totally different, for the ether is now assumed to have no mechanical properties, and hence its field is sharply distinguished from the field of mechanical stresses which pervades an elastic body.

The withdrawal of all mechanical properties from the ether has caused some physicists to regard the word "ether" as merely another name for empty space. According to this conception the field of the ether is a field which is due to the properties of space itself (or space-time); it is thus the field of space, and so space must be regarded as an active agent instead of a passive void.

The point to be kept in mind for the present is that field theories differ from mechanistic theories of the continuous type since they do not involve the mechanical categories; and that they differ still more from the mechanistic theories of the discrete type (such as the kinetic theory of gases) because of the additional contrast between the continuity of the field and the discreteness of matter viewed atomistically.

Several considerations account for the relatively recent appearance of field theories. First of all, fields, since they are continuous, require for their mathematical formulation the introduction of partial differential equations. Accordingly, field theories (as also mechanistic theories of the continuous type) could not have been constructed until the early part of the nineteenth century, when the properties of partial differential equations began to be understood. But even after the purely mathematical difficulties were overcome and continuous distributions of a physical condition could be studied mathematically, field theories did not suggest themselves immediately. The fact is that the first kind of continuous distribution which attracted the attention of physicists was the continuous distribution of stresses in an elastic medium; and these stresses are obviously mechanical magnitudes. Only when the electromagnetic field was discovered by Maxwell, did physicists suspect that a new, non-mechanical physical existent had been revealed. Maxwell's discoveries could not have arisen at a much earlier date, for Maxwell was guided by many of the experimental and theoretical results established by

the pioneers in electrical research. Thus, considerations of an experimental and of a mathematical nature both conspired to prevent the development of field theories in the seventeenth and eighteenth centuries.

But we might be tempted to inquire why a field theory was not developed in connection with the gravitational field. The very appellation "gravitational field" would seem to imply a belief in the existence of a field. In point of fact, however, prior to Einstein, a gravitational field was not meant to suggest a condition of space or of the ether. When physicists spoke of a gravitational field surrounding matter, they meant that a mechanical field of force surrounded the matter; the space itself around the matter was assumed to be inert and to have no direct connection with the field. For this reason Newton's theory of gravitation was not a field theory. These considerations will be better understood when we examine some of the peculiarities of field theories.

Let us imagine that a massive body, such as the sun, suddenly comes into existence. A planet situated in its vicinity is then subjected to a gravitational pull. We wish to understand how the gravitational action of the sun may have been transmitted to the planet. One way of accounting for the phenomenon is suggested by commonplace observation. Thus let two pieces of cork be floating on the tranquil surface of a pond. We move one of the corks up and down; a train of expanding circular waves is generated; and when these waves reach the second cork, it is set into motion. There is no direct influence of one cork on the other; the transmission of the action occurs from place to place through the intervening water, and a certain period of time necessarily elapses before the action is communicated from one cork to the other. Any transmission of the type just discussed is called a transmission by "contact action."

We now attempt to interpret gravitational action along the same lines. We must suppose that in our hypothetical illustration the sudden birth of the sun disturbs the condition of surrounding space (or of the medium filling space). The disturbance is then propagated in expanding spherical waves by contact action until it reaches the planet. At this instant the attractive force manifests itself on the planet. In our present analysis we need not be concerned with the precise nature of the disturbance sustained by the intervening medium, whether it be a mechanical compression or anything else; we may therefore refer to the disturbance by the vague name "gravitational condition." The intensity of this condition will vary from place to place in the space surrounding the sun, and the distribution of this condition defines what we have called a field. The attitude here tentatively adopted leads to a field interpretation of gravitation: for the sun does not act directly on the planet; it merely generates

a change in the condition of the ether, and it is this condition of the ether which betrays itself in a gravitational pull. The ether, or space, surrounding the sun thus plays an active part in the phenomenon. We also note that, according to the foregoing field interpretation, a certain period of time must elapse between the instant at which the sun comes into existence and the instant at which the gravitational pull is experienced by the planet. Attempts were once made to view gravitation as a field manifestation; and experiments were devised to illustrate the situation. We refer to the experiment of Bjerkness's pulsating spheres. Two rubber spheres which contract and expand rhythmically are placed in water. Theory and experiment show that the spheres act as though attracted to each other. Such mechanistic interpretations of gravitation were soon abandoned, however.

The problem of gravitation may be approached in a different way. We may suppose that the action of the sun on the planet is direct and that the intervening medium remains passive, playing no part in the phenomenon. This interpretation compels us to assume that the transmission of the action is instantaneous, so that the planet is submitted to the gravitational pull at the very instant the sun comes into existence. Herein consists the hypothesis of "action at a distance." It is incompatible with a field theory.

A third interpretation, which differs essentially from the former two, is the one advanced by Lesage. According to Lesage, gravitational attraction is due to the incessant bombardment of particles moving through space in all directions. Two bodies which would normally be said to be attracting each other are in reality inert, but as they exert a mutual screening, the resultant effect of the incessant bombardment is to draw them together. Lesage's hypothesis was subsequently investigated and found to be untenable. We mention it because it illustrates the possibility of avoiding "action at a distance" without invoking the activity of the intervening medium—the latter remains passive.

A decision between the hypotheses of action at a distance and of contact action is not an easy one if we are to be guided solely by experimental tests. We cannot, of course, bring a large mass into existence suddenly (or annihilate it suddenly) and then measure the time taken by the gravitational force to manifest itself (or to cease) at a distant point. Newton himself was averse to action at a distance, and at one time invoked the pressure of the surrounding medium as a vehicle for the transmission of the gravitational action. But in any case the mathematical difficulties alone would have prevented him from developing a field theory, and in addition the hypothetical nature of the assumptions

which he would have been compelled to make did not satisfy his cautious nature. His celebrated dictum "Hypotheses non fingo" may be interpreted, in our opinion, by the statement: If we cannot furnish an interpretation which is in agreement with a large number of facts, it is better to give none at all. The net result was that Newton's theory of gravitation, in its actual application, assumed action at a distance and was not a field theory.

We said that a direct test of the opposing interpretations is exceedingly difficult. Calculation does, however, furnish some information. Thus in the solar system, a given planet, say the earth, is submitted to the gravitational actions of the sun and of all the other planets. If the gravitational action is instantaneous, the force exerted by the planet Jupiter on the earth at a given instant is directed toward the *present* position of Jupiter. If, on the other hand, a certain time, say half an hour, is required for the gravitational attraction to be propagated, the earth will be attracted towards the position which Jupiter occupied half an hour earlier. The motion of the earth cannot therefore be quite the same in the two cases.

Laplace submitted the problem to a test by assuming a delay in the transmission of the gravitational force. His aim was to determine how great a delay was compatible with the observed planetary motions. As in all problems of celestial mechanics where more than two bodies are involved, exact calculations are hopelessly difficult, and so Laplace resorted to approximations. These he obtained by omitting to consider influences which he assumed to be of relatively small importance. Unfortunately, the effects he neglected were subsequently shown by Poincaré to be as important as those he had retained, and so Laplace came to the erroneous conclusion that the force of gravitation was transmitted instantaneously or, at any rate, with a velocity far greater than that of light. Laplace's results were accepted in his day and were regarded as affording an important confirmation of "action at a distance." Any attempt to develop a field theory for gravitation was therefore deemed useless.

In the study of electromagnetic phenomena, the problem of deciding whether to adopt a theory of action at a distance or a field theory came up for consideration as it did in the case of gravitation. The first empirical law established was Coulomb's law of electrostatic actions, which states that opposite electric charges attract each other in accordance with the Newtonian law of the inverse square, and that charges of the same sign repel according to the same law. The law of the inverse square also holds for magnetic poles. Possibly by reason of the similarity between

these laws and the law of gravitation, but more probably because as yet no field theories had been constructed, the hypothesis of action at a distance was retained. Likewise, the action of currents on currents and on magnets was viewed as an instance of action at a distance.

Faraday and Maxwell were the first to suspect that the seat of electromagnetic phenomena should be sought in the surrounding medium. The next step was therefore to construct a field theory of electromagnetism. But first let us recall what consequences a field theory will entail.

Suppose that a current is made to pass through a straight wire. We know that a magnetic needle placed in the neighborhood of the current is acted upon by a force and sets itself crosswise to the direction of the current. If we accept a field theory, we must assume that the needle is subjected to the deflecting force, not instantaneously, but only some time after the current has started. Similarly, if the current is stopped suddenly, the needle continues to be deflected until the disturbance has passed it; and this will necessarily be some time after the current itself has ceased. The disturbance, or field, then spreads out into space, decreasing in intensity as it proceeds. A successful field theory of electromagnetism must be in agreement with the foregoing conclusions. Maxwell took into consideration the empirical laws already known, and, by combining them with an assumption of his own, constructed his celebrated field equations of electromagnetics.

In Maxwell's theory, the condition of the ether is determined at each point by two magnitudes: the electric intensity (a vector), and the magnetic intensity (a tensor). The distribution of these two magnitudes at an instant determines the electric and the magnetic fields respectively. The two fields are not independent, for any variation in the one entails a change in the other. Maxwell's equations express the interconnection between the distributions and variations of the two fields. The intimate connection between the electric and the magnetic fields suggests that we view them as forming one single field, the electromagnetic field. Maxwell's equations may therefore be said to express the mathematical laws controlling the changes of the electromagnetic field. As is well known, a direct consequence of Maxwell's theory is that an electromagnetic disturbance should be propagated *in vacuo* with the velocity of light. Hertz, by submitting this important prediction to a successful experimental test, vindicated the theory and secured its acceptance.

There is considerable similarity between the condition of the ether when an electromagnetic field is present and that holding in an elastic solid submitted to deformations. The similarity is still more striking when the evolution of the field assumes the form of waves—a situation

which is realized when we cause a current to oscillate back and forth along a straight wire, as in a radio antenna. But the similarity is superficial, for on closer analysis the mechanical interpretations prove cumbersome and often contradictory. Maxwell therefore abandoned his initial attempts at giving a mechanical interpretation of the field. His theory thus assumed the aspect of a field theory in the modern sense. The condition of the ether betrayed by the field was a physical condition, but it could not be interpreted in terms of the mechanical categories; and so physicists came to regard the electromagnetic field as a new type of physical reality.

In Lorentz's theory of electromagnetics, this non-mechanical attitude was stressed, and all mechanical interpretations of the field were ruled out on principle. The Lorentzian ether was rigid, indeformable, and so it was taken to define the reference frame of absolute rest. Lorentz's ether is thus but another name for Newton's absolute space. These attributes of the ether automatically exclude the possibility of our viewing the field as due to displacements or to stresses or indeed to modifications of any type that is familiar. So long as we are dealing with a region of space free from matter, the field is the only physical existent to be considered, but in many cases matter and ordinary electric currents may also be present. For the reasons mentioned in the preceding chapter, Lorentz assumed that electricity was corpuscular and that matter was built up from the electrons and protons. The corpuscular electrons are, to all intents and purposes, matter, for they have mass and their motions are controlled by the ordinary laws of mechanics. The word "electricity" refers to the electrons rather than to the field; and the electric current, which is due to the rushing of electrons along a wire, is regarded as something entirely distinct from the field and its changes. Lorentz's theory thus endorses a dualistic conception. On the one hand, we have the continuous non-mechanical field, represented by physical magnitudes existing on their own account. On the other, we have matter, represented by the corpuscular electrons and protons, the motions of which are controlled by the laws of mechanics. The field and the electrons are not altogether independent, for the presence and the motions of the electrons affect the field; and conversely, the field, even though it is non-mechanical, develops forces on the electrons. The mechanical force exerted by the field is called by Lorentz the "ponderomotive force"; its mathematical expression is obtained by combining in an appropriate way the electric and magnetic magnitudes of the field. The ponderomotive force therefore acts as a link between the field and the mechanical electrons.

The dualistic aspect of Lorentz's theory was distasteful to those monistically inclined, and attempts were made to secure a fusion of the

mechanical and the electromagnetic categories. In the original attempts the field magnitudes were treated as fundamental and the mechanical ones as derived therefrom. According to this program the corpuscular electrons and protons were to be built up from field magnitudes alone, and quite generally the mechanical laws were held to be consequences of the laws of electromagnetics. Any theory which professes to interpret all the categories in terms of non-mechanical field magnitudes is called a "pure field theory." Consequently, the attempts we are here mentioning represented an endeavour to transform Lorentz's dualistic field theory into a pure field theory.

Abraham, to whom we owe the first attempt of this kind, concentrated on the mechanical concept of mass. He assumed that an electron was a rigid structure and he showed that the entire mass of an electron in motion could be built up from the field magnitudes. Mass, according to this view, was fundamentally an electromagnetic manifestation. Since it was the electron's motion through the field which generated its mass, an electron at rest should have no mass. Lorentz himself proceeded along different lines. For reasons connected with certain experiments in electrodynamics, he postulated the so-called FitzGerald contraction for bodies moving through the ether. Thus an electron, which was assumed to be spherical when at rest, became flattened when in motion. In this assumption consists the hypothesis of "the contractile electron." When the problem of mass was investigated for the contractile electron, calculation showed that though a part of the mass was electromagnetic in nature, there remained a residue which could not be interpreted in terms of the field. A remnant of mechanical mass was therefore inevitable.

Experiments were soon devised for the purpose of deciding between the opposing theories. The most precise ones performed at the time were those of Bücherer; and they were favorable to Lorentz's anticipations. Abraham's pure field theory was abandoned in consequence. But even had Abraham's views been supported by experiment, other difficulties would still have barred the way to a pure field theory. For instance, the electric charges on the surface of an electron, or in its interior, being of the same sign, should repel one another, and hence the electron should explode. To account for the electron's stability, we must postulate some intense extraneous pressure forcing back the escaping charges. But this pressure (Poincaré pressure) cannot be built up from the field quantities of Maxwell and Lorentz; and so it must be of non-electromagnetic origin. The electromagnetic field by itself is thus incapable of accounting for the stability and the existence of the electron. For these reasons and

many others, any pure field theory founded on the Maxwell-Lorentz electromagnetic field seems doomed to failure from the start. A more successful attempt was made at a later date by Mie, and at one time Weyl favored Mie's theory; but he has since abandoned it. In view of these failures physicists were thrown back on the dualistic theory of Lorentz, and it is the development of this theory and its connection with the theory of relativity that we shall now examine.

The theory of relativity, whether in the Lorentzian or the Einsteinian form, arose from the consideration of a group of phenomena known under the name of the electrodynamics of moving bodies and of moving fields. Lorentz's electron-field theory, even when supplemented by the FitzGerald contraction, was unable to account for the experimental results. In order to bring theory and practice into agreement, Lorentz was compelled to postulate revolutionary relations between space and time. These relations, which automatically embody the FitzGerald contraction, are expressed by the celebrated Lorentz transformation. Had Lorentz realized that the space and time, which he had subjected to the new relations, were the real space and time of physical science, he would have been the creator of the theory of relativity. But the progress of science is gradual, and before revolutionary changes are accepted, attempts are usually made to interpret the facts at our disposal in terms of those classical notions which have proved their worth in other situations. Lorentz accordingly viewed his transformation as purely formal and as having no bearing on the space and time of physics.

Then a year later, Einstein adopted the more revolutionary course, contending that the Lorentz transformation expressed the relations between physical space and time. This departure marked the start of the special theory of relativity. The ether, which had already been deprived by Lorentz of all its mechanical attributes, was now shorn of its ability to define a frame of reference in which velocity could be measured. The new views received a tremendous impetus from Minkowski's discoveries. Minkowski showed that if Einstein's contention were correct, this world of ours could not be the familiar world in which space and time are separate, but must be a strange non-Euclidean 4-dimensional world, since called *space-time*.

With the introduction of space-time, all physical magnitudes, formerly represented in the 3-dimensional space and separate time of classical science, now had to be expressed as 4-dimensional magnitudes of the new world. The new representation thereby entailed the crediting of additional components to the classical 3-dimensional magnitudes. The extra

components turned out to be well-known magnitudes, so that the novelty of the space-time form of representation resided more particularly in its fusion of physical magnitudes formerly viewed as distinct. The net result was that many such physical magnitudes revealed themselves as the various components of the same 4-dimensional existents. The simplest illustration of this fusion is that of space and time, but there are others. Thus momentum and energy are unified, and the separate laws of the conservation of momentum and of energy are seen to express but one law of conservation in a 4-dimensional world. Force and work are also merged; so are electric current and charge.

It is particularly instructive to examine the fusion in the case of the electromagnetic-field magnitudes. As we have seen, in classical 3-dimensional space the electromagnetic field may be viewed as due to the superposition of two separate fields, the electric field, which is a vector field, and the magnetic field, which is a tensor field. Both fields are of course 3-dimensional since they are situated in 3-dimensional space. We also mentioned that these two fields are not mutually independent, for variations in the one entail modifications in the other. Maxwell's four equations express the mutual relations between the two fields. We must now transcribe the field magnitudes into 4-dimensional space-time. Calculation shows that the 3-dimensional electric vector and the magnetic tensor merge into a single 4-dimensional tensor, called the electromagnetic tensor. In other words, the 3-dimensional electric and magnetic fields become fused and appear as mere partial aspects of the same underlying 4-dimensional tensor field—the electromagnetic field. The interdependence of the electric and magnetic fields (which in 3-dimensional space had to be accepted as a fact of experience) is thus an immediate consequence of the new outlook. At the same time the exact quantitative expression of the connection between the two partial fields can now be obtained, even in the more involved situations which deal with the electrodynamics of moving bodies. Maxwell and Lorentz had to guess at this connection in the more complicated cases, and in point of fact guessed wrongly. Finally, we may mention that Maxwell's four 3-dimensional equations are reduced to two 4-dimensional ones.

In the special theory of relativity, Minkowski's space-time is viewed as flat (uncurved), rigid, and indeformable; it is thus the 4-dimensional counterpart of Newton's absolute space. But it has over Newton's space the advantage of showing clearly why uniform motion cannot be detected. This point may be understood when we note that in space-time the history of a particle, whether it be in a state of rest or of uniform motion, is

represented by a straight world-line,* and that there is no means of differentiating one straight world-line from another. We also understand why acceleration may be distinguished from rest or uniform motion; it is because, in the event of acceleration, the space-time point describes a curve and no longer a straight line. Though the general concept of an ether, filling space-time, may be retained, no apparent justification can be advanced in its favor, and so the electromagnetic field is usually regarded as symbolizing a peculiar condition (non-mechanical) existing throughout space-time itself. The special theory of relativity is as dualistic as Lorentz's theory, because a successful interpretation of matter and electrons in terms of the field magnitudes seems as remote as ever.

The special theory is not, however, confined solely to electromagnetic and to mechanical actions; it also sheds some light on the phenomenon of gravitation. For example, a feature of Minkowski's space-time is that a velocity exceeding the velocity of light *in vacuo* cannot be attained, a fact which suggests that no physical action can be transmitted with a speed greater than that of light. Action at a distance even in the case of gravitation, seems therefore to be ruled out by the basic tenets of the special theory. The condition of Lorentzian invariance, which is characteristic of the special theory, also restricts the form of all natural laws and thereby affords a clue to the correct law of gravitation. Poincaré set himself the problem of amending Newton's law of gravitation so as to bring it into agreement with the Lorentzian restrictive condition. He obtained a slightly modified form of the law of gravitation, a form in which the action is propagated with the velocity of light. Incidentally, Laplace's error (see page 75) was disclosed in this investigation, and the way opened for a field interpretation of gravitation. If Poincaré's treatment is accepted, we must assume that Minkowski's rigid space-time may be permeated by a gravitational field (presumably of a non-mechanical nature), which may be superposed on the electromagnetic field.

Einstein, in his general theory of relativity, followed a different line of attack. He discarded his original conception of a perfectly rigid, flat space-time and assumed that the space-time structure, though normally flat, would become warped in the neighborhood of matter. It was this warping which was responsible for the presence of a gravitational field in the ordinary sense. The gravitational field, in contradistinction to the electromagnetic field, thus received a geometric interpretation.

* The successive points in space occupied by a particle over a period of time are represented in 4-dimensional space-time by a succession of space-time points. The aggregate of these space-time points defines the world-line of the particle.

So as to obtain a clearer understanding of the new views, let us revert to our hypothetical illustration of the sun suddenly coming into existence. For convenience we shall express the situation in terms of the more familiar categories of ordinary space and time. The sudden creation of the sun generates a warping of the immediately surrounding continuum. The warping then spreads in all directions with the velocity of light, decreasing in intensity as it proceeds. When the warping reaches a planet, the planet acts as though it were attracted towards the sun. It is, however, the warping of space-time in the immediate neighborhood of the planet which is responsible directly for the force. As in all field theories of gravitation, the sun is thus only indirectly responsible for the action.

In point of fact the 4-dimensional gravitational field, which Einstein identifies with the space-time structure, is of a far more general nature than is a gravitational field in the ordinary sense. In common parlance we should not speak of a gravitational field in a region of space infinitely distant from matter. Yet, inasmuch as in a distant region, space-time still has a definite structure (flat), Einstein's gravitational field is still present. To avoid confusion on this score, it is usual to refer to Einstein's 4-dimensional space-time field as the "metrical field" or the "guiding field." The name "metrical field" is justified by the fact that Einstein's field determines the geometry which we obtain when we perform measurements with rods or with rays of light. The name guiding field is given because a free body, *e.g.*, a planet, follows a geodesic world-line through the space-time structure, as though it were guided by a groove.

Einstein's celebrated field equations express the general law of varying (or permanent) structure which space-time may assume; a special case of this structure is the flat structure, which holds far from matter. In the language of ordinary space and time, the equations embody the laws which control the permissible distributions and changes of the gravitational field. As such, Einstein's gravitational equations are the gravitational analogues of Maxwell's equations of the electromagnetic field.

We now examine in what respects Einstein's law of gravitation differs from Newton's. One way of illustrating the difference is to express Einstein's law in terms of the artificial, though more familiar, categories of ordinary space and time. But before doing this, we must mention one of the formulations of Newton's law of gravitation. Suppose we are considering the space surrounding the sun. In Newton's treatment, no appeal is made to a field as connoting a condition of space, but we may nevertheless speak of a field of force in a mathematical sense. Let us then

consider the field of force which surrounds the sun. Since the force at any point is represented in magnitude and in direction by an arrow, we are dealing with a vector field. Instead of considering the gravitational force, we may introduce a magnitude which is related to the force and which is called the "Newtonian potential." Newton's gravitational field is then represented equivalently by the distribution of the Newtonian potential. The potential, in contradistinction to the force, is expressed by a simple number, *i.e.*, a scalar. The field defined by the potential distribution is therefore a scalar field. The mathematical relation which determines the possible distributions of the potential around matter is given by a celebrated equation studied by Laplace (Laplace's equation).

With this preliminary information disposed of, we may return to Einstein's 4-dimensional gravitational field. Let us picture it in ordinary 3-dimensional space. If we assume that the field is permanent, as it would be around a permanent body such as the sun, the representation of the space-time field may be regarded as due to the superposition of two separate 3-dimensional fields: a scalar field and a tensor field. If the mass of the body (the sun) generating the field is not too great, this scalar field is found to be none other than the field of the Newtonian potential. We conclude that the first field of the theory of relativity, if it existed alone, would entail Newton's law of the inverse square. We now pass to the second relativistic field: the tensor field. If, as before, the mass generating the gravitational field is not too great and if, in addition, the velocity of the planet is not too high, the second field has only a small influence. To the second field are due the small deviations between the anticipations of Einstein and of Newton. In particular, the secular advance of the perihelion of Mercury must be attributed to it. To avoid misconceptions, we must mention that this splitting of 4-dimensional space-time into 3-dimensional space and a separate time, which we have here considered for motives of simplicity, is not only artificial but also inaccurate. Furthermore, it obscures the fact that the two gravitational fields mentioned form a single 4-dimensional unit, just as do the two fields in electromagnetics.

The entire general theory revolves around Einstein's 4-dimensional gravitational field-equations, and hence around the structure, or field, of space-time. For this reason the general theory of relativity is a field theory. But it is not a pure field theory, because, in addition to the field, we have to consider matter under the corpuscular form of electrons, protons, and the like. Insofar as the theory of relativity is concerned, the atomicity of matter is accepted as an empirical fact, just as is the rotundity of the earth; no theoretical reason for this atomicity is even

suggested. Although a pure field theory would require that we account for the atomicity of matter in terms of the field magnitudes, this aim has yet to be realized. The situation does, however, appear more promising than it did in electromagnetics, for now we have two fields at our disposal—the electromagnetic field and the gravitational, or metrical, field.

Various attempts were made by Mie and others to construct a pure field theory, but they were only partly successful. Einstein then attempted the reverse course, seeking to account for the field in terms of matter distributed more or less uniformly throughout space. He was thus led to the hypothesis of the "cylindrical universe" and to the introduction of the cosmological constant λ, which is proportional to the mean density of matter. In view of the existence of competitive cosmological theories and owing also to recent astronomical discoveries, this latter part of the relativity theory is still extremely speculative, and we shall not dwell on it. Turning to the more solidly established aspects of the theory, we may say that it is as dualistic as Lorentz's. Its primary concern is the field, but, for empirical reasons, electrons and matter atomistically constituted must be introduced subsequently.

Unified Field Theories—The general theory of relativity deals with two separate fields, the electromagnetic field and the metrical (or gravitational) field. The electromagnetic field acts on electrified bodies and is itself influenced by their presence. Similarly, the metrical field acts on masses and is influenced by them. Up to this point, the analogy between the two fields is very great; but there are also important differences which we shall now consider.

In the first place, the metrical field receives a geometric interpretation since this field is a mere expression of the geometrical structure of space-time. The electromagnetic field, on the other hand, cannot be attributed to the space-time structure. It is true that the presence of an electromagnetic field modifies the structure, but we cannot regard this modification as accounting entirely for the presence of the field. The electromagnetic field must therefore be regarded as resulting from the presence in space-time of another type of physical condition, the electromagnetic tensor. Thus, whereas the metrical field is a characteristic of space-time itself, the electromagnetic field appears as a foreign intrusion.

Let us also observe that the entire relativity theory would collapse were space-time to have no structure: in the absence of a structure the distinction between space and time would vanish; there would be no geometry, no geodesics, and hence there would be nothing to regulate the

paths and motions of bodies. The absence of a space-time structure being incompatible with the theory of relativity, we conclude that a metrical field is a necessity in this theory and cannot possibly be non-existent. But the electromagnetic field, being divorced from the structure, may quite well be absent without endangering the theory. A further dualism is thus revealed: the metrical field is essential and the electromagnetic field contingent. The unity of the theory of relativity is marred by this dualism.

The subsidiary rôle played by the electromagnetic field, as contrasted with the fundamental nature of the metrical field, entails interesting consequences. Thus, when we wish to determine the path and motion, say, of a planet, we first determine the space-time curvature around the sun and we deduce therefrom the lay of the geodesics (straightest lines). The world-line of the planet will then lie along the geodesic corresponding to the initial position and velocity of the planet. When the geodesic is known, the path and motion of the planet are obtained. This is what is meant when we say that the space-time structure guides the planet. But the situation is entirely different when, in place of a massive body moving in a gravitational field, we consider an electrified particle moving in an electromagnetic field: the world-line of the electrified particle no longer coincides with a geodesic and, indeed, exhibits no connection with the space-time structure. A considerable gain in unity would be obtained if we could establish some connection between the world-line of the electrified particle and the space-time structure, for then the world-lines of the massive body and of the electrified particle would both be dependent on the same underlying space-time structure. Now this unification would be secured if we could merge the two fields, identifying them both with the structural peculiarities of space-time. We should then have what is known as a "unified field theory."

The earlier development of the theory of relativity gives a clue to the method whereby the unification may be obtained. We saw that, in 3-dimensional space, the electromagnetic field may be regarded as resulting from the superposition of two fields, which are different though causally related. We then found that when 3-dimensional space is replaced by 4-dimensional space-time, the fusion of the two fields is realized. This circumstance suggests that any attempt to unify the electromagnetic and the metrical fields will presumably involve an extension of the fundamental 4-dimensional continuum. In Kaluza's theory, the extension is secured by the addition of a fifth dimension; in Weyl's, no extra dimensionality is invoked, but, as against this, the continuum is held to be

susceptible of a richer variety of structures. Weyl's theory was subsequently extended by Eddington. Kaluza, as a result of his introduction of a fifth dimension, was able to identify a geodesic of the fundamental continuum, not solely with the path and motion of a mass in a gravitational field, but also with the path and motion of an electron in an electromagnetic field. Einstein himself, after investigating a different generalization from those suggested by Weyl and Kaluza, is now said to be working on a modified form of Kaluza's suggestion. Though retaining 4-dimensional space-time, he nevertheless introduces field magnitudes that were thought to be possible only in a 5-dimensional world.

Any one of these attempts, if successful, would furnish a unified field, both gravitational and electromagnetic. The next step would then presumably be to secure a pure field theory, by interpreting matter and electrons in terms of the field magnitudes. The foremost theoretical physicists, however, have expressed the belief not only that a pure field theory is impossible, but also that field theories in general can give only a macroscopic picture of reality. The fact is that all field theories are rigorously deterministic, or at least have always been assumed as such. But quantum mechanics leads to the suspicion that rigorous determinism is a myth, and that the determinism accepted by classical science (and the theory of relativity) refers merely to average results, not to individual processes. If this be so, the field theories of the past can be compatible only with a macroscopic survey of phenomena.

In addition to these objections of a general nature, to which all field theories are open, there are others which pertain to Maxwell's theory in particular. According to Maxwell's theory of the electromagnetic field *in vacuo,* any disturbance emitted from a point-source should be propagated equally in all directions. Consequently, radiation from a point-source of light should consist of spherical waves. Maxwell's field theory is thus in conflict with the corpuscular nature of radiation, which recent experiment appears to demand. On the other hand, Maxwell's equations must certainly contain a large part of the truth, for their mathematical consequences express physical results which are nearly always verified by experiment. The difficulty may be explained away if we suppose that Maxwell's field theory represents an approximation to reality—an approximation which is sufficient for our needs so long as we view phenomena from a macroscopic standpoint.

One of the first successful attempts to interpret Maxwell's continuous electromagnetic field as the macroscopic manifestation of fundamentally atomistic processes is the theory developed by Dr. Leigh Page in 1914.

He postulated particles moving with the velocity of light and showed that a macroscopic condition which could be represented by Maxwell's field equations would arise. Dr. Page writes:

"Instead of being propagated through an elastic medium in the manner of material waves, electromagnetic waves are to be thought of as carried by discontinuities emitted from the source with the velocity of light." *

With the discovery of Heisenberg's uncertainty relations, the conflict between the continuous field and the discrete corpuscles of radiation assumes a novel aspect. The continuity of the field is connected with the theoretical uncertainty which prevents the precise localization of particles. In place of discrete particles we have blurred regions of space which give the impression of continuity. But it should be added that the entire subject is still extremely obscure.

* Leigh Page. The Emission Theory of Electromagnetism. Transactions of the Connecticut Academy of Arts and Sciences, 26, Jan. 1924; pp. 213-243.

CHAPTER XI

PHENOMENOLOGICAL THEORIES

THE source of our knowledge of the physical world resides in the sensations of light, heat, sound, touch, and the like, which we experience directly. The world of the infant consists in bare sensations. Then gradually, through the inferences to which cumulative experience gives rise, sensations are sorted out, connected, and formed into a mental picture of the outside world. The more or less conscious efforts of our infancy are soon forgotten or ignored, and unless we be inclined towards introspection and analysis, we find it natural to say that concrete objects, such as the table in our room, are observed directly.

It is useless to quibble over the meaning of words, and so we shall agree that, when no conscious inferences are made, a phenomenon will be said to have been observed directly. But suppose now that on peering through a microscope we see a strange pattern of rods and circles. Any person acquainted with a microscope will contend that he has observed the texture of the preparation placed in the slide. The inferential nature of this conclusion is, however, undeniable, for it would not be made by a savage. If, then, we insist on speaking of "direct observation" in connection with the microscope, we are certainly extending the previously accepted meanings of words.

Let us proceed a step further. The quantum theorists regard ultra-violet light as observable. Yet ultra-violet light is invisible. When we speak of ultra-violet light as observable, what we really mean is that certain observable occurrences (*e.g.*, those affecting a photographic plate) are attributed to the presence of ultra-violet light. But this conclusion is manifestly inferential, and to class ultra-violet light among the observable existents is to accept a still more extended meaning of words than was done in the example of the microscope.

The preceding analysis shows that a large number of the observables in physics are revealed through inference. If the inference is so plausible that it would be accepted without question by any man regardless of his training, the distinction between the observables of physical science and of ordinary life is not so striking. But when the inference forms a characteristic feature of some sophisticated theory, the magnitudes which

are observable in principle according to this theory may cease to be so when, at some subsequent date, the theory is rejected.

As an example, let us consider the kinetic theory of gases, in which elastic molecules are assumed to be rushing hither and thither in accordance with the laws of mechanics. The pressure which the gas exerts against the walls of the enclosure may be observed directly, for instance, by the effort we have to exert on a piston to compress the gas. The inference characteristic of the kinetic theory is to ascribe this pressure to the impacts of the invisible molecules. In virtue of this inference the molecular impacts become observable, at least in much the same indirect sense that ultra-violet light is claimed to be observable. On the other hand, if we reject the kinetic theory, the impacts cease to be observable.

A further illustration is afforded by the concept of absolute acceleration. Both the classical theory and the general theory of relativity recognize that the so-called forces of inertia can be detected and measured. But whereas the classical theory ascribes these forces to an acceleration through absolute space, the general theory of relativity connects them with the contingent structure of space-time. The net result is that, from the classical viewpoint, absolute acceleration is revealed through the forces of inertia; whereas, from the standpoint of the theory of relativity, these forces betray something entirely different and absolute acceleration is meaningless.

Planck expresses the situation in the following passage:

"For the question whether a physical magnitude can in principle be observed, or whether a certain question has a meaning as applied to Physics, can never be answered *a priori*, but only from the standpoint of a given theory. The distinction between the different theories consists precisely in the fact that according to one theory a certain magnitude can in principle be observed, and a certain question have a meaning as applied to Physics; while according to the other theory this is not the case. For example, according to the theories of Fresnel and Lorentz, with their assumption of a stationary ether, the absolute velocity of the earth can in principle be observed; but according to the theory of Relativity it cannot; again, the absolute acceleration of a body can be in principle observed according to Newtonian mechanics, but according to Relativity mechanics it cannot. Similarly the problem of the construction of a *perpetuum mobile* had a meaning before the principle of the conservation of energy was introduced, but ceased to have a meaning after its introduction. Hence it is not sufficient to describe the superiority of Quantum-mechanics as opposed to classical mechanics, by saying that it confines itself to quantities and magnitudes which can in principle be observed, for in its own way this is true also of classical mechanics. We

must indicate the particular magnitudes or quantities which, according to Quantum-mechanics, are or are not in principle observed; after this has been done it remains to demonstrate that experience agrees with the assertion." *

Planck's lucid presentation stresses the difficulty of deciding *a priori,* in the absence of a theory, whether a postulated magnitude is observable or not. We must remember, however, that in this discussion the word *observable* is used in a highly sophisticated sense. If, on the other hand, we take the word "observable" with its commonplace connotations, neither absolute space nor ultra-violet light is observable. The usual practice is to adopt neither the highly sophisticated definition of the word "observable" nor its commonplace meaning. A compromise is accepted, and the distinction between magnitudes that are observed and those that are not observed is taken to rest on the more or less intuitive appeal that the inferences from the accepted theories would have for the average man. Ultra-violet light, for instance, will be said to be observable, whereas absolute space and the motions of the molecules in the kinetic theory will be classed as unobservable. Some degree of vagueness is inevitable, as in all situations where a compromise is made, but in the main no serious disputes need arise. At all events, it is in this last, half-way sense that the word observable will be applied in the remainder of this chapter.

The purpose of a physical theory is to establish causal relations between the facts observed, so that, from given facts, others may be predicted. But a causal connection between two observed magnitudes, A and B, need not express a direct relation; more often we have a causal chain extending from A to B and involving connections between intermediary magnitudes. Here a difference in method may arise.

1. We may stipulate that all the magnitudes in the causal chain must be observable. Theories in which this restriction is imposed are called "phenomenological"; thermodynamics is the outstanding example. In this science the quantities that appear in the equations, or causal relations, are the observable magnitudes: pressure, volume, quantity of heat, temperature; and also auxiliary magnitudes, such as entropy, built up from the former.†

2. We may disregard the foregoing restriction and accept intermediary magnitudes which are not observed, but which are postulated so

* Max Planck. The Universe in the Light of Modern Physics. W. W. Norton and Co., New York, 1931, p. 49.

† Another magnitude which enters into the equations of thermodynamics is "internal energy," and it is questionable whether this magnitude can be said to be observable.

as to secure a causal connection between the facts that are observed. Theories of this type involve hidden occurrences and are obviously more speculative, for the evidence in favor of the precise hidden occurrences postulated is necessarily indirect and inferential.

Illustrations of the latter type of theory are found in the mechanistic theories and in the field theories discussed in the preceding chapters. For instance, the kinetic theory of gases postulates invisible occurrences; they consist in the disorderly movements and collisions of the molecules. Einstein's theory of the gravitational field affords another example. Thus, if we confine ourselves to what is actually observed, all we can say is that a stone, when released, falls to the ground. Our attitude is then equivalent to supposing that there is a direct relationship between the earth and the stone. But in the general theory of relativity the causal connection is indirect: the earth causes a warping of 4-dimensional space-time; and this warping, which is not observable, is then assumed to be the cause which directs the stone in its fall to the earth.

In many cases the same group of phenomena may be investigated either by a phenomenological theory or by one which introduces hidden occurrences (*e.g.*, both thermodynamics and the kinetic theory can be applied to gases). In the phenomenological theories, our attention is restricted to the macroscopic properties which appear on the commonplace level of experience. These macroscopic properties are viewed as fundamental, and no attempt is made to analyze them further. In the opposing theories, the macroscopic properties are dissected, as it were, and usually (but not always) microscopic entities are postulated to represent the fundamental reality behind the macroscopic appearances. For this reason we may often refer to the latter theories as "microscopic theories."

Lorentz, in connection with electromagnetic phenomena, developed both kinds of theory. Let us first examine the macroscopic aspect of electromagnetic phenomena in the presence of matter. These phenomena are represented by electric currents, charged bodies, and fields of force. Matter appears under the form of conductors and non-conductors (dielectrics). A macroscopic theory would involve systems of relations connecting these magnitudes directly. Lorentz, however, in his attempt to construct a macroscopic theory, proceeded indirectly by first concentrating on the microscopic theory. He assumed that, from the microscopic standpoint, an electromagnetic phenomenon would only include electrons, protons, and fields. No fundamental unit of magnetism was invoked, so that magnetism was viewed as a macroscopic effect, much as temperature is shown to be in the kinetic theory of gases. Similarly, the electric currents of the macroscopic level were ascribed to streamings of

electrons. Lorentz's microscopic theory, represented by his microscopic equations, expressed the relations connecting the electrons and fields. Lorentz then obtained the macroscopic magnitudes by submitting the microscopic ones to an averaging process. Magnetism and electric currents reappeared, and the differences in the behavior of conducting and of nonconducting media were accounted for. Lorentz's macroscopic laws or equations were derived in a similar fashion from the microscopic equations by an averaging process, the microscopic equations being regarded as fundamental.

Mach and Ostwald in the course of the last century insisted that theories which postulate hidden occurrences should be rejected and that phenomenological theories alone should be retained. A somewhat similar philosophy has been defended in recent years by the quantum theorists, but the incentives of the quantum theorists differ so greatly from those of Mach that it will be advisable to examine them separately. We first consider the views of the Machian school; those of the quantum theorists will be discussed towards the end of the chapter.

According to the school of Mach and Ostwald, theories which introduce hidden occurrences smack of metaphysics and should be discarded in principle lest we fall into the errors of the metaphysicians. It seems difficult to accept this view, for the repugnance which the average man of science experiences towards the writings of the metaphysicians is due, not to the speculative spirit manifest in their schemes, but to their disregard or ignorance of the most elementary facts of mathematics and of physics. Inasmuch as Mach and Ostwald did not clarify their idea further, their criticisms appear to have been dictated by preference rather than by reason. At all events their views found favor with many physicists who, though of first rate ability in the experimental field, never evidenced any particular liking for mathematics. Many of them have expressed their mathematical dislike openly, and one cannot help suspecting that the more difficult mathematical technique of the microscopic theories may have influenced their philosophy.* The fact, however, that the critics most averse to mathematics usually sponsor thermodynamics as a theory suiting their taste would seem to mitigate the force of their criticisms, for if we accept thermodynamics we are automatically accepting mathematics as an adjunct to physical research. Under these conditions the degree of difficulty of the mathematics incorporated in a theory should play no part in a philosophical discussion, especially so since the difficulty of mathematics depends largely on the amount of study we have

* Ostwald was a chemist and Mach made no contribution to theoretical physics.

devoted to the subject. The only fair way to determine the merits of the two opposing kinds of theories is to examine the influences they have exerted on the development of science. We shall find that theories of both types have been of extreme importance and that they have supplemented one another. We consider the phenomenological theories first.

Thermodynamics, which is the prototype of these theories, requires for its development the most elementary kind of mathematics; a few superficial notions of the calculus suffice. Furthermore, owing to its neglect of microscopic differences, thermodynamics deals with relatively simple situations. All this has an important bearing on the nature of the problems that can be solved by thermodynamical methods. Many discoveries which have been arrived at with comparative ease in thermodynamics would have been arrested by hopeless complications had a microscopic theory been adopted. Gibbs's celebrated phase rule is a case in point.

Because of its rejection of speculative hidden occurrences, a phenomenological theory is on safer ground than a microscopic one. This may be an advantage from a certain point of view, but it would not seem to constitute any decided superiority. Scientific progress is not realized by following dogmatically the motto, "Safety First." If it were, the best course would be to do nothing and to speculate on nothing; and this would lead to utter stagnation. Besides, the safety and reliability of a phenomenological theory is usually more apparent than real.

For example, so long as we are guided solely by the methods of thermodynamics, we have no reason to cast doubt on the law of entropy. Yet when this law is investigated by means of the kinetic theory of gases, a microscopic theory, its validity is seen to be only statistical. We may readily understand why thermodynamics failed to reveal the true situation, and why similar limitations are to be expected in all phenomenological theories. Thus when we restrict ourselves to a phenomenological, or macroscopic, theory, the only relations and properties that we are concerned with are those disclosed by direct observation. The mathematics of the theory then enables us to derive the necessary consequences of these observed relations, and by this means predictions can be made. But the drawback to this procedure is that the mere refusal on our part to take unobservable occurrences into consideration does not suppress the possibility of their existence; and if such occurrences do exist, they too will entail phenomena, some of which may be perceptible. If these latter perceptible phenomena are generated only on very rare occasions, our original direct observations may fail to detect them; and the net result

will be that our macroscopic theory will leave us in complete ignorance of their existence. For these reasons a phenomenological doctrine is not necessarily as reliable as we might suppose.

Other illustrations may be given. The theoretical values of the chemical constants of thermodynamics, the reasons why some chemical reactions progress more slowly than others, catalytic effects and the like, are beyond the province of thermodynamics as such. Here speculative assumptions which are certainly not observable (in the sense we have adopted) must be introduced.

A similar situation occurs in the quantum theory. Physicists were seeking the law of equilibrium radiation. Thermodynamical arguments proved of considerable help by imposing a restriction on the form of the unknown law.* But thus far and no further could they go. Planck's microscopic assumptions were required to derive the radiation law which bears his name.

We have mentioned the most conspicuous advantages and disadvantages of the phenomenological theories. Let us now pass to the microscopic theories. Much light is cast on the various problems at issue when we examine the reasons that gave rise to these theories. We recall that physical science originated from the accumulation of observational and experimental facts. Since it was the magnitudes and conditions observed on the commonplace, or macroscopic, level which forced themselves most immediately on the attention of physicists, the earlier facts necessarily pertained to macroscopic properties; and as a result the empirical laws secured by connecting these facts were likewise of the macroscopic variety. But investigators did not rest content with this mode of procedure, and even in Newton's day we find microscopic theories of the corpuscular and also of the continuous type developed in connection with optical phenomena. In the eighteenth century Daniel Bernoulli conceived the kinetic theory of gases. Bernoulli was undoubtedly guided to his theory by the empirical facts which suggested a corpuscular constitution for matter. Even so, in view of the hypothetical nature of his assumptions, we may well inquire what advantage he saw in taking this leap in the dark. Why should Bernoulli not have been satisfied with the empirical laws and causal relations established by direct experiment? Presumably, his desire was to obtain a better understanding of things—to peer behind the scene with the hope of tracing the various macroscopic properties to a common cause. The appreciation of what constitutes the most satisfac-

* The restriction leads to the conclusion that the law of radiation must have the general form expressed by Wien's relation (page 453).

tory form of understanding is, however, largely a personal matter which differs from one individual to another. We shall not dwell therefore on the motives that may have dictated Bernoulli's course. Instead, let us judge his kinetic theory by its results.

Thanks to the mathematical development of the kinetic theory at the hands of Maxwell and Boltzmann, unsuspected peculiarities of gases were revealed; for instance, Maxwell showed that the viscosity of a gas is independent of its pressure. More important still, as we mentioned previously, the law of entropy was shown to have but a statistical validity. The usefulness of microscopic theories is further revealed in electromagnetics, for Lorentz's microscopic theory enabled him to obtain his macroscopic equations and to supplement the deficiencies of Maxwell's equations in the presence of matter. To this microscopic theory also is due the prediction of the Zeeman effect and other phenomena.

From this analysis we conclude that both kinds of theories have their advantages and their disadvantages, and that the type of theory to be preferred depends largely on the viewpoint we propose to adopt. When we deal with matter in bulk, thermodynamics is sufficient for our needs, the exceptions to the principle of entropy occurring on such rare occasions that as yet no one has ever observed them. But if our concern is with minute volumes, the microscopic theories cannot be dispensed with. The fame acquired by an investigator in developing a theory of one type or the other is not necessarily a proof that his preferences lean in one direction. More often it is the circumstances surrounding his discoveries which govern his course. Indeed many of the greatest theoretical physicists have devoted their efforts to the development of both kinds of theories. Maxwell, Gibbs, Planck, and Einstein are among the most distinguished.

The New Quantum Theories—The new quantum theories, developed by Heisenberg, Born, Pauli, Dirac, and Bohr, are usually classed as phenomenological. When Heisenberg first formulated his Matrix Theory, considerable stress was placed by him on the necessity of introducing none but observable magnitudes in the equations. Heisenberg's attitude might be thought to be the same as Mach's. But a closer examination of the new theories and of their recent developments shows that this belief would be unwarranted.

In the first place, we must remember that the quantum theorists are among the foremost theoretical physicists of the present day, so that the mathematical difficulties which may have colored the philosophy of the Machian school need not be reckoned with, Furthermore, we shall see

that the quantum theorists did not start out by rejecting hidden occurrences on principle. In fact their initial attempts were directed towards the microscopic theories, and only after repeated failures was a more phenomenological outlook adopted as a last resort. Finally, we are scarcely justified in calling the new quantum theories phenomenological; for though their equations connect only observable magnitudes,* these magnitudes are not connected directly, but only indirectly by means of an underlying network of intricate mathematical relations. The relations are of a degree of abstraction never before encountered in physical theories, and they cannot be visualized in terms of familiar notions. What we are witnessing is an extreme mathematicizing of physical Nature. It would therefore seem that the new quantum theories are more adequately expressed by the appellation "mathematical theories." From the standpoint of the disciple of Mach and Ostwald, the present situation is worse than ever before.

The motives and characteristic features of the new theories will best be understood if we follow the historical order of their appearance. This leads us back to Bohr's mechanistic theory of the atom, because it was the failure of Bohr's theory to give a satisfactory account of subatomic phenomena that precipitated Heisenberg's matrix method. When Bohr initiated his theory, the empirical evidence indicated that an atom was similar to a miniature solar system. Thus an atom was supposed to be represented by a central nucleus (built up of electrons and protons) which had a net positive charge and around which electrons were circling in accordance with the laws of mechanics. In this model the electrostatic attraction of the nucleus on the electrons played the part of the sun's gravitational attraction on the planets. The simplest of the atoms was the hydrogen atom, in which a single electron was circling around a proton as nucleus. The model just described represents the hidden microscopic occurrences, the radiations emitted from the atom constituting the observed events which must receive their interpretation from the invisible motions. The foregoing conception of the atom was adopted by Bohr, so that his theory was a mechanistic theory of the corpuscular type.

The application of the classical laws of mechanics and of electrodynamics to this model did not, however, account for the precise radiations observed. Even in the atom of hydrogen, which because of its simplicity offered the best hope of success, the distinctive spectral series could not be interpreted. The originality of Bohr's theory consists in imposing so-called quantum restrictions on the classical mechanical laws,

* These include ultra-violet light.

and also in rejecting classical electromagnetics in its entirety. Thanks to Bohr's restrictions and postulates, the spectral series of hydrogen were accurately predicted, and since Bohr's theory was the only theory then known which was able to secure this result, it was accepted. The restrictions imposed by Bohr were not exactly hypotheses *ad hoc* postulated for the sole purpose of interpreting the hydrogen spectrum. They were natural extensions of the quantum assumptions previously introduced by Planck in a totally different field of investigation; namely in the study of black-body radiation. Bohr's results served therefore as an indirect confirmation of Planck's original assumptions and strengthened the belief that, with quantum phenomena, a new body of physical occurrences had been uncovered.

At first sight it would seem that two different sets of mechanical laws should be assumed: firstly, the classical (or relativistic) laws which hold for the planets of the solar system and more generally for bodies moving outside the atoms; secondly, the quantum-restricted mechanical laws which regulate motions inside the atom. In itself this dualism is unsatisfying; besides, since an atom has no definite boundary, we cannot decide under what conditions an electron will be moving inside or outside a given atom. Bohr proved, however, that the dualism was more apparent than real. He showed that the effect of the quantum restrictions would tend to vanish when the electron was describing a larger orbit round the atomic nucleus, or when the dimensions of the mechanical system were increased.* According to this view, classical mechanics supplemented by the quantum restrictions constituted the correct mechanics; classical mechanics in the absence of the restrictions was a mere approximate doctrine tending to become valid only on the macroscopic level of experience.

Thus far we have been concerned with the hydrogen atom, but the situation became hopelessly complicated when the heavier atoms were considered and when a more careful examination was made of the atomic spectra. It was found necessary to introduce, as and when needed, additional quantum restrictions on the mechanical laws. The expression of some of the mechanical magnitudes had to be modified and the mechanical principles revised accordingly. Even when all these changes were made, Bohr's theory was unable to predict with precision the phenomena actually observed. The belief generally entertained at this stage may be expressed in the following statement:

There is but one mechanics holding within and without the atom, and it differs considerably from the classical variety formerly thought to be

* A similar situation was shown to hold for the electromagnetic modifications which Bohr had postulated.

correct. On the macrascopic level we may accept classical mechanics without danger, for the refinements of the new mechanics then cease to exert perceptible effects; but if we are dealing with a microscopic mechanism, the refinements become perceptible, and we obtain totally incorrect results unless the correct mechanics is applied. Unfortunately the precise form of the correct mechanics is unknown, so that we cannot construct a mechanistic theory of the atom. Thus, though in principle a mechanistic theory of the hidden-occurrence type is conceivable for the atom, it must be rejected in practice.

This brief summary gives an idea of the situation that was responsible for the rejection of Bohr's mechanistic theory and that prompted Heisenberg to devise a new method of attack. Heisenberg, in his matrix theory, did not deny that an atom was constituted as had been assumed by Bohr, namely, by electrons circling round a nucleus. But he pointed out that since the laws regulating the electronic motions were utterly mysterious, the best course was to disregard these motions and to concentrate on the radiations emitted from the atomic system. Accordingly in Heisenberg's method, the positions, motions, and orbits of the electrons in the atom were not mentioned, and the aim was to connect the radiations by appropriate mathematical schemes. Thus, it was under the stress of circumstances, and not by choice, that Heisenberg and the other quantum theorists came to formulate a phenomenological theory of the atom. No *a priori* rejection of the hidden-occurrence theories is found in the writings of any of these men. This last point is sufficiently obvious when we note that in addition to Bohr himself many of the other quantum theorists, chiefly Born, Pauli, and Heisenberg, had devoted considerable time to Bohr's original mechanistic atom.

The following passage, written by Born in 1926, exhibits the idea we are expressing.

"We therefore stand before a new fact which forces us to decide whether the electronic motion or the wave shall be looked upon as the primary act. After all theories which postulate the motion have proved unsatisfactory we investigate if this is also the case for the waves." *

A study of Heisenberg's matrix method shows that it is not so phenomenological as it is sometimes claimed to be. Thus, when we treat the problem of the hydrogen atom by the matrix method, we assume that the atom is formed of a proton and an electron which attract each other

* M. Born. Problems of Atomic Dynamics. Massachusetts Institute of Technology; 1926; p. 70.

In the above passage Born uses the word "waves" in place of "radiations."

according to Coulomb's law of the inverse square. To this extent, invisible occurrences are postulated exactly as they were in Bohr's theory. In the second place, we find that although Heisenberg never introduces the position and momentum of the electron in his equations, he nevertheless appeals to magnitudes which play exactly the same rôle as do the mechanical magnitudes in Bohr's theory.* If, then, by a phenomenological theory we mean one which postulates no hidden occurrences, Heisenberg's matrix method cannot be said to be of this type. Besides, it is questionable whether all the magnitudes which Heisenberg introduces in his equations are truly observable. The phases of the vibrations, for instance, cannot be observed, and yet we cannot avoid considering them in the application of the theory. Finally, there is a last point which differentiates the matrix theory from the commonly-accepted standard of a phenomenological theory, such as thermodynamics. In thermodynamics the observed magnitudes are linked directly; but, as we mentioned on a previous page, in the new quantum theory the observed magnitudes are connected by an elaborate mathematical substructure of the most abstract type. It can be visualized only when we represent it in a space which has an infinite number of imaginary dimensions.

A question that might be asked is whether the matrix theory has been successful in disclosing the unknown mechanical laws of the subatomic world. The answer is in the affirmative, but at the same time it assumes an unexpected form, for the matrix theory and the other quantum theories lead to the conclusion that there are no exact microscopic mechanical laws; the true mechanical laws merely restrict uncertainties. One of the mistakes in Bohr's theory was to assume rigorous microscopic laws and then attempt to guess at them. The new theories also show that the mechanical restrictions which Bohr introduced in his theory of the atom are mere manifestations of the fundamental natural uncertainties discovered by Heisenberg.

At about the time Heisenberg was developing his new theory, de Broglie suggested a theory in which mysterious waves (not to be confused with waves of radiation) play a controlling part. De Broglie accepted the simultaneous presence of waves and particles; the office of the former was to guide the latter. In the hydrogen atom, de Broglie retained Bohr's mechanical model of an electron circling around a proton, but he did not rely on the classical mechanical laws to determine the motion; instead, he appealed to the controlling waves. The waves were regarded as physic-

* We refer to the matrices q and p and to the adoption of the classical Hamiltonian function of the problem.

ally real, so that de Broglie's original theory was a compromise between a mechanistic theory based on corpuscles and a theory involving continuity. Furthermore, since neither the electrons and their motions, nor the waves, are observable, we must conclude that de Broglie's theory is not a phenomenological one. Schrödinger developed some of the implications of de Broglie's waves and came to the conclusion that these waves must in many cases be supposed to be propagated in multidimensional spaces. This circumstance in itself prevents the waves from being physically real, as de Broglie had assumed them to be. Gradually, and in spite of de Broglie's resistance, the wave theory lost more and more of its physical attributes and became an abstract theory in which the substructure was entirely mathematical. Finally, Schrödinger's wave theory was proved to be mathematically equivalent to Heisenberg's matrix theory. In our opinion the fusion of these two theories, which were initially conceived in a totally different spirit, is of considerable importance. It shows that the individual preferences of the various investigators (whatever they may have been) have had no effect on the final results. The highly mathematical scheme to which both theories lead would thus appear to spring from the nature of things rather than from some passing vogue.

CHAPTER XII

A RETROSPECT ON THE DIFFERENT KINDS OF THEORIES

In the preceding chapters we have examined the more important features of mechanistic theories, of field theories, and of phenomenological theories. As for the modern quantum theories, they are best described by the qualification "mathematical." Jeans, commenting on the status of modern physical science, illustrates this idea in the words: "The Great Architect of the Universe now begins to appear as a pure mathematician."

A characteristic feature of the mechanistic theories, whether of the corpuscular or of the continuous type, is that they postulate underlying occurrences which are relatively simple to visualize. This feature is only natural since the occurrences postulated were suggested by phenomena which are familiar to us on the macroscopic level, for instance by swarms of particles or by elastic jellies. We cannot so easily form a mental picture of the field theories, especially when we are dealing with a field in 4-dimensional space-time. In particular, attempts to visualize the curvatures of space-time in the general theory of relativity would probably lead to incorrect conclusions. Nevertheless the difficulty we experience is due solely to the four dimensions of the warped continuum: if space-time were two-dimensional, its warpings could easily be imagined. The practical impossibility of representing the situation is therefore more one of degree than of essence.

In the phenomenological theories of the classical period, *e.g.*, in thermodynamics, the urge to visualize did not come into consideration because the equations of the theory expressed direct connections among the magnitudes observed. The situation is different in the modern quantum theories. Here we have an underlying mathematical scheme which we might wish to render more concrete by picturing it as the expression of physical occurrences. However, efforts in this direction have failed, and we must be content with mathematical symbols. It is true that quantum phenomena have been interpreted by means of waves and that the wave form of representation might seem to afford a more physical model. But the similarity of the underlying occurrences to waves is

deceptive, for these waves differ in too many respects from those we see moving over the surface of a pond.

Whatever may be said of the urge to visualize, we must agree that it is a natural psychological trait, issuing from the desire to interpret the unfamiliar in terms of the familiar. Any man, when faced by an unfamiliar situation, first tries to cope with it according to the familiar methods which have been of service to him on other occasions. He may subsequently discover the inefficacy of his measures, but only by testing them can he convince himself of this fact. Now in the physical world, mechanical categories are the ones with which we are best acquainted. We are surrounded by them on the commonplace level, and they cannot well escape our notice. Mechanics, moreover, was the first of the sciences to attain a considerable mathematical development, and so the earlier theoretical physicists, on looking around for hidden occurrences on which to base their theories, naturally selected mechanical ones. Indeed, no others were known before the days of Maxwell.

Today, for reasons which will appear presently, mechanistic theories are discredited. But Jeans and Eddington, in their popular writings (Jeans especially), convey the impression not only that mechanistic theories must be rejected in view of the discoveries in modern science, but also that these theories are totally wrong departures and never should have been considered. The claim is made that these mechanistic theories have led physicists astray, forcing them into a blind alley whence science has had to retrace its steps.

Jeans's main contention is undoubtedly correct, for there is little doubt that mechanistic theories are gone, never to return. But this does not necessarily imply that *a priori* cosmological beliefs were in any sense responsible for the mechanistic theories of the past, and that, had the earlier physicists been blessed with greater wisdom, theories of this type would never have arisen. Mechanistic theories found protracted favor simply because they seemed to account, with a high degree of accuracy, for the phenomena observed. In the kinetic theory of gases, for example, theory and experiment were in remarkable agreement in the days of Maxwell and Boltzmann. Only recently, thanks to the improvement in experimental technique, has the behavior of gases at exceedingly low temperatures been studied and have important discrepancies been observed. When these discrepancies, and also others relating to kindred mechanistic theories, were established beyond any reasonable doubt, mechanistic theories were immediately abandoned by an overwhelming majority of theoretical physicists without any particular sign of reluctance—manifest proof that no deep-seated beliefs were being discarded.

But there is also another aspect to the problem. The recent developments of the quantum theory show that as we near the commonplace level of ordinary experience, the mathematical scheme underlying the new quantum theories tends to merge into the mathematics derived from the classical mechanistic theories. This is but another way of saying that as we near the macroscopic level, the underlying occurrences may be represented by schemes that are mechanistic in the ordinary sense.* Indeed it is surprising to find how far the classical mechanical analogies may be accepted without involving us in perceptibly incorrect anticipations.

Compton for example, in his original interpretation of the effect which bears his name, assimilated the collision of a photon and electron to that of two perfectly elastic billiard balls. The Compton effect is no longer treated in this crude mechanistic way today; nevertheless, Compton's mechanistic interpretation gives very approximate results and it prompted him to further discoveries. The important rôle played by the mechanistic theories in the progress of science must also be borne in mind. The mechanistic methods developed by Boltzmann and by Gibbs in the classical kinetic theory of gases were transplanted into the new gas theories, so that to this extent the classical theory served as a model for the modern theories. It would be difficult to imagine how, in a single leap, science could possibly have passed from the macroscopic gas laws of Boyle and Charles to the extremely abstract mathematical schemes of the new gas theories. Similarly, the modern quantum theories of the atom could scarcely have arisen had it not been for Bohr's prior mechanistic theory.

From this analysis we conclude that the mechanistic theories of the past should be viewed as first approximations, and that subsequent discovery has served to refine but not altogether to destroy them. Furthermore, dogmatic assertions on the finality of the modern theories are out of place. For all we know, the modern theories in turn may be only approximations which will appear strangely crude twenty years hence. The field theories, in particular, cannot be the last word—they may be accepted only provisionally till a higher stage of approximation is reached. The progress of science must be viewed as the forging of a chain; each link has its place and cannot be suppressed without breaking the continuity of the chain; and the mechanistic theories represent one of these indispensable links.

Although we believe that Jeans's condemnation of the mechanistic

* The essence of this discovery is embodied in Bohr's Correspondence Principle.

theories is too severe, his stand becomes comprehensible when we take into consideration the general attitude of English physicists in the last century. The possibility of devising a familiar model of underlying physical occurrences was long regarded in England as an argument in favor of a theory. Kelvin once said that he could not understand a phenomenon unless he could devise a mechanical model of it. With these promptings, he proceeded to construct models of atoms by postulating vortices in the ether; the idea was suggested to him by Helmholtz's discovery of the great stability of vortex motions. Even more recently we find Larmor interpreting the electromagnetic field as due to motions of the continuous ether. Finally, at the Solvay Congress of 1911, the only two scientists to dissent strongly from Planck's quantum-interpretation of radiation, and to suggest in its place theories based on familiar models, were the two English representatives, Jeans and Lord Rayleigh.* But mechanistic interpretations did not find the same favor among the scientists of other countries. These thinkers leaned more towards analytical (mathematical) theories, and although viewing models as having a certain utility, claimed that on no account should the models be taken too seriously. Boltzmann expresses this view when he writes in the introduction to his Kinetic Theory of Gases:

"In England mechanistic ideas on the nature of heat, and on atomism in general, have been adhered to. But on the Continent the same is not true. . . ."

It is also instructive to examine Boltzmann's own attitude in connection with his kinetic theory. In the introduction previously mentioned he writes:

"Who can prophesy the future? Let all ways remain open therefore, and away with all dogmatic statements, either for or against atomistic ideas! Besides, when we present the theory of gases as a complex of *mechanical analogies,* we are in fact indicating, by the choice of this expression, how far we are from admitting, in any certain fashion and as a reality, that bodies are from all standpoints constituted by very small particles."

Elsewhere, discussing the significance to be attached to models, Boltzmann continues:

"The question of the utility of atomistic conceptions is of course quite independent of the fact on which Kirchhoff has insisted, that our theories bear the same relationship to Nature as symbols do to the entities they

* Lord Rayleigh could not attend the congress but wrote a letter supporting the stand taken by Jeans.

stand for, or as the letters of the alphabet bear to the voice, or as notes do to music. It is also independent of the question of ascertaining whether it would not be better to call these theories simple descriptions, so as to remind us always what their status is with regard to Nature.''

In a similar vein, Poincaré writes: ''Maxwell's theory is Maxwell's system of equations.'' Poincaré's statement implies that the mathematical relations embody what is essential in a theory and that the conceiving of a model does not contribute anything of importance. The reader is referred to Poincaré's chapter on ''Hypothesis''.

We have mentioned these points because, in our opinion, the revulsion against mechanistic theories heralded by Jeans and other English physicists arises from a belated recognition on the part of English science that mechanism is not the final solution.

CHAPTER XIII

PSYCHOLOGICAL DIFFERENCES AMONG PHYSICISTS

In the vernacular, the term physicist is applied both to the experimenter and to the theoretical investigator; and the awarders of the Nobel prize make no distinction between the two. Among the physicists to receive the prize in recent years are Chadwick, M. and Mme. Joliot, Anderson, Hess, Heisenberg, Dirac, Schrödinger, and de Broglie. The first five are experimenters, whereas the last four are purely theoretical men who have never contributed to the experimental field.

There are significant differences between the qualifications of these two types of investigators. The successful experimenter must be gifted with infinite patience, dexterity, and mechanical ingenuity in the devising and constructing of apparatus. On the other hand, he need have but little knowledge of mathematics. The theoretical investigator may be manually clumsy; a pencil and paper are his sole instruments. He must, however, be fully conversant with the experimental results obtained in all spheres of physics and must have a wide knowledge of those parts of mathematics which are of use to him in his work. His interest is primarily in coordinating results, if necessary by means of the introduction of auxiliary hypotheses. To the philosophic syntheses he creates, the name "natural philosophy" rightfully applies. As may be gathered, the qualifications necessary for the experimenter and the theoretical physicist differ so widely that very few men have been gifted and learned enough to distinguish themselves in both fields. Fresnel and Hertz were among the privileged few.

During the course of the present century the gap between the two groups of investigators has rapidly widened: on the one hand, successful experimentation is making ever-increasing demands on laboratory technique; on the other, the more recent physical theories make use of mathematical doctrines which are far from elementary. Thus, whereas we may study thermodynamics or even Newton's theory of gravitation with a minimum of mathematical knowledge, we could never expect to gain any profound insight into the theory of relativity or into the quantum theory without an extensive mathematical preparation demanding years of preliminary study. When it comes to determining which of the two types of investigators has contributed most to the advance of natural

philosophy, no answer can be given; theorists and experimenters supplement each other. Without the experimenters, or gatherers of facts, there would be no facts on which to speculate; and without the theoretical investigators, or amalgamators of facts, there would be no possibility of interpreting the various facts and of establishing their relations.

As may well be imagined, the experimenters as a whole evince little sympathy with the modern theories of mathematical physics. The advanced mathematics involved, and the impossibility of obtaining concrete pictures of the underlying occurrences, act as a deterrent to many of the laboratory workers. The layman who wishes to gain a general idea of the reactions of the foremost physicists to the modern theories should therefore be careful to ascertain whether the authority, whose opinions he is weighing, has acquired his reputation in the experimental or in the theoretical field.

But even if we restrict our attention to the theorists, we find sharp divergences. When confronted with two possibilities, some theorists will pursue the first alternative while the others will follow the second. We have encountered several illustrations in the preceding chapters. Kelvin favored mechanical models, whereas other physicists preferred the more mathematical theories. Planck and Einstein refuse to countenance the rejection of the causal doctrine, whereas Bohr, Heisenberg, and Dirac do not regard this doctrine as unassailable and experience no particular qualms in abandoning it. Similarly, absolute space was not repugnant in principle to Newton, Euler, Lagrange, Laplace, and many others including Planck among the moderns; whereas it was abhorrent to Mach. Action at a distance seems to have had no defenders among the leaders in natural philosophy, though the classical theory of gravitation accepted it implicitly. Finally, there have always been investigators who have preferred the phenomenological outlook of thermodynamics to the more speculative one which prevails in the kinetic theory of gases.

In this matter of psychological differences, it is of interest to note the marked monistic tendencies of Einstein. By his monistic tendencies, we mean his desire to interpret what appear to be different physical categories in terms of some common category. Of course, this tendency is general among natural philosophers, but it is more apparent in Einstein than in others: dualism seems to be definitely repugnant to him. For example, when discussing the coexistence of material corpuscles and of Newton's corpuscles of light, he writes:

"Moreover, the introduction into the theory of the two entirely different kinds of material particles that are necessary for the representation of ponderable matter and of light is in itself most unsatisfactory; and later

there appear the electric corpuscles as a third kind of particle, with again entirely different properties." *

In another passage where he discusses the coexistence of the mechanical and field categories in electromagnetics, he says:

"But this change led to a fundamental dualism which in the long run was insupportable." †

Einstein's formulation of the general theory of relativity in which he identifies inertial mass with gravitational mass is another case in point. So also is his attempt to rid the general theory of the dualistic conception it confers on the structure of space-time—a structure which appears to be part due to matter and part due to the nature of space-time itself. We may recall that it was in furthering this attempt that he imagined the cylindrical universe.

In this connection, Einstein writes:

"The physical properties of space would not then be wholly independent, that is, uninfluenced by matter, but in the main they would be, and only in small measure, conditioned by matter. Such a dualistic conception is even in itself not satisfactory." ‡

On the other hand, neither de Sitter nor Eddington have objected to the dualism to which Einstein refers. In his latest attempt to fuse the metrical and the electromagnetic fields, Einstein is furnishing another illustration of his monistic tendencies.

In some cases the conflict in the points of view of the various thinkers is striking. The following passage, taken from the records of the Solvay Congress of 1927, illustrates Lorentz's reactions to the new concepts of the quantum theory which were there under discussion:

"We wish to obtain a representation of phenomena and form an image of them in our minds. Till now, we have always attempted to form these images by means of the ordinary notions of time and space. These notions are perhaps innate; in any case they have been developed by our daily observations. For me, these notions are clear, and I confess that I am unable to gain any idea of physics without them. . . .

"For me, an electron is a corpuscle which at any given instant is situated at a determinate point of space, and if I believe that at the following instant this corpuscle is situated elsewhere, I attempt to imagine its path, which is a line in space. And if this electron meets an atom and

* James Clerk Maxwell; Cambridge, 1931; p. 68.

† Einstein. Sidelights on Relativity. E. P. Dutton and Co., 1920; p. 8.

‡ Einstein. The Meaning of Relativity. Princeton, 1923; p. 109.

penetrates in its interior and, after several adventures, leaves the atom, I attempt to construct a theory in which this electron has retained its individuality . . .

I would like to retain this ideal of other days and describe everything that occurs in the world in terms of clear pictures." *

Contrast this passage with Dirac's statement:

"The only object of theoretical physics is to calculate results that can be compared with experiment, and it is quite unnecessary that any satisfying description of the whole course of the phenomena should be given.†

The conflict between Lorentz and the quantum theorists brings to light an interesting point. It is the extraordinary rapidity with which science has developed in the last forty years. The theory of relativity was the work of two men in their twenties, Einstein and Minkowski. With the sole exception of Poincaré, no outstanding theoretical physicist past fifty years of age would countenance it. Those who did not oppose it outspokenly, complained that they were too told to be interested in such revolutionary ideas, and that the task should be entrusted to the younger generation. In those days Einstein was viewed as an extreme radical in science. As time passed, the theory of relativity was accepted, and today no natural philosopher of repute has expressed his dissent or suggested any modification of it (except in the most speculative part pertaining to the universe as a whole).

Then came quantum mechanics, also the work of men in their twenties, and history has repeated itself. Planck and Einstein, the extreme radicals of yesterday, now appear as conservatives, taking little part in the recent developments though opposing the new doctrines. It is not that Planck and Einstein have changed; it is merely that the new theories are so revolutionary that the supposedly revolutionary theories of thirty years ago (theory of relativity and earlier quantum theory) now appear to be almost classical by contrast. If the past is any guide to the future, we may safely predict that, thirty years hence, Heisenberg and Dirac, the extreme radicals of today, will in turn be regarded as conservatives and will resist the still newer doctrines of tomorrow.

On the whole, the psychological differences that we have noted among natural philosophers have not occasioned incompatible scientific developments. For unless the divergences relate to problems which are insoluble

* Electrons et Photons, p. 248. Solvay Congress of 1927.

† Dirac, The Principles of Quantum Mechanics, Oxford, 1930; p. 7.

(such as we suspect the doctrine of determinism to be), there eventually comes a time when new discoveries or some crucial experiment settle the dispute. No one conversant with modern science would suggest a return to the mechanistic theories; we doubt whether even Kelvin, if he were alive, would champion them.

Closely related to the personal preferences of the different investigators is the emphasis these men place on one or another of the various features that may characterize the same conceptual notion. For instance, a notion, such as absolute space, may under ordinary conditions be associated with different attributes *A, B,* and *C*. Suppose now the attribute *A* is lacking. The writer who feels that *A* is less important than *B* or *C* may not deem it necessary to change his terminology and may still speak of absolute space. Another writer who views the attribute *A* as essential will hold that, in its absence, space can no longer be called absolute. For those who are in a position to follow the mathematical explanations which accompany the presentation of a physical theory, differences of terminology are of minor importance, because the meaning the writer wishes to ascribe to his words is sufficiently clear. But the layman and the metaphysician, who may be interested in science and who must rely on popular books, are likely to be misled by seemingly contradictory statements made by the various theoretical physicists.

The theory of relativity, even more so than the quantum theory, appears to have provoked much confusion in this connection. For lack of space, we shall not discuss exhaustively the various meanings that are attributed to such words as relative, absolute, force, real, fundamental. However, as an illustration of the confusions that may occur, we shall examine some of the seemingly contradictory statements that have been made by leading physicists on the nature of 4-dimensional space-time. All theoretical investigators agree that the 4-dimensional space-time of the special theory of relativity is a four-dimensional amplification of Newton's absolute space. But whereas Weyl views the space-time of the general theory as deprived of all absolute features, Planck still stresses its absolute nature. Einstein's attitude is somewhat different and we shall examine it later. A few quotations will clarify the situation. We start with the space-time of the special theory. Einstein writes:

"The principle of inertia, in particular, seems to compel us to ascribe physically objective properties to the space-time continuum. Just as it was necessary from the Newtonian standpoint to make both the statements, *tempus est absolutum, spatium est absolutum,* so from the standpoint of the special theory of relativity we must say, *continuum spatii et*

temporis est absolutum. In this latter statement *absolutum* means not only "physically real," but also, "independent in its physical properties, having a physical effect, but not itself influenced by physical conditions." *

Weyl expresses himself in equivalent terms. Thus he writes:

"Where do the centrifrugal and other inertial forces take their origin? Newton's answer was: in absolute space. The answer given by the special theory of relativity does not differ essentially from that of Newton. It recognizes as the source of these forces the metrical structure of the world and considers this structure as a formal property of the world." †

We now pass to the space-time of the general theory. To quote Weyl:

"The falling over of glasses in a dining-car that is passing round a sharp curve and the bursting of a fly-wheel in rapid rotation are not, according to the view just expressed, effects of 'an absolute rotation' as Newton would state but whose existence we deny; they are effects of the 'metrical field' or rather of the affine relation associated with it." ‡

The significance of this statement is to rule out absolute rotation and thereby the absolute continuum with respect to which rotation would acquire a meaning. But let us see what Planck has to say. Planck, commenting on the space-time of the general theory of relativity, writes:

"If from space and time we should take away the concept of the absolute, this does not mean that the absolute is thereby banished out of existence, but rather that it is referred back to something more fundamental. As a matter of fact, this more fundamental thing is the four-dimensional manifold which is constituted by the welding together of time and space into a single continuum. Here the standard of reference and measurement is independent of the observer." ¶

Weyl and Planck appear to be at loggerheads; and yet, when the mathematics of the theory is understood, we find that Weyl and Planck are merely stressing different aspects of the same situation. The fact is that the attributes of an absolute continuum are varied. We leave out of account the mystical significance attributed to absolute space in the writings of Henry More and other metaphysicians, for it has nothing in common with the situation we are here considering. The features of

---

* The Meaning of Relativity, p. 61.

† H. Weyl. Space-Time-Matter. Methuen, London; p. 218.

‡ Space-Time-Matter, p. 221.

¶ Planck. Where Is Science Going? W. W. Norton and Co., New York, 1932; p. 197.

an absolute continuum (such as Newton's absolute space or Lorentz's stagnant ether) that are relevant in the present discussion are:

A. Its ability to exert physical effects.
B. The rigidity of its structure; and by this we mean the impossibility of our influencing its structure in any way through the action of matter or of other physical agents.
C. The existence of geometric properties which transcend the frame of reference or the point of view of the observer who surveys them.

The attribute *A* does not appear to reflect so much on the absolute nature of the continuum as on its physical existence. A continuum which would exert no physical effects would be a mere void having the sole attribute of extension; it would fail to manifest itself physically, and there would be no particular reason to postulate its existence in any physical sense. In Newton's absolute space, in Lorentz's stagnant ether, and in the 4-dimensional space-time of the special theory of relativity, the attribute *A* is illustrated by the forces of inertia that the continuum exerts on any accelerated body.

No particular comment is required on the attribute *B*; its significance is clear.

The significance of the attribute *C* is best understood by an illustration. The length of a shadow cast on a plane varies with the inclination given to the plane. On the other hand, the distance between two points in space is determined solely by the geometry or structure of space; it thus betrays an intrinsic property which a change in our frame of reference cannot affect. It seems trivial to mention this feature in the case of Newton's space, for it appears obvious. But Minkowski's discovery of the property of invariance for the distance between two points in 4-dimensional space-time was very important. By welding together space and time, it gave reality to the 4-dimensional continuum. The 4-dimensional continuum thus became a physical existent endowed with a structure transcending the contingencies of space and of time, that result from the varying point of view of the observer. The geodesics of space-time, in particular, appear as intrinsic features of the continuum.

Since the space-time of the special theory of relativity has the three attributes mentioned, no question has arisen as to its absoluteness. Einstein, however, objects to the simultaneous presence of the attributes *A* and *B*, and he expresses the belief that the attribute *B* should be discarded. To quote Einstein:

"In the first place, it is contrary to the mode of thinking in science to conceive of a thing (the space-time continuum) which acts itself, but which cannot be acted upon." *

* The Meaning of Relativity, p. 62.

Einstein postulates therefore in his general theory of relativity not only that the space-time structure acts on matter, as had previously been assumed, but that, conversely, matter must exert an action on the structure of space-time. The attribute B is thus withdrawn; the attributes A and C alone remain. Weyl, who views the attribute B as the one truly characteristic of an absolute continuum, is accordingly consistent in his contention that space-time has lost its status as an absolute; whereas Planck, to whom the attribute C appears the most important, is justified in holding that space-time is still absolute.

Lastly, we come to Einstein's contention. He expresses the opinion that the space-time of the general theory, having lost the attribute B while retaining the attributes A and C, is a hybrid concept which is neither absolute nor relative. Einstein's monistic tendencies urge him to remove this dualism by withdrawing also the attributes A and C. We must understand exactly what this implies. In the general theory the structure of space-time is affected by the proximity of matter, but, even at an infinite distance from matter or in a universe totally devoid of matter, a definite structure still remains—the flat structure of the special theory. Suppose, then, we assume the world to be empty of matter, a small test body alone being present. The flat structure will manifest itself by acting on the test body, and the forces to which the body is submitted can no longer be attributed to the action of other masses. Thus, in an empty world, space-time would still retain the attributes A and C.

Commenting on this dualism of the space-time of the general theory, Einstein, in a passage previously quoted, writes:

"The physical properties of space would not then be wholly independent, that is, uninfluenced by matter, but in the main they would be, and only in small measure, conditioned by matter. Such a dualistic conception is even in itself not satisfactory."

If we are to remove these attributes, as Einstein here aspires to do, we must assume that, in a world empty of matter, space-time would lose all trace of a structure. Matter would then be the sole cause of the structure of space-time, and space-time in the absence of matter would lose all its attributes and would cease to have any physical existence. It was to further this idea that Einstein devised his theory of the cylindrical universe.

The analysis we have given of the different aspects of an absolute continuum is typical of many concepts that might be mentioned. In mathematics also we find examples of a concept which has various attributes and for which the popular terminology is retained even when

the more familiar attributes are lacking. The concept of perpendicularity is a case in point. We all have a clear understanding of what is meant by two perpendicular lines; and we should deem it absurd to speak of a line that is perpendicular to itself. The mathematician, however, is not concerned with the visual aspect of perpendicularity. His interest lies in the mathematical relation that is satisfied by perpendicular lines; and so, from his standpoint, the realization of this relation constitutes the true characteristic of perpendicularity. Now there is a certain kind of line, called a minimal line or a null-line, for which the above relation is satisfied, not in connection with a second line, but in connection with the null-line itself. The mathematicians therefore feel justified in saying that a null-line is perpendicular to itself. Null-lines play an important part in Minkowski's space-time; they define the world-lines of particles (photons) moving with the velocity of light.

PART II

PHYSICAL THEORIES OF THE CLASSICAL PERIOD

CHAPTER XIV

MATHEMATICAL PRELIMINARIES

(FUNCTIONS AND GROUPS)

IN THE previous pages we mentioned that the progress of physical theories is dependent on the mathematical development of the day. Thus the general theory of relativity could never have arisen had it not been for the prior discoveries of Gauss and Riemann in non-Euclidean geometry; and almost all theories draw upon the calculus. For these reasons a few general considerations on various aspects of mathematics may not be out of place.

The Greeks were highly proficient in geometry, furnishing mathematical geniuses of the first rank, notably Euclid, Appolonius, and Archimedes. But ignorant as they were of the more powerful methods of analytical geometry and the calculus, their demonstrations were laborious, and the majority of the theorems they discovered can be established today by much simpler means. In point of fact analytical geometry and the infinitesimal methods were both foreshadowed by Archimedes, but the lack of a simple notation and other causes precluded the great Syracusan from deriving full benefit from his ideas. Not until the seventeenth century did these two great instruments of research become available; and, thanks to them, the scope of mathematical investigations was extended to a degree which would have seemed unbelievable to the ancients.

The outstanding mathematician of the scientific Renaissance in the seventeenth century was Newton. The eighteenth century produced three giants, Euler, Lagrange, and Laplace. The nineteenth century witnessed a much larger number of masters. In the chronological order of their birth, the great mathematicians of the nineteenth century were Gauss, Cauchy, Abel, Jacobi, Hamilton, Weierstrass, Hermite, Riemann, Sophus Lie, and Poincaré. To this list we might add Poisson and Cayley. A host of other mathematicians, such as Fourier, Dirichlet, Liouville, also made discoveries and evolved methods of the highest importance; and indeed these mathematicians, taken as a whole, have probably contributed more than any single one of the great masters.

The chronological order in which the names of the masters has been written is, very approximately, the order in which the student of mathematics will come across their discoveries. Thus Euclid is the first of the great mathematicians whose theorems we study; whereas it would require many years of hard work before we could become conversant with the contributions of Poincaré, or even understand in what they consist. The fact is that each successive creative mathematician stands, so to speak, on the shoulders of his predecessors, utilizing their discoveries.

In a broad way we may classify mathematical advances under two headings. Firstly, new methods are discovered, methods which render possible the solving of problems formerly insoluble and which open up new avenues of research. Secondly, considerable advances may ensue from the application of known methods. Analytical geometry, the calculus, the theory of groups, vector analysis, and tensor analysis illustrate new methods. Newton's deduction of the law of the inverse square from Kepler's laws is an example of the successful application of a known method to a particular problem.

We might suppose that the mathematician who originates a new method deserves greater credit than the one who applies it. But this is not necessarily so, for the application of a known method may make greater demands on the mathematical ingenuity of the investigator than may the bare discovery of the method. Fourier, for instance, constructed the celebrated series which bear his name, and the importance of these series is today so considerable that we could scarcely do without them. Yet no mathematician would think of classing Fourier with the greatest masters. Luck may also play a part in the fame a new method or theory receives. Of two methods equally deep, one may prove relatively sterile whereas the other has a number of applications. Hadamard, commenting on the theory of integral equations developed by Volterra, Fredholm, and Hilbert, writes:

"The theory of integral equations, born yesterday, is already classical. It has been introduced in several university courses. There is no doubt—perhaps after further improvements—that it will soon impose itself as of current use in mathematics. This is a rare piece of good fortune for a mathematical doctrine, for mathematical doctrines so often become museum exhibits."

The great masters have been prolific not only in the application of known methods, but also in the discovery of new ones. The caliber of a mathematician is usually determined by the volume of his productions and by the diversity of the subjects he has treated. This statement may

seem strange when we recall that many of the masters have compared their science to an art; and an artist is not gauged by the quantity of his productions, but by their quality. The paradox is attenuated, however, when we realize that there can be no such thing as mediocre mathematics. A mathematical demonstration is either right or wrong, and if it is wrong, it is not mathematics and finds no place in a mathematical publication.* To be sure, there may be a slip in a demonstration, even when furnished by a master like Riemann; but the slip is eventually detected and remedied, or else, if it vitiates the whole, the demonstration is condemned and forgotten. On the other hand, it would be misleading to say that, in mathematics, quality plays no part at all. In many cases a theorem may originally be demonstrated in a cumbersome way and then later receive a more simple demonstration characterized by that subtle quality which is referred to as "mathematical elegance." Nevertheless in such cases, the theorem is usually named after the original discoverer; to him is given credit for the discovery rather than to the investigator who has simplified the demonstration.

Pure and Applied Mathematics—A distinction is usually made between pure and applied mathematics. According to the prevalent classification, the subject matter underlying a mathematical doctrine is assumed to establish the distinction. When this subject matter pertains to the physical world, the mathematics is said to be of the applied variety; and when it is selected without regard for the world of our physical experience, the mathematics is called pure. On this basis theoretical physics and mechanics must be regarded as branches of applied mathematics, whereas the theory of functions and the theory of groups are representative of pure mathematics. Euclidean geometry was at one time classed with pure mathematics because its basic postulates were thought to express requirements of the human mind. But the discovery of the non-Euclidean geometries has since dispelled this belief; it has shown that the Euclidean postulates are mere idealizations of the results of our physical experience.

The accepted distinction between pure and applied mathematics is far from satisfying. In the first place, it permits no permanent classification. As an example let us consider the doctrine of classical mechanics. When the basic postulates of classical mechanics were estalished by

* We are not referring here to the discussions on the nature of mathematics and of mathematical reasoning. In this latter sphere of investigation, appropriately christened "metamathematics," it seems to be a case of every man for himself.

Galileo and Newton, they were thought to express physical characteristics of the world. Classical mechanics was thus regarded as a branch of applied mathematics. But today, as a result of the theory of relativity, we know that the classical postulates do not correspond to physical reality. Strictly speaking, we should therefore reverse our former stand and view classical mechanics as an abstract doctrine pertaining to pure mathematics. From this illustration we see that the prevalent classification of the mathematical doctrines can never aspire to any degree of permanency since it may be modified by every new discovery.

More important than these objections is the fact that the prevalent classification pays no heed to one of the essential characteristics of mathematics: namely, in a mathematical doctrine the subject matter is of no importance. What the mathematician is concerned with are systems of relations, and his aim is to extract from given relations their necessary consequences. In celestial mechanics, for example, we may be concerned with determining how two or more bodies will move under their mutual gravitational attractions. The physical part of the problem is exhausted when we have utilized the mechanical postulates and have transcribed our physical problem into the mathematical form of a system of differential equations. The answer to the physical problem is then obtained if we can solve these equations. In this attempt the physical significance of the equations may be lost sight of entirely; the difficulties in our task become purely mathematical, and they are not dependent upon the fact that the equations may or may not be the mathematical expressions of physical occurrences. The purely mathematical difficulties in obtaining the solutions of seemingly simple mechanical problems are often so great that they have challenged the efforts of the foremost mathematicians. Many of the most abstract methods in mathematical research have been devised for the purpose of obtaining these solutions; and we find that the men who have contributed most to the development of mechanics and celestial mechanics are the same masters who have made some of the greatest discoveries in pure mathematics.

In our opinion the distinction between pure and applied mathematics should rest not on the subject matter of a mathematical doctrine, but on the method of procedure of the investigator. The true difference, according to this view, must be sought between pure and applied mathematicians rather than between pure and applied mathematics; and the name "applied mathematician" should be given to the investigator who utilizes established mathematical results without perhaps knowing, or even caring, how they are arrived at. For instance an applied mathematician may, in the course of his work, be confronted with some differential equa-

tion which he does not know how to solve. He turns to a textbook or calls upon some pure mathematician and asks him to solve the equation. His sole interest lies in the cut and dried solution and not in the complicated theory that may be involved in obtaining the solution.

Usually, theoretical physicists are of this type or, at least, they used to be until recent years. But in the new physical theories, some of the mathematical problems raised are of a novel kind, and they cannot be solved by the classical methods. For this reason the modern theoretical physicist has been thrown back on his own resources and, under the force of circumstances, has been compelled to become a pure mathematician. Nevertheless, in the majority of cases, the requirements of the theoretical physicist even today are not those of the pure mathematician. As an example, let us examine the most celebrated equation of wave mechanics, *i.e.*, Schrödinger's equation of the hydrogen atom.

Schrödinger sought to solve this differential equation. He proceeded along standard lines by splitting the equation into two simpler ones. The first of these is a classic to mathematicians and is solved by means of functions introduced by Laplace. The second equation is of a standard form investigated by Fuchs in the last century. Schrödinger solved it by applying a method devised by Laplace. Subsequently, he noticed that the solution of this equation could be obtained more simply by solving another equation previously studied by the mathematician Laguerre. The solution of Laguerre's equation being well known, the problem of the hydrogen atom was solved automatically. From this account we see that the solution of Schrödinger's wave equation did not involve any new mathematical discoveries, but was arrived at in a routine way by a direct application of the methods previously established by the pure mathematicians.

The foregoing analysis is not meant to disparage Schrödinger's mathematical ability; it merely proposes to show that the contribution of the mathematical physicist, as such, resides not in the solution of purely mathematical problems, but in the construction of physical theories which enable him to transcribe physical processes and phenomena into the form of mathematical equations.

Functions of a Real Variable

Functions—The Greeks, in addition to their study of conics (on which they stumbled when considering the plane intersections of a circular cone), invented a number of other curves: the cissoid, the strophoid, the cycloid, epicycloids, spirals. Many of these curves were introduced

for the purpose of solving some particular geometrical problem, *e.g.*, the trisection of an angle. Then, when in the seventeenth century analytical geometry was invented by Descartes, an infinite variety of new curves was brought to the attention of mathematicians. At a still later date, these curves, which had been accepted as mathematical by the investigators of the seventeenth and eighteenth centuries, were found to form only an insignificant minority among the mathematical curves. It is with the evolution of the concept of a mathematical curve that we shall be concerned in the first part of this chapter.

We must first examine what is meant by a function. In a broad sense, if one thing depends on other things, the former is said to be a function of the latter. Thus the growth of a plant is a function of the nature of the soil, of its moisture content, and of the amount of sunshine the plant receives. The dependences here involved are, however, vague and ill-defined; and only when the relationships are precise and susceptible of numerical expression are we properly dealing with a mathematical function. Nevertheless it is certain that man was acquainted with the concept of function long before this concept was introduced explicitly into mathematics during the seventeenth century. Thus the notion of function, like so many others in mathematics, is plainly taken over from the world of common experience.

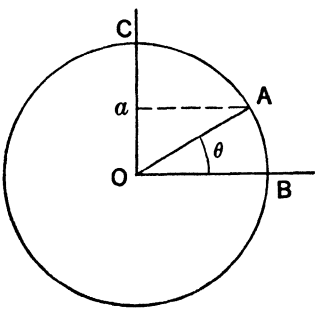


FIG. 1

Geometry affords many illustrations of functions. As an example, consider a circle. Draw any radius OA making any arbitrary angle θ with the horizontal OB.

Project OA on the vertical OC. We thus obtain the length Oa. The magnitude of this length will depend on the value of the angle θ; and to each angle θ will correspond a definite length Oa. We say that the length Oa is a function of the angle θ. This relationship is perfectly definite; there is nothing vague about it; and so a definite functional relationship is determined. Symbolically, the function is written sin θ, and the functional relationship is expressed by

$$\text{length } Oa = \sin\theta \times \text{length } OA.$$

Quite generally, if we call y the magnitude which depends on the value of x, and if we wish merely to express the existence of a functional relationship without having to know its exact nature, we write:

$$y = f(x).$$

Under this form the exact functional relationship is left unspecified. The magnitude y, whose value is determined by the value ascribed to x, is called the *dependent variable*; whereas x is called the *independent variable*. It may happen that y depends on more than one independent variable, *e.g.*, on three, which we may represent by x, u, v; the functional relationship would then be written

$$y = f(x, u, v).$$

A physical illustration of this more complicated type of function is given by the dependence of the apparent height of the sun on the time of the day, the season of the year, and the latitude of the observing station. In what follows, we shall confine ourselves to functions of a single independent variable.

A graphical representation of a function of one variable is obtained when we follow the Cartesian method of plotting the value ascribed to x along a horizontal and the corresponding value of y along a vertical. When a given value is ascribed to x, we thus obtain a point P, as is seen in the figure. If now we repeat this operation for all values of x in the range considered, *e.g.*, from $-\infty$ to $+\infty$, we obtain an infinite aggregate of points in the plane. This aggregate of points gives a graph, or geometrical representation, of the function. If the points determine a continuous curve, the function is said to be continuous.

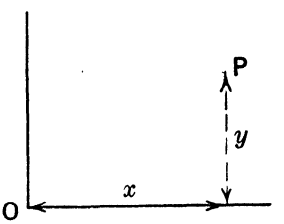


FIG. 2

Next we must examine how a function will be defined in mathematics. According to the modern conception, y is viewed as a known function of x when we have stated clearly the quantitative law of correspondence between the two variables. The law may be expressed purely verbally, according to our fancy; it need not be defined by known mathematical operations. But the mathematicians of the seventeenth and eighteenth centuries were more exacting. The only functions which they accepted as susceptible of mathematical study were those which could be defined by means of well-known mathematical operations. We shall refer to such functions as *mathematical functions*. All functions which did not satisfy this condition were regarded as beyond the scope of mathematics. For example a function such as

$$y = \frac{\sqrt{x^2 + 1}}{x^3 - 4} - x + 2$$

was accepted as mathematical, because the expression on the right involves mere algebraic operations. We need scarcely mention that the curve which represents this function is obtained when we make x run over a continuous range of values; then calculate the corresponding values of y; and, finally, proceed according to the Cartesian method. The curve is a continuous one. Similarly, the relationship expressed by

$$y = \sin x,$$

since it can be determined by a geometric construction, was also accepted as defining a mathematical function. The curve obtained is called a sinusoid

Since a continuous curve was associated with each of the continuous functions that could be imagined, and since these functions were legion, the mathematicians of the seventeenth century were conversant with a vast collection of continuous curves. Among these, all the curves known to the Greeks readily found a place. For instance,

$$y = x^2 \text{ and } y = \pm\sqrt{R^2 - x^2}$$

represent, respectively, a parabola and a circle of radius R. It was impossible to coin a name for each one of the new curves which any mathematician could invent, and so the majority of curves went unnamed. Some few, however, received names, either on account of their simple geometric properties or by reason of their connection with the solution of some conspicuous problem. Thus we have the oval and the folium (Descartes), the lemniscate (Bernoulli), Pascal's curve, and at a later date Gauss's curve of errors.

Towards the end of the seventeenth century, infinite series were studied by Newton and Leibnitz, and, by utilizing these series to define new functions, a considerable extension was given to the class of continuous curves accepted as mathematical. This leads to a digression on series.

Series—Suppose we take the algebraic sum of a billion finite numbers. We assume that the numbers are positive or negative, or that some are positive and others negative. The sum is obviously finite. But suppose that in place of a billion numbers, we consider an infinite succession of numbers. We now have an *infinite series*. We cannot speak of the *sum* of an infinite series, since infinity is never exhausted; but we may consider a magnitude which plays the same rôle with respect to an infinite series as a sum does to a finite succession of terms. This magnitude is the one towards which the sum of the first n terms of the infinite series

tends when n is made to increase indefinitely. It is called the *limit* of the infinite series.

Important differences distinguish the sum of a finite number of numbers from the limit of an infinite number of numbers. Let us mention a few. We know that the sum of a billion numbers, some of which are positive and others negative, has exactly the same value whether we effect the summation in the order given or in any other order. Yet, as was proved by Dirichlet in the nineteenth century, this conclusion is not necessarily correct when we are dealing with infinite series: the limit may have one value or another according to the way in which the terms are ordered. Here is a first important distinction between the sum of a finite number of terms and the limit of an infinite number of terms. But there are others. Thus the sum of a finite number of numbers is always finite and, as just explained, is perfectly well determined. On the other hand, the limit of an infinite series may be finite or infinite, and sometimes the series has no limit. In the two latter cases the series is said to *diverge*. We may illustrate these various situations by elementary examples. If we add the number 1 a billion times in succession, we obtain a perfectly definite finite sum, *i.e.*, one billion. But if we have the unending series

$$1+1+1+1+1+\ldots\ldots,$$

the limit is infinite—the series diverges. Similarly, if we have a finite sequence of terms such as

$$1-1+1-1+1, \text{ or } 1-1+1-1+1-1,$$

the sum will have the value 1 or 0 according to whether the number of terms is odd or even. But if we are dealing with an infinite series of this type, there is no definite limit, for we have a kind of oscillation between 1 and 0—the series diverges.

As an illustration of a series the limit of which is well defined and finite, we may mention the series

$$1+\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\ldots\ldots.$$

Its limit is 2. Such a series is said to *converge*. The series

$$1-\frac{1}{2}+\frac{1}{3}-\frac{1}{4}+\frac{1}{5}-\ldots\ldots$$

also converges, but here we must be careful not to change the order in which the terms are written. If we change the order of the terms, we

may find that the series diverges. These brief illustrations on numerical series will suffice for our purpose. What we shall now be concerned with are series of functions.

Series of functions differ from numerical series in that the successive terms are functions of a magnitude x whose precise numerical value is not specified. Thus the infinite series

$$f_1(x) + f_2(x) + f_3(x) + \ldots\ldots,$$

where $f_1(x)$, $f_2(x)$, $f_3(x)$, . . . are known functions of x, represents a series of functions. Let us assign the same arbitrary numerical value to x in each one of these functions. The functions now define numbers, and our series of functions becomes a numerical series. By changing the value ascribed to x, we obtain one numerical series after another. In most cases the numerical series converges if certain values of x are substituted, whereas it diverges (*i.e.*, has no limit or becomes infinite) when other values of x are taken. Suppose for argument's sake that the series converges for all values of x between 0 and 1, and diverges for all other values of x. The limit of the series then has a well-determined value only when x is situated between 0 and 1, and usually this value will depend on the value assigned to x in the interval. The limit of the series thus defines a function of x. But note that it is only when x is situated between 0 and 1 (where the series converges) that the series defines a function; elsewhere the series is meaningless and defines nothing.

We thus realize the importance of determining for what range of x (if any) a series of functions converges. At the same time we see that if a series converges for a certain range of x, the series *via* its limit enables us to define a function, of which we may obtain a graphical representation by Descartes' method. This introduction of series as a means of defining functions enables us to obtain new functions and therefore new curves. The problem of determining for what range of x any given series of functions converges is one which even today is often of extreme difficulty. The mathematicians of the seventeenth and eighteenth centuries, handicapped as they were by their ignorance of the properties of infinite series, were restricted to the simplest kinds of series. Such series were the elementary geometric progressions and the more general *power series* or *Taylor series.*

Now the functions defined by Taylor series are functions of a very special kind. These functions, today named *analytic functions*, form but a very small sub-class among the functions that can be defined by other types of series and by other kinds of mathematical expressions. But this

fact was not realized by the mathematicians of the seventeenth and eighteenth centuries; and so they made no distinction between mathematical functions and analytic ones. They erroneously believed that any mathematical function (any function that could be defined by a mathematical expression) was necessarily analytic, *i.e.*, susceptible of being expressed by means of a Taylor series.* The mistake made by the earlier mathematicians is comprehensible when we are told that all those mathematical functions whose properties they were able to investigate happened to be analytic.

Taylor Series—In view of the importance of Taylor series in the historical development of mathematical analysis, it will be advantageous to examine their principal characteristics. Consider the two following infinite series:

$$(1) \qquad a_0 + a_1x + a_2x^2 + a_3x^3 + \ldots\ldots$$

and

$$(2) \quad a_0 + a_1(x-k) + a_2(x-k)^2 + a_3(x-k)^3 + \ldots\ldots,$$

in which a_0, a_1, a_2, a_3 . . . . . and k are arbitrary, fixed finite numbers, and x is a variable. It will be noted that both series contain the successive integral powers of $(x-k)$, the only difference between the two series is that, in the first, k is assumed to have the value zero. Series of this kind are called "power series."

Now a power series, such as (2), always converges for the value $x=k$, since it then reduces to its first term, a_0. If the power series diverges for all other values of x, it merely defines a point, *i.e.*, the point $x=k$. But the power series may also converge for some other value of x. Mathematical analysis shows that in this case the series will necessarily converge for a continuous range of values of x and that the point $x=k$ will occupy the centre of the range. Within this range of convergence (which may be limited or may comprise all values of x), the power series defines a function of x which can be shown to be a continuous function, all of whose derivatives are also continuous.

---

* It was in this connection that the appellation "analytic" originally arose; it was regarded as synonymous for mathematical. Subsequently, when it became apparent that not all functions that could be expressed mathematically were susceptible of being expressed by means of Taylor series, the word "analytic" was retained only for the more restricted type of mathematical functions.

A power series of this last kind constitutes a Taylor series,* so that a continuous function defined by a power series is what we have called an analytic function of x.

As mentioned in a previous paragraph, all the functions known to the mathematicians of the eighteenth century happened to be analytic; and we have said that this fact prompted the idea that all mathematical functions were analytic. An example may now be given. The function

$$y = \sin x$$

expresses a relation between x and y, which is defined by means of a geometric construction. Such a function is therefore mathematical. But it was soon discovered that this geometric function could be expressed equivalently by means of the Taylor series

$$y = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots\ldots,$$

and that as a result the function was analytic. In all other cases examined by the mathematicians of the seventeenth and eighteenth centuries, the same situation was found to occur.

Analytic Curves—If we plot, according to the Cartesian method, the x and y of a function defined by a Taylor series (*i.e.*, an analytic function), we obtain a curve which we also call "analytic." We have said that the function defined by a Taylor series in the interval of convergence is always a continuous function with all its derivatives continuous. This implies that when we plot the function by Descartes' method, the analytic curve obtained will be continuous and will have no sharp bends in it, and that the radius of curvature will also vary continuously as we move along the curve. In a crude way, we may say that analytic curves are smooth curves. But there are other important features to analytic curves; these features are elucidated when we revert to the Taylor series that define the curves.

As a first illustration, let us examine the trigonometric function

$$y = \sin x. \tag{3}$$

* From this account it would appear that a Taylor series is always an infinite series. As a special case, however, we may suppose that the series contains only a limited number of terms. In this case it reduces to a polynominal and it converges for all finite values of x.

Let us construct the curve by plotting along a horizontal the values we ascribe to x, and along a vertical the corresponding values for y. We obtain a snake-like curve (Figure 3) extending from $-\infty$ on the left

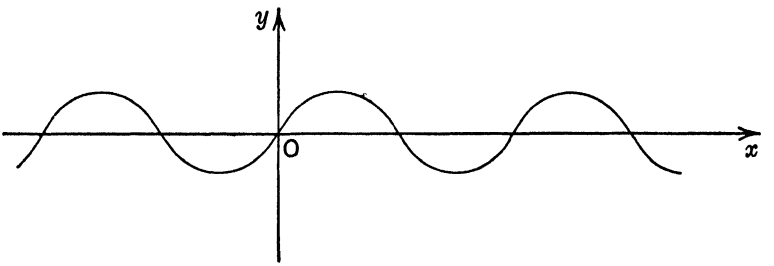


FIG. 3

to $+\infty$ on the right; the curve is called a sinusoid. We have seen that this function may be expressed by the Taylor series

$$y = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots\ldots, \tag{4}$$

and that this fact proves the function, or curve, to be analytic. Now the series just written converges for all values of x; hence we conclude that the series also represents the curve over its entire domain of existence, from $-\infty$ to $+\infty$. The two expressions (3) and (4) of the function are equivalent in all respects. By this we mean that if in (3) and in (4) we give the same current value to x, exactly the same value is obtained for y in the two cases. The Taylor series just written is not the only one that can represent the function; an infinity of such series can be obtained. For instance,

$$y = \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}}\left(x - \frac{\pi}{4}\right) - \frac{1}{2\sqrt{2}}\left(x - \frac{\pi}{4}\right)^2 - \frac{1}{6\sqrt{2}}\left(x - \frac{\pi}{4}\right)^3 + \ldots\ldots \tag{5}$$

is a Taylor series in power of $\left(x - \frac{\pi}{4}\right)$, which represents the same function. Like the series (4) it converges for all values of x.

Let us pass to a second illustration. The function

$$y = \frac{1}{x^2 + 1} \tag{6}$$

is analytic, as is obvious since it is defined by mere algebraic operations. If we construct the curve by giving all values to x and calculating from (6) the corresponding values for y, we obtain a curve (Figure 4) extending from $-\infty$ to $+\infty$.

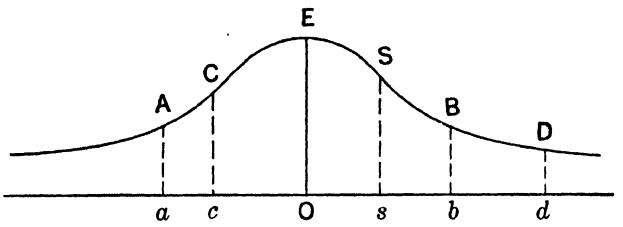


FIG. 4

Since the function, or curve, is analytic, we may also express it by means of a Taylor series. Two such series are

$$y = 1 - x^2 + x^4 - x^6 + x^8 - \ldots \tag{7}$$

and

$$y = \frac{4}{5} - \frac{16}{25}\left(x - \frac{1}{2}\right) - \frac{16}{125}\left(x - \frac{1}{2}\right)^2 + \frac{384}{625}\left(x - \frac{1}{2}\right)^3 - \ldots \tag{8}$$

Both of them represent the curve; but there is an important difference between the properties of these two Taylor series and those of the series (4) and (5), that we considered in the previous example. The two series (4) and (5) converged for all values of x and therefore represented the entire curve. With the two series (7) and (8) a new situation arises, for these series converge only when x is situated in certain ranges. Thus the series (7) converges only when x is taken between -1 and $+1$ (*i.e.*, from a to b in the figure); whereas the series (8) converges only when x is within the interval defined by the two values $\frac{1-\sqrt{5}}{2}$ and $\frac{1+\sqrt{5}}{2}$. The second interval of convergence is marked by the points c and d in the figure (Figure 4). Owing to the peculiarities just mentioned, the first series represents only the arc AB of the curve, and the second series represents only the arc CD. Neither series defines the whole curve; hence neither series is completely equivalent to the function defined by (6). Incidentally, we note that our two series are in powers of $(x - k)$, where k

has the values $k = 0$ and $k = \frac{1}{2}$ respectively, and that in either case the point $x = k$ is at the midpoint of the interval of convergence. (In the figure these two centres of convergence are represented by the points O and s respectively). The positions of the two points $x = k$, at the centres of the intervals of convergence in the two foregoing Taylor series, illustrate the general rule mentioned on a previous page.

We may wonder why these Taylor series do not always converge. In particular, why does not the first series converge to the right of b? The point b does not appear to constitute any fundamental obstacle, since the second series converges beyond it. These questions could not be answered by the mathematicians of the eighteenth century. Not until the complex variable was systematically introduced by Cauchy, was the reason for the strange breaks in the convergence fully understood. We shall revert to this problem on a later page. For the present we wish to examine a different point.

Let us suppose that we are not given the equation (6) of the curve or any of the series which define stretches of the curve. We shall assume, however, that we are informed with mathematical precision of the lay of a small arc of the analytic curve—an arc which may be as tiny as we choose. Mathematical analysis then shows that it will be possible to construct the Taylor series which defines the curve over equal stretches to the right and left of the given arc. To take a definite example, let us assume that the tiny arc is situated at the central point marked E (see Figure 4). In this event, the Taylor series (7), which defines the curve from A to B, can be constructed. As a result, the curve is known in the interval AB, and hence it is known in particular over any tiny arc within this interval. Let us then fix our attention on the arc in the neighborhood of the point S. Since this arc is now known with mathematical precision, we may proceed as before and construct the Taylor series that converges over a stretch of which S, or more properly s, is the centre. The Taylor series in question is the series (8). The curve is thus known from C to D, and hence altogether from A to D (since the part from A to C is already known). Let us now consider a tiny arc between B and D. Proceeding as before, we may construct a third Taylor series converging symmetrically about the new arc and extending to the right of the point D. Thanks to this third series, the construction of the curve to the right is extended. We may continue this method indefinitely, deducing one Taylor series from another, and by this means we may ultimately construct the entire curve in piecemeal fashion from A to a

point infinitely distant to the right. In a similar way we might construct the curve to the left.*

The method of progressive construction, just explained, is called *analytic continuation*. If we recapitulate what we have done, we see that, starting as we did from an assumed precise mathematical knowledge of an arc (however small) of the analytic curve, we have been able to construct the entire curve. No other curve than the one obtained could have been constructed by this method. This discussion brings to light an important characteristic of analytic curves. It shows that an analytic curve is completely determined (subject to the reservations mentioned in the note) when any arbitrarily small part of the curve is specified.†

The foregoing properties of analytic curves are reminiscent of those of physical systems evolving in a rigorously deterministic world. In Chapter VII we mentioned that the specification of an initial state of a physical system sufficed to determine the entire history of the system. The specification of an initial state of a physical system is therefore similar in its significance to the giving of a small arc of an analytic curve, and the determinism associated with physical processes is seen to be shared by analytic curves.

An immediate consequence of the properties of analytic curves is that if two analytic curves coincide over an arc, however small, they will necessarily coincide over their entire lengths. A straight line is the simplest illustration of an analytic curve, and it may be referred to as such in spite of the contradiction in the popular meanings of the words curved and straight. Since two analytic curves cannot coincide along an arc without coinciding everywhere, we conclude that if even a

* The foregoing explanations, which we have illustrated in connection with an analytic curve, assume a more rigorous form when we reason in terms of analytic functions. Thus the giving of a small arc of an analytic curve is equivalent to the specifying of the values of an analytic function $f(x)$ for all values of the variable x within a small interval. This information enables us (in theory) to determine the values of all the derivatives of the function for any value x_0 of the variable x within the given interval. We may then construct the Taylor series, in powers of $(x - x_0)$, which represents the function, and from this series determine the values of all the derivatives of $f(x)$ for any other value x_1 of x within the interval of convergence of the series. This latter information in turn permits us to construct a second Taylor series, in powers of $(x - x_1)$, which may converge beyond the original one, and so on.

† If a singular point is present on the real axis (*e.g.*, if the function becomes infinite for a certain value of x), the giving of an arc of the curve does not enable us to construct the entire curve, but only a part thereof. This is because the Taylor series will not converge beyond the singular point. However, this restriction can be removed by the introduction of the complex variable (see page 146), so that the general statement made in the text is correct.

tiny arc of an analytic curve is straight, the entire curve must necessarily be straight.

Consider then two straight segments forming a V. In view of the previous statement, the V-curve cannot possibly constitute a segment of an analytic curve. This fact entails an interesting consequence, illustrated in the problem of a vibrating string. As d'Alembert remarked:

"The ordinary way, indeed practically the only way, to set a string into vibration, is to hold it at a point, and then pull. And this yields two straight lines which form a broken line."

Whence d'Alembert concluded that the initial position of the string did not define an analytic curve, so that more generally a vibrating harp string could not always define an analytic curve at any current instant during its vibration. The consequences of this conclusion will be examined presently.

We may also understand why a curve traced at random with a pencil cannot in general be analytic. For if the curve were analytic, its entire length would be determined by the initial arc we might trace; and this is absurd, because, having started a curve with our pencil, we are free to continue it to our fancy. Nor need many other curves encountered in physics be analytic. For example, there is no reason why the graph traced by a registering thermometer should be analytic, for there is no particular reason why the sequence of temperatures over a given small interval of time should determine the temperature at all future instants.

We have dwelt at some length on analytic curves, because they were the only curves or functions which the mathematicians of the seventeenth and eighteenth centuries thought it possible to express by means of mathematical relations. All other curves, in the opinion of these early mathematicians, were incapable of mathematical representation and were ruled out as non-mathematical and as beyond the scope of mathematical investigation.

Non-analytic Curves—The inability of the earlier mathematicians to deal with non-analytic functions and curves precluded these investigators from attacking problems in which such functions appeared. Fortunately, the functions which they encountered in mechanics and in celestial mechanics were analytic, and because of this circumstance many mechanical problems were solved in the eighteenth century. But in general physics, the situation was different: non-analytic functions were often met with; and so mathematicians believed that the mathematical

instrument was inapplicable to the majority of physical processes. As an illustration, we recall that the shape of a vibrating string at any given instant is not usually represented by an analytic function. Accordingly, the mathematicians of the eighteenth century came to the conclusion that the problem of a vibrating string could not be investigated in its full generality by mathematical means.

It was therefore a momentous event in the history of mathematical physics when Daniel Bernoulli, towards the middle of the eighteenth century, claimed to have achieved the impossible. He constructed mathematical series which, so he maintained, were able to define the curve traced by a harpstring at any instant of its vibration. Needless to say, these series were not Taylor series, since the latter define none but analytic curves. The astonishment of Euler and d'Alembert may well be imagined; it was tempered, however, by the fact that Bernoulli's demonstration was lacking in rigor, so that the results he thought he had established remained in dispute. We may wonder why the dispute was not settled then and there; for is not a mathematical demonstration correct or incorrect? The trouble was that although Bernoulli had obtained series which formally defined the shape of the string, he was unable to establish the convergence of his series—and we have seen that unless a series converges, it defines nothing.* Not until the following century were Bernoulli's results vindicated by Fourier and by Dirichlet. His series (whose terms are the trigonometric functions *sine* and *cosine*), are today known as Fourier series.

The advance made by Fourier consisted in the formulation of a general method for constructing the Fourier series which, in his opinion, would represent any arbitrarily assigned function, analytic or not. Unfortunately Fourier, like Bernoulli, failed to give a rigorous proof of the convergence of his series, so that it was questionable whether they truly represented the functions they were supposed to represent. Owing to this lack of rigor, Fourier's memoir, though it received the prize offered by the French Academy, was nevertheless severely criticized by the judges, Lagrange, Legendre, and Laplace. Towards the middle of the nineteenth century, however, Dirichlet stated conditions under which Fourier series would certainly converge, and he thereby established the kinds of functions that could be represented by these series. Dirichlet's conditions have since been refined by Camille Jordan and others. Many of the masters, including Riemann, spent considerable time in investigating the various properties of Fourier series.

* In modern mathematics, however, divergent series are not entirely useless.

Although Fourier gave no rigorous proof of the convergence of his series, he proceeded on the assumption that, in all the cases contemplated, convergence would be realized, and that the series would thus define curves, or functions, whether analytic or non-analytic. These considerations were then extended by him to series of several variables. At the same time, Fourier established the empirical laws of heat conduction and showed that problems of heat conduction and of temperature equilibrium could be solved by means of his series. He also treated in more amplified form the problem of vibration, extending it from the case of strings to that of membranes and elastic solids. Thanks to Fourier, theoretical physics in its modern sense was born.

Before we proceed to a discussion of Fourier series, two important points may be noted. In the first place, the discovery of these series arose from an attempt to solve a physical problem; many other mathematical discoveries have been brought about in this indirect way. Indeed, the statement has often been made that the progress of mathematics would soon have come to a standstill had it not been for the suggestions derived from the physical world. Another feature illustrated in the discovery of Fourier series is the procedure whereby great advances in mathematics have often been obtained; namely, by stumbling and guessing rather than by pursuing a rigorous chain of deductive reasoning. Though rigor is essential in mathematics, yet in view of the limitations of the human mind it is often preferable not to demand too much of it in the early stages. A premature appeal to rigor might stifle discovery. Its effect might be to cloud the vision of the whole by concentrating our attention on some minor detail. History shows that detail may wait; the important thing is to explore the territory first and to establish its general outline. In Poincaré's words: "Of what use is it to admire the work of the mason if we cannot appreciate the design of the architect?" At a later date, subsequent mathematicians may advance along the grand avenues opened by the pioneers and concentrate on the details left in abeyance. Sometimes the working out of the details reveals errors invalidating the whole structure, and in this event the structure must be abandoned or considerably modified. But more often than not, the faith of the pioneers has been found justified. Besides, the subsequent attempts to vindicate or to invalidate their results have often led to other discoveries.

Let us pause to consider the kinds of functions that Fourier series can represent. Fourier himself believed that all continuous curves could be expressed by his series. In this, however, as was subsequently proved by Dirichlet and by Jordan, he was mistaken; for although Fourier

series can represent all analytic curves and also a large class of non-analytic continuous curves (*e. g.*, curves scribbled at random with a pencil), many continuous curves cannot be represented. But the interest of Fourier series is not restricted to continuous functions. These series can also express a large class of discontinuous curves, or functions. Discontinuous curves are curves which have breaks in them, *e.g.*, the curve traced in the figure. The number of discontinuities (or breaks) over a finite stretch may be so large that a graphical representation by means of a succession of tiny arcs is impossible in practice. In such cases the word "curve" is no longer appropriate, and we must adhere to the more technical appellation, "function." A Fourier series cannot, however, represent any arbitrary discontinuous function; restrictions must be imposed on the nature of the discontinuities.

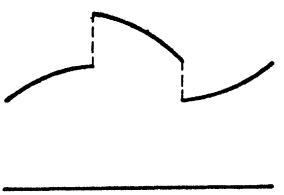


FIG. 5

Following Fourier's discoveries, mathematicians sought to construct functions by utilizing other developments in series. Weierstrass constructed a series more general than that of Fourier and showed it to define a continuous function, or curve, which exhibits the remarkable property of having no tangent at any of its points. The Weierstrass curve is described most simply when we say that it contains waves within waves *ad infinitum*; obviously it cannot be drawn with a pencil. We have here an illustration of a continuous function which cannot be represented by a Fourier series. In modern mathematical physics, *e.g.*, in the theory of heat conduction, theorists utilize constantly the more general series devised by the mathematicians who followed in Fourier's steps. All this development of theoretical physics would have been impossible had our mathematical knowledge been restricted to that of the mathematicians of the early eighteenth century.

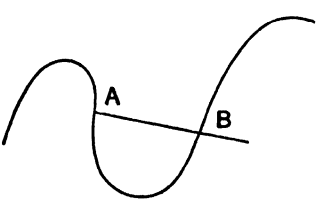


FIG. 6

The gradual incorporation of all continuous curves into mathematics led to a study of continuous curves the properties of which baffle our geometric intuition. Here rigor is essential. We have already mentioned Weierstrass's curve which has no tangents; and it may be of interest to understand how such a curve can exist. As is well known, the tangent to a continuous curve at a point A is defined by the limiting position of the straight segment drawn from A to a current point B on the curve, when this point B is made to approach A indefinitely. In order, then, that the curve should have a tangent at A, this segment must

tend to a limiting position. (The concept of limit previously referred to in connection with series thus crops up again.) But suppose that in the neighborhood of the point A, the curve contains waves within waves *ad infinitum*. In this case, when B moves towards A, the segment AB oscillates incessantly without tending to a limiting position; and as a result there will be no tangent to the curve at the point A. Now this is precisely the situation that holds at each point of the Weierstrass curve, and so we understand why this curve has no tangents. The foregoing discussion shows that a curve must exhibit a certain smoothness if it is to have a tangent, or slope, at every point.

Weierstrass's curve is but one among the strange curves. Thus we may conceive of continuous curves which are sufficiently smooth to have a tangent at every point and yet which have no definite radius of curvature at every point. We may also mention Peano's curve, which passes through every point of a given area and thereby fills the area completely. The collapse of our geometric intuition that is betrayed when we seek to visualize these strange curves, led mathematicians to question the reliability of this intuition in other cases. And so they raised such questions as: Is it always possible to attribute a definite length to an arc of a curve limited by two points A and B? May we always view as determinate the area contained within a closed continuous, plane curve? Does a closed continuous, plane curve (which has no multiple point) necessarily divide the plane into an outside and an inside part?

Geometric intuition would prompt us to give an affirmative answer to all these questions. But past experience shows that we must refrain from jumping at conclusions. Thus Camille Jordan has shown that the length of a curve, between two points A and B taken arbitrarily on the curve, may be indefinite. The Weierstrass curve is of this type. It can also be shown that a closed curve need enclose no definite area (in the ordinary sense of the word). On the other hand, even though a closed curve may have no definite length, it may enclose a definite area. As for the question of deciding whether a closed continuous curve which does not intersect itself will always divide the plane into two parts, it has been settled by Jordan affirmatively. In this case therefore our geometric intuition has not led us astray. But to obtain this assurance Jordan was compelled to resort to a lengthy demonstration, and the problem has many pitfalls. The fact that alternative demonstrations have been given by de la Vallée Poussin and by Brouwer shows the importance mathematicians attribute to this seemingly self-evident theorem.

Thus far we have been concerned with continuous curves. But we have said that Fourier series can also represent a certain class of discontinuous functions, or curves. Thanks to this possibility, discontinuous

functions have acquired a mathematical status which was formerly denied them. Following Fourier's discoveries, mathematicians sought to express a wider variety of discontinuous functions by means of series of one kind or another. In this connection we must mention that a discontinuous function $y = f(x)$ can have an infinite number of discontinuities corresponding to a finite range of values credited to the variable x. But here important differences arise according to whether the type of infinity contemplated is the one exemplified in the infinite sequence of integers or is the one associated with the aggregate of all real numbers (the mathematical continuum). A point of historic interest is that the study of the discontinuities of discontinuous functions led Cantor to his classification of the various kinds of infinity and thereby to his celebrated theory of infinite aggregates.

In another direction also, the developments originally suggested by Fourier series precipitated an important generalization in our understanding of mathematical functions. We saw that the earlier mathematicians regarded as mathematical only those functions which were susceptible of being expressed by means of mathematical operations. Prior to the advent of Fourier series, this restriction narrowed down the mathematical functions to the analytic ones. Then, as a result of Fourier's discoveries and of the developments to which they gave rise, many non-analytic functions were found to be mathematical in the sense just defined. But in the middle of the last century Dirichlet suggested that we generalize our conception of a mathematical function. According to Dirichlet, y will be a mathematical function of x whenever a precise law of correspondence between x and y can be stated clearly. Dirichlet no longer regards as indispensable that this law of correspondence be defined by mathematical operations; he also accepts any purely verbal definition provided it be precise.

An illustration of a function defined verbally is the function called Dirichlet's function. It is defined as follows: In the relationship, $y = f(x)$, the dependent variable y is assigned the value $+1$ or 0 according to whether the value of x is a rational or an irrational number. If, then, we assume that x is made to vary continuously between, say, 2 and 3, the value of y jumps back and forth between 1 and 0. It so happens that Dirichlet's weird function can receive a mathematical expression by means of a series, but this circumstance is accidental and in no wise restricts the general significance of Dirichlet's extended conception of a function. Discontinuous functions manufactured so to speak artificially were thus introduced into mathematics. The modern advance in our knowledge of discontinuous functions is due to René Baire. He

classes them into two groups according to whether they are *pointwise* discontinuous or *totally* discontinuous. Since the considerations he develops involve an understanding of the theory of aggregates, we shall dispense with further details.

As may well be imagined, the introduction of the stranger curves and functions has necessitated a generalization of the concept of integration. The method of integration was invented for the purpose of evaluating the area bounded by any given plane closed curve. In the seventeenth and eighteenth centuries, when only analytic curves were known, the areas to be calculated were always bounded by curves of this type, and the method of integration was that of Newton and Leibnitz. No particular difficulty was experienced in its application. But with the advent of the more general curves of the nineteenth century, Newtonian integration often became impossible. Indeed the very conception of an area enclosed by a curve becomes blurred when the bounding curve is a discontinuous one. However, despite the breakdown of geometric intuition in such cases, it is possible to generalize our understanding of integration in such a way that areas limited by discontinuous curves acquire at least a formal significance. The first generalization was accomplished by Riemann. A second one of greater scope has more recently been elaborated by Lebesgue.

The reader may well wonder whether these excursions of the mathematicians into the weird world of discontinuous functions are of any practical interest in theoretical physics. In this connection Baire, in the year 1905, prior to the development of the quantum theory, expressed himself as follows:

"Some physical theories in chemistry and mineralogy bear a certain resemblance to mathematical discontinuity. In any case, in spite of the old dictum, happily discarded, there is nothing which allows us to assert that 'Nature makes no jumps'. Under these conditions, is it not the duty of the mathematician to begin by studying *in abstracto* the relationships between the two notions of continuity and discontinuity, since these, while opposed, are intimately connected? This is perhaps the best way to prepare the advent of a mathematical physics in which the rôle of hypothesis will be reduced to a minimum."

The Complex Plane

Analytic Functions of a Complex Variable—In view of the fact that all the functions investigated during the eighteenth century were analytic, we might suppose that, at the close of the century, the properties

of analytic functions were well known. But this belief would be erroneous. As Volterra has remarked, the knowledge which the eighteenth century mathematicians had of analytic functions was comparable to the knowledge we might have of a book by merely viewing its cover.

Strictly speaking, the opening of the imaginary book in Volterra's analogy occurred in the eighteenth century, when the variable x in an analytic function $y = f(x)$ was allowed to assume complex as well as real values. In this way analytic functions of a complex variable were obtained. But no particular advantage was derived from this generalization till 1825, when Cauchy established a series of theorems on analytic functions of a complex variable and showed that the properties of these functions revolutionized our understanding of the analytic functions of a real variable. Cauchy's investigations opened up a new field in mathematical research, a field which is far from being exhausted even today. Some of the more important points established by Cauchy and his successors will now be examined.

We saw in connection with analytic functions of real variables, that if we allow x to assume all real values and then consider the corresponding values for y, we may represent the connection between x and y graphically by means of a curve. This method becomes impracticable when x is allowed to assume complex values, and so a different method of representation must be followed. We adopt the Gauss form of representation for complex numbers. Thus, if x is given a complex value $a + ib$, where a and b are both real numbers and where i stands for $\sqrt{-1}$, we agree to represent this complex value of x by a point P of abscissa a and ordinate b.* The figure is self-explanatory. If x assumes only real values, then $b = 0$, and a alone varies. The point P then describes the horizontal line in the figure. This line is called the *real axis*; and only when x has a real value is the representative point P situated on this axis. Now for each value ascribed to x, and hence for each point P of the complex plane, the functional relationship imposes a value (or sometimes several different values) on y. This value (or values) is usually also complex. We require therefore a second complex plane to represent the value of y. However, we need not represent the value of the function y geometrically, for we may confine ourselves to associating mentally a value (or values) of y with each point P of the first plane.

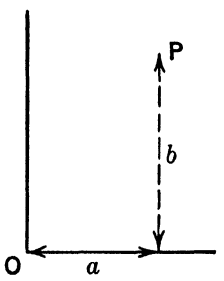


FIG. 7

* The plane on which the point P is represented is called the "complex plane."

Cauchy restricted his attention to *analytic* functions of a complex variable. Such functions are easily obtained if we select any analytic function of a real variable, *e.g.*,

$$(9) \qquad y = \sin x, \quad y = \frac{1}{x}, \quad y = \frac{1}{1+x^2},$$

and then agree that x may assume complex as well as real values. In some cases, when we make x (*i.e.*, P) move over the plane, we find that at certain points the corresponding value of y is infinite. Such points are called the *poles* of the function. For instance in the case of the second function (9), the point, $x = 0$, is a pole. This point is situated on the real axis, at the origin. The third function has two poles: they are given by the values of x which make the denominator $1 + x^2$ vanish, *i.e.*, the values $x = \pm\sqrt{-1}$, or $x = \pm i$. These two points are no longer situated on the real axis of the complex plane, but on the vertical at unit distances above and below the origin. As for the first function written, it has no poles.

In addition to poles, other important kinds of points must be considered. So called *branch points* may arise; these will be present only when to a given value of x correspond two or more values for y. Let us place ourselves in this case, and as an illustration let us consider the function

$$y = \sqrt[3]{x - (2 + 3i)}.$$

If in this function we give some value to x, we obtain, in general, three possible values for y. The point B, defined by $x = 2 + 3i$, at which the three values of y coalesce into one, *viz*, $y = 0$, is called a branch point of the function. To investigate the properties of branch points, let us suppose that we start from some definite value of x, say $x = 0$. We represent the value ascribed to x by a point P (here situated at the origin O). To this value of x correspond, as we have said, three values for y; we take one of them and disregard the others. Suppose now that starting from the point O we make x vary continuously over a range of complex values, and then make x return to O. In this event the point P, which represents the value of x at any instant, describes a closed path over the complex plane; and while P moves continuously along the path, the value of y also changes continuously. If the closed path does not surround the branch point B, we shall find that when P has described the closed path and is back at its starting point O, the value of y will also resume its original value. On the other hand, if the closed path surrounds the

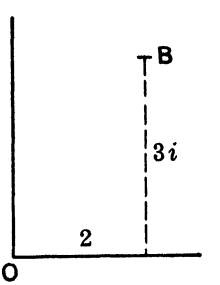


FIG. 8

branch point, the value of y will no longer be the initial value with which we started but will be one of the other two values that we disregarded. A second circulation round the branch point, in the same general sense of circulation and with a return to the same point O, makes y change values once again; and we obtain the third possible value. Finally, a third circulation restores the original value selected for y.

Branch points are thus present only when to any given value of x more than one value of y corresponds. Functions in which this situation arises are called *many-valued*, or *non-uniform*. When y has only one value for each value of x, the function is called *one-valued*, or *uniform*. There are no branch points in this latter case; and any closed path followed by x will always bring y back to the same initial value. Non-uniform functions introduce complications, for their non-uniformity interferes with our associating a definite value of y with each value of x, *i.e.*, with each point P on the complex plane. Even if we agree on the initial value for y when x is at some point P, and then make x follow an open path reaching some other point P', the value of y at P' will differ according to whether the path followed by x passes on one side or on the other of a branch point.

Cauchy sought to rid non-uniform functions of this ambiguity. So he made cuts in the plane, extending them from the branch points to infinity. The cuts act as barriers which cannot be crossed by continuous paths, and they thus prevent us from circling completely around a branch point. Thanks to this artifice, if we now make x describe any closed path which is not barred by the cut, we shall invariably return to our point of departure with the same value for y with which we started. But the drawback to this method is that having started with one of the possible determinations for y, we are unable to obtain the other determinations, so that only a part of the function can be represented at a time. Riemann improved on Cauchy's method. He replaced the Cauchy plane and its cuts by a number of superposed planes, or sheets, in which appropriate cuts were made. Strips passing through the cuts joined one sheet to the next and thereby acted in the capacity of staircases connecting the aggregate of superposed sheets. Suppose, then, we attempt to make the point x circle around a branch point. Owing to the communication strip, the cut radiating from the branch point no longer interferes with the circling; the point x passes through the cut, moves along the strip, and thus reaches the sheet situated immediately below the original one. During this process the value of y changes and we find ourselves with a new value for y at a point which is situated just underneath the original point of departure, and is on the next lower sheet. The net result is that to each

point *on each sheet* corresponds only one value for y, so that when we take into consideration the individualities of the various sheets, the requirements of uniformity are satisfied. Riemann's device has thus for effect to restore uniformity, or one-valuedness, to the function; at the same time it enables us to obtain all the various determinations of y.

This famous invention of Riemann's, called the method of Riemann surfaces (not to be confused with Riemann spaces in non-Euclidean geometry), may appear trivial. In point of fact it has revolutionized the theory of functions. We may readily understand that once the branch points of a function are known, it is possible, in theory, to construct the corresponding Riemann surface with its communicating strips, and thereby render the function uniform. We also sense the close relationship between the peculiarities of the communicating Riemann sheets and the general theory of the connectivity of surfaces, *i.e.*, that branch of mathematics called Analysis Situs (initiated by Listing and developed by Riemann). We have here an unexpected correlation between the theory of functions and Analysis Situs, two branches of mathematics which at first sight might seem to have nothing in common. Incidentally, the method of Riemann surfaces, by giving so to speak a geometrical model for the elaboration of an abstruse theory, illustrates one of the characteristics of Riemann. He seeks to visualize. Nothing of this sort is found in the works of Riemann's great contemporary, Weierstrass, whose major achievements are also found in this same theory of functions.

Poles and branch points are called *singular* points. There is also another kind of singular point, studied by Weierstrass, but it need not detain us here.

Analytic functions of a complex variable elucidate many of the features which mystified the mathematicians of the eighteenth century. We recall that analytic functions of a real variable may always be expressed by Taylor series, and that these series usually have only limited ranges of convergence. The underlying reason for these limited ranges becomes clear when complex variables are introduced. To understand this point we must mention that analytic functions of a complex variable, similarly to those of a real variable, may be expressed by Taylor series, *i.e.*, by series the successive terms of which contain the increasing powers of $(x - k)$, where k is any number, real or complex. Cauchy showed that, if we are given an analytic function of a complex variable and express it by means of a Taylor series in powers of $(x - k)$, where k is any number, real or complex, then this Taylor series will converge only so long as x remains within the circle whose centre is situated at the point k and whose circumference passes through the nearest singular point of the

function. This *circle of convergence,* as it is called, obviously contains no singular point in its interior. Of course, if our function has no singular points, all circles of convergences will be the same. all of them will then extend over the entire plane to infinity.

Let us apply these considerations to the analytic function

$$y = \frac{1}{x^2 + 1}, \tag{6}$$

which we have previously discussed from the standpoint of real variables.

Two singular points are apparent immediately: they are the two poles, $x = \pm i$, at which the function becomes infinite. On the complex plane these two poles, P and P', are situated at unit distance above and below the origin O, on the vertical passing through O. There are no other singular points. Let us consider the Taylor series in powers of $(x - k)$ that represents the function (6). According to Cauchy's rule this series will converge in a circle whose centre is the point $x = k$, and whose circumference passes through the nearer of the two points P or P'. In particular, let us consider the Taylor series in powers of x and the one in powers of $(x - \frac{1}{2})$. We have here two Taylor series in which k is taken to have the real values zero or $\frac{1}{2}$. The point, $x = k$, centre of the circle of convergence, is then situated at the origin O or at a point on the real horizontal axis separated from the origin by a distance $+\frac{1}{2}$. Circles of convergence having these points as centres may then be drawn, as in Figure 9.

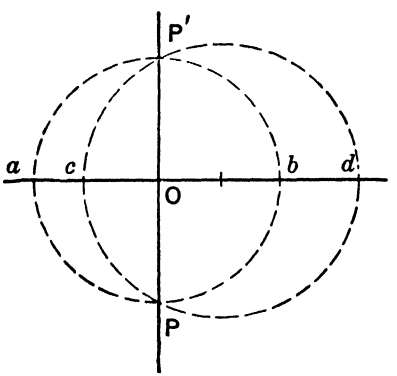


FIG. 9

If we confine our attention to real values for x, we see that the first of the two Taylor series converges only when x is between the points a and b, whereas the second Taylor series converges only when x is between c and d. The two Taylor series we are here considering are the two series (7) and (8), which we discussed on a previous page. In our earlier discussion we could give no reason for the strange limitations in the ranges of convergence of these series. The reason is now clear: the limited ranges of convergence are due to the presence of the singular points, P and P', situated on the complex plane. The mathematicians of the eighteenth century, by disregarding the complex values of x, were unable to understand the situation.

We also see why, in the case of the real function, $y = \sin x$, the Taylor series representing the function converges for all real values of x. It is because this function has no singular points (at finite distance), and so the one and only circle of convergence is the infinite circle, which covers the entire plane.

These elementary examples show how a knowledge of a function's singular points on the complex plane may yield important information on the behavior of the function, even when only real values of x are considered. In other words, the situation over the complex plane influences profoundly the occurrences along the real axis, and we cannot obtain any deep insight into the latter if we ignore the former. The limitations of the mathematicians of the eighteenth century now become obvious. All they knew was the behavior of their analytic functions along the real axis; they ignored the complex plane which extends above and below. As a result, they were in the situation of men wearing blinkers who walk along a road unable to see to the right or left. To such men the crossing of a bridge or the presence of a tunnel would be unpredictable events. Could they but look away from the road, they would perceive the river in the distance and the chain of hills, and they would expect the bridge and tunnel before reaching them. To the introduction of the complex variable, mathematicians owe the removal of the blinkers.

An important property of analytic functions of a complex variable is the following: If we are given the values of the function over an arc, as small as we choose, in the complex plane, we may deduce therefrom a knowledge of the function over its entire domain of existence.

Let us understand how the foregoing statement may be justified.

Consider a small arc of a curve in the complex plane, and suppose that we are informed of the value of an analytic function $f(x)$ at each point x of this arc. This information enables us to determine the values of all the derivatives of the function at any point A on the arc. The knowledge of the values of the function and of its derivatives at the point A permits us to construct a Taylor series which represents the analytic function, and which converges within a circle having the point A as centre. The circumference of this circle will pass through the nearest singular point of the function. If there is no singular point at finite distance, the circle will extend to infinity, covering the entire complex plane, and the function will thus be known in its entire domain of existence. But let us suppose that there is a singular point at finite distance, so that the circle of convergence has a limited radius. The function is then known only within this circle and is unknown at exterior points. Nevertheless, as we shall now see, the function can be con-

structed outside the circle. Consider any point B inside the circle. At this point B the values of the function and of all its derivatives are furnished by the Taylor series. The knowledge of these values enables us to construct a second Taylor series which also represents the function and which converges within a circle having the point B as centre. This second circle will usually overlap the first one and will thus determine the function beyond the limits of the first circle. Continuing in this way, we may determine the function progressively over more and more extended regions, until eventually we have determined it over its entire domain of existence. In this consists the method of *analytic continuation* as applied to analytic functions of a complex variable.

The domain of existence of the function may cover the entire complex plane or it may be restricted to a finite part. If the latter situation arises, the boundary line of the domain of existence acts as a barrier preventing further analytic continuation. This implies that the boundary line is formed of points all of which are singular—the boundary line itself is then singular. Beyond this line the function is meaningless.

These considerations on analytic continuation in the case of analytic functions of a complex variable clarify certain points we mentioned in connection with analytic functions of a real variable. Thus we mentioned that if a tiny arc of an analytic curve was given, we could deduce therefrom the entire curve by analytic continuation, provided, however, no singular point was present on the real axis (see note page 132). But we added that the introduction of the complex variable enabled us to remove this restriction. The justification for this statement is now clear. The singular point on the real axis prevents the real Taylor series from converging beyond this point, so that if we are restricted to real variables, the method of analytic continuation will enable us to construct the curve up to the singular point but not beyond. On the other hand, if we utilize the complex variable, we may proceed by analytic continuation around the singular point and then regain the real axis on the other side of the point. Thus the singular point ceases to constitute an obstacle. At the same time we see that the rigorous determinism which characterized analytic functions of a real variable (or analytic curves) finds its full expression in the analytic functions of a complex variable.

Applications of the Complex Plane—A particularly beautiful illustration of the utility of the complex plane, even when our interest is confined to functions of a real variable, is afforded by complex integration. For instance, consider the analytic function

$$y = \frac{1}{(1+x^2)^2}. \tag{10}$$

We assume that x takes only real values, so that the function may be represented by a curve. We thus obtain the curve traced in Figure 10; it extends from $-\infty$ to $+\infty$. Suppose now we wish to evaluate the area comprised between this curve and the Ox axis. The straightforward method is to integrate the function along the whole extent of the real axis (the horizontal line). Integrations of this sort are often tedious, but, as we shall now see, the introduction of the complex variable simplifies the problem considerably.

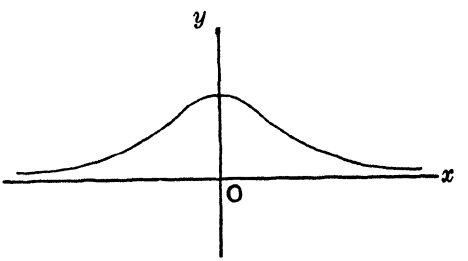


FIG. 10

Thus in the present example let us assume that x takes all complex values; our function must then be considered over the whole extent of the complex plane. We note that there are two poles present in the complex plane, *i.e.*, the two points $x = \pm i$. We now replace our original problem of integration along the real axis by a different problem. We integrate the complex function around a closed curve enclosing the pole $+i$. In particular, we select as closed curve the path defined by the entire real axis and by an infinitely extended semicircle having the point O as centre. The path of integration is illustrated in Figure 11, except that the dotted semicircle must be assumed infinitely large and hence the points A, B, and C infinitely distant.

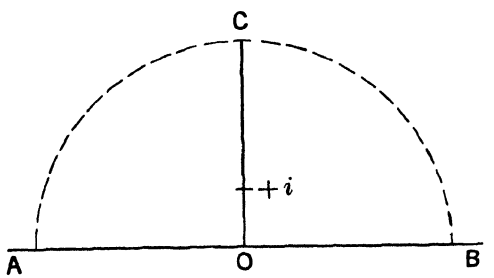


FIG. 11

Calculation shows that the integral along the circumference of the infinitely expanded semicircle has a zero value; and so the value of the integral around the entire closed curve has exactly the same value as the integral along the unlimited real axis AB. This latter integral, we recall, is the one we originally proposed to calculate. As a result, our problem will be solved if we can compute the integral around the entire closed curve. At first sight it might appear that we have merely replaced one problem by another equally difficult. But this is not so, for thanks to a theorem established by Cauchy, known as the "theorem of residues," the value of an integral along a closed curve surrounding a pole is obtained with little difficulty. In the present

case the value is $\frac{\pi}{2}$. Here, then, is also the value of the integral along the real axis, and our problem is solved with a minimum of effort. Our example again shows how useful is the introduction of the complex plane, even in problems dealing with real variables.

The artifice we have just resorted to is called the *method of complex integration*; it is of current use for the calculation of integrals. Many of the integrations met with in theoretical physics, in optics, in celestial mechanics, and elsewhere are most easily calculated by this method; in particular, complex integration is utilized in the computation of the orbits in Bohr's atom. The complex plane has also been the source of powerful methods in the theory of differential equations. Applications of these methods are numerous in mathematical physics. For instance, in the problem of the hydrogen atom in wave mechanics, Schrödinger utilizes a theorem due to Fuchs and a method invented by Laplace, both of which involve the complex plane.

Elliptic Functions—We now propose to examine various special kinds of analytic functions of a complex variable. During the eighteenth century, mathematicians sought to calculate the lengths of the more important analytic curves. The length of an arc of a curve is furnished by an integral limited by the end-points of the arc. In particular if the curve is an ellipse, we obtain an integral of a special kind, which, in view of its connection with an ellipse, was called an *elliptic integral.* The calculation of the arcs of other curves may also lead to elliptic integrals, the curve known as the lemniscate of Bernoulli being a case in point.

Let us take on the lemniscate curve a fixed point A and a variable point B. The length of the arc AB depends solely on the position of B (since A is fixed). The elliptic integral which gives the length of the arc AB is therefore a function of the position of the variable point B. But we may reverse the relationship and view the position of B as a function of the length of the arc AB. This latter function, called an *elliptic function,* is of course not the same as the function of the end point defined by the aforementioned elliptic integral; the elliptic function is, so to speak, the inverse of the elliptic integral, and for this reason we say that it is obtained by inversion from the elliptic integral.

Euler, and principally Legendre, studied these elliptic integrals and functions, concentrating on the integrals rather than on the functions. The entire subject was revolutionized by a single remark of Abel, who pointed out that Legendre's procedure was unnecessarily laborious. Instead of deriving the properties of the function from the integrals, Abel

maintained that the opposite procedure should be followed and that the study of the elliptic functions should be given priority. Abel and Jacobi soon put this idea into practice, and the results they obtained have entitled them to be regarded as the real founders of the theory of elliptic functions. Commenting on elliptic functions, Volterra writes:

"The story of elliptic functions is well known. It has often been told, for it is one of the most interesting chapters in mathematics. Surprise followed surprise as investigators passed from one stage of development to the next, which in turn brought new discoveries and new marvels."

Some of the more important properties of elliptic functions may easily be understood. Thus let

$$y = f(x)$$

represent an elliptic function. The variable x, as before, is allowed to move over the complex plane, and the corresponding value of y is then determined by the function. Only one value for y corresponds to each value of x, for there are no branch points and so the function is uniform, *i.e.*, one-valued. Elliptic functions betray important periodic properties, which we shall now examine. Elementary trigonometry makes us familiar with periodic functions. For instance, in the trigonometric function

$$y = \sin x,$$

where x is restricted to real values, we find that y (*i.e.*, the function) resumes exactly the same value when x is increased or decreased by 2π. Equivalently, we may say that y remains unchanged when $x \pm 2\pi$ is substituted for x. The magnitude 2π is called the *period* of the function; and since our function has only one period, it is called *simply periodic*. These results are expressed analytically by the equalities

$$y = \sin x = \sin(x \pm 2\pi);$$

and since the substitution of $x \pm 2\pi$ for x may be repeated any number of times in succession without the value of y being affected, we may write more generally

$$y = \sin x = \sin(x \pm m\, 2\pi),$$

where m is any integer, positive or negative.

Elliptic functions differ, however, from the trigonometric ones in that they are *doubly periodic*; by this is meant that they have two independent periods. These two periods, which we shall designate by ω and ω', may have any values, but not more than one of the two periods can be a real

number. In general, both periods are complex. The double periodicity of elliptic functions is expressed by the statement: If $y = f(x)$ is an elliptic function, then y will resume the same value when $x + \omega$ or when $x + \omega'$ is substituted for x in the function. Inasmuch as these substitutions may be repeated consecutively any number of times, the double periodicity leads to the relations

$$y = f(x) = f(x + m\omega + m'\omega'),$$

where m and m' are any two integers, positive, negative or vanishing. Obviously, elliptic functions are, from the standpoint of periodicity, a generalization of the simply-periodic trigonometric functions.

The double periodicity of elliptic functions is susceptible of a geometrical representation on the complex plane. To understand it, consider a point P on the complex plane. This point defines some complex number. If we add another complex number $a + ib$ to the original number, we obtain a point Q situated at a distance a from P along a horizontal and at a distance b along a vertical. Figure 12 is self-explanatory. The length and slant of the segment PQ are thus determined by the value of the complex number we have added.

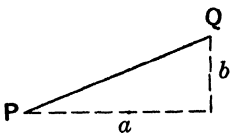


FIG. 12

We may now apply these considerations to the property of double periodicity of the elliptic functions. Let the point P (Figure 13) represent the value ascribed initially to x in the elliptic function $y = f(x)$. The two complex periods ω and ω' are represented by two different segments of definite slant and length. If, starting from P, we draw in all possible ways successions of segments equal and parallel to the period segments ω and ω', we obtain an aggregate of points in the complex plane. Each one of these points represents some numerical value for x, which differs from the value at P by integral numbers of time the magnitudes of the two periods. The aggregate of all such points defines the summits of a network of juxtaposed parallelograms, which extends over the entire complex plane; all these parallelograms are equal, their sides being equal and parallel to the period segments ω and ω'. Some of them are represented in Figure 13. The double periodicity of the elliptic function then shows that y resumes the same value when x is made to pass from the initial summit P to any one of the other summits. More generally, the value of the function y is repeated at all similarly-situated points in the various parallelograms. An immediate consequence is that if we know how the function behaves in any one of the parallelograms, we know thereby how it behaves over the entire complex plane.

In Figure 13 both periods ω and ω' are complex, but we have said that one of the two periods may be real. When one of the periods is real, the lower side of a parallelogram is parallel to the real axis (*i.e.*, the horizontal axis), as in Figure 14. In this case the elliptic function resumes the same value when we start, say, with a real value for x and then add to x an integral number of times the value of the real period. Only when we consider complex increments to the value of x is the second periodicity detected.

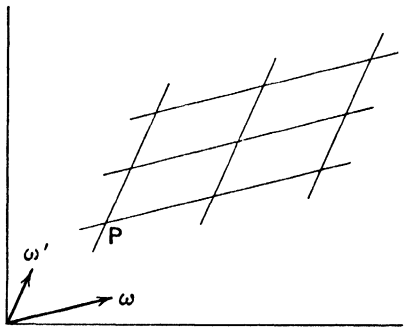


FIG. 13

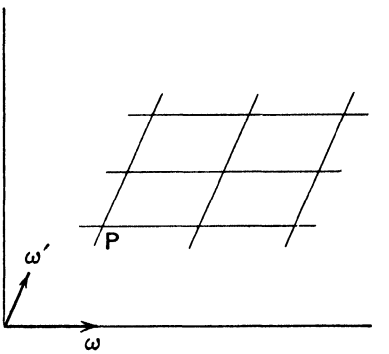


FIG. 14

The elliptic functions known to Legendre were of the type last mentioned, and since Legendre confined his attention to real values, only one period (the real one) appeared to be associated with his elliptic functions. Later Abel and Jacobi, by introducing the complex variable, were led to the discovery of the second periodicity and gained thereby an understanding of elliptic functions, undreamed of by Legendre. Abel, Jacobi, Weierstrass, Riemann, and Hermite are the great names associated with elliptic functions. Various series and functions were introduced by these mathematicians in order to coordinate results, *e.g.*, the *Theta functions* of Jacobi and the *Zeta and Sigma functions* of Weierstrass. The range covered by the theory is so vast that voluminous treatises have been written on this single department of the theory of analytic functions.

Elliptic functions furnish the solutions of several mechanical problems. The angle of oscillation of a swinging pendulum, for example, is expressed at any instant by an elliptic function of the time. The problem of the spinning top also involves elliptic functions, and so does the calculation of the planetary orbits in the theory of relativity. But quite aside from their physical applications, elliptic functions have yielded the solution of many problems in geometry and in the various fields of pure mathematics. Furthermore, the methods devised in the study of elliptic func-

tions served later as a guide for the elucidation of the properties of the more complicated functions that were soon to be discovered.

Modular Functions—Hermite, from a study of the elliptic functions, was led to the discovery of new uniform analytic functions with a more complicated kind of periodicity. These are the elliptic modular functions. If we designate an elliptic modular function by

$$y = f(x),$$

the function will resume the same value whenever x is replaced by

$$\frac{ax + b}{cx + d}, \tag{12}$$

where a, b, c, and d are any real integers satisfying the relation $ad - bc = +1$.* We must remember that x is assumed to receive all real and complex values. The periodic properties expressed in (12) yield a division of the complex plane into repeating domains limited by straight lines and circular arcs. These domains are therefore no longer parallelograms; neither are they equal in area. Modular functions, like elliptic ones, afford additional means of solving mathematical problems. It was thanks to the introduction of modular functions that Hermite was able to solve the algebraic equation of the fifth degree.

Automorphic Functions—More general than the modular functions are the *automorphic* functions of Poincaré. The same substitution that causes a repetition in the value of a modular function still holds, but now the numbers a, b, c, d are no longer restricted to real integers. The discovery of these new functions, which Poincaré classes under the headings of Fuchsian and Kleinian in the order of their generality, has opened a new continent in mathematics, much of which is still virgin territory. The foremost contributor in this field, after Poincaré himself, was Klein.

Let us observe that both trigonometric and elliptic functions are particular cases of automorphic functions. Elliptic functions repeat when x is replaced by $x + \omega$ or $x + \omega'$, where ω and ω' are the two periods. That this property also holds as a special case for automorphic functions is seen by setting in

$$\frac{a\,x + b}{c\,x + d}$$

the values $a = 1$, $b = \omega$ or ω', $c = 0$, $d = 1$.

* We may also say that an elliptic modular function resumes the same value when x is replaced by $-\frac{1}{x}$ or by $x + 1$; for when we perform these substitutions in succession in all possible ways, the general substitution (12) is obtained.

In his discovery of the new functions Poincaré was guided by the theory of elliptic functions. The method of inversion, noted in connection with elliptic functions, and also the construction of theta series were carried over by Poincaré and generalized to suit the new conditions. The applications of automorphic functions in pure mathematics are numerous and varied. Poincaré showed that an important class of differential equations can be integrated when automorphic functions are utilized. Interesting relations between the new functions and problems of higher arithmetic and of non-Euclidean geometry were established. Finally, Poincaré observed that the properties of some of these functions suggested the possibility of solving by their means the algebraic equations of degree higher than five.

There is one further application that we should like to mention. It concerns the representation of algebraic curves. By an algebraic curve is meant a curve defined by any algebraic relation between x and y. If we assume real values for x and y, the curve may be represented graphically by plotting x as abscissa and y as ordinate. Thus

$$x^2 + y^2 - R^2 = 0, \tag{13}$$

or equivalently

$$y = \pm\sqrt{R^2 - x^2}, \tag{14}$$

is an algebraic curve. It is of the second degree, since 2 is the highest power appearing in the equation (13). It represents a circle of radius R and with centre at the origin.

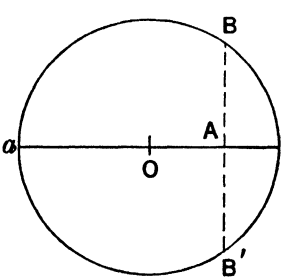


FIG. 15

Now we see that for a given value of x between $-R$ and $+R$, there are two values for y; these correspond to two symmetric points, such as B and B', which appear above and below the Ox axis. In certain cases we may wish to avoid this many-valuedness or non-uniformity. Since it is the radical in (14) which, owing to its double sign, generates the non-uniformity, we must rid ourselves of the radical. This may be done in the case of the circle by introducing an auxiliary variable called a *parameter*. Thus with t as parameter, the equations

$$x = R\,\frac{1 - t^2}{1 + t^2}, \quad y = \frac{2Rt}{1 + t^2} \tag{14}$$

define exactly the same circle when t is made to vary from $-\infty$ to $+\infty$. When t is given the value $-\infty$, the values of x and y are $-R$ and zero

respectively, so that we obtain the point a on the circle. Then, as t increases to $+\infty$, the point moves once around the circle in a counterclockwise direction. To each value of t, corresponds only one point on the circle; accordingly the ambiguity in the previous method of representation is removed. We now inquire: Can a similar procedure be followed for all algebraic curves? Can we, in every case, however complicated the curve, define a point of the curve without ambiguity when the value of some appropriate parameter is given?

Prior to the discovery of automorphic functions, such a representation was usually impossible. But by utilizing these functions, Poincaré showed that all algebraic curves could be given a parametric representation. As one commentator has said, this discovery gave Poincaré the keys to algebra. All these results are extremely important owing to their connection with the theory of Abelian integrals, which itself leads to one of the pinnacles of analysis.

Infinite Processes—From a psychological standpoint we first obtain an understanding of finite processes; only at a later stage are infinite processes considered. Thus the conception of the sum of a finite number of terms or of a finite number of areas preceded that of infinite sums, illustrated by series and integrals. The same is true of finite and infinite products. In this passage from the finite to the infinite, we cannot always extrapolate results uncritically from one domain to the other. The danger of a hasty extrapolation is apparent when we recall the important differences between the properties of the sum of a finite number of terms and those of the limit of an infinite series. We mentioned, in particular, that if a series which contains an infinite number of positive and negative terms converges, the limit of the series may change when the order of the terms is modified; the series may even be made to diverge. No such analogy holds for finite sums. A similar situation arises in the case of a convergent series of continuous functions. If there were only a finite number of continuous functions to be added, say a trillion, their sum would still define a continuous function. But this conclusion is not necessarily correct when we are dealing with an infinite series of continuous functions; the limit may now define a discontinuous function. It was by considering series of this latter type that Baire was led to his theory of discontinuous functions, referred to on a previous page. Many other properties that hold in the finite case may fail when infinity is considered. For these reasons, the passage from the finite to the infinite is always a delicate matter, and hasty conclusions are often found to be erroneous. Other examples of the passage to infinity are

illustrated by Poincaré's study of infinite determinants and by Hilbert's investigations on infinite quadratic forms. Infinite determinants occur in the theory of integral equations (these equations are of frequent use in modern mathematical physics). As for the theory of infinite quadratic forms, it has proved to be at the basis of the new quantum theory of the atom. These points will be understood when we discuss Heisenberg's Matrix Method.

Galois' Theory of Groups

Groups of Substitutions—Galois appears to have been the first mathematician to realize the importance of the group concept in mathematics. A few hours before meeting his death in a duel, he jotted down a summary of his discoveries on groups of substitutions and showed that his ideas revolutionized the entire theory of algebraic equations. The following brief explanations will indicate the essentials of Galois' theory.

Let us consider a set of different mathematical operations, which we may designate by A, B, C, ... If any two of these operations performed in succession are equivalent in their effect to one of the operations of the set, then the set of operations is said to have the group property. But for the operations to form a group, the inverse of every operation must be present in the set, and one of the operations of the set must be the identical operation (*i.e.*, an operation which changes nothing).

Suppose, then, we have a number n of different objects a, b, c, . . . placed along a straight line. They may be ordered in $n!$ different ways. Each different method of ordering is called a *permutation*. For instance, if there are three objects a, b, c, there are six permutations, namely,

$$abc,\ acb,\ bac,\ bca,\ cab,\ cba.$$

To pass from the first permutation to the second, we must replace b by c, c by b, and a by itself. To pass from the first permutation to the last but one, we must replace a by c, b by a, and c by b. Each one of these sets of replacements is called a *substitution*. There are as many different substitutions as there are different permutations. The identical substitution, which changes nothing, is the one in which a is replaced by a, b by b, and c by c.

We may readily verify that the aggregate of the $n!$ different substitutions connected with n magnitudes a, b, c, . . . forms a group. Thus, if a substitution is performed on a given permutation, and if to the permutation obtained a second substitution is applied, we obtain a permu-

tation which could have been derived directly from the original one by utilizing some appropriate substitution of the set. Furthermore, the set of substitutions comprises the identical substitution; and each substitution is accompanied by its inverse. The foregoing group, which contains $n!$ substitutions, is called the *symmetric group* of degree n.

The *degree* of the group is defined by the number of objects $a, b, c, \ldots$ on which the substitutions operate. The *order* of the group is defined by the number of distinct substitutions comprised in the group. Thus the symmetric group just considered is of degree n and of order $n!$

It may happen that some among the substitutions of a group can be withdrawn and that the remaining ones still form a group. The smaller group which is thus obtained is said to be a *sub-group* of the original group. For instance, the symmetric group of degree n always contains a sub-group, of degree n, represented by $\frac{n!}{2}$ of the original substitutions. This sub-group is called the *alternating group* of degree n. The alternating group in turn contains a sub-group. This sub-group may contain another, and so on.

Some sub-groups are characterized by a property described in the note. Such sub-groups are called *invariant sub-groups.** A given group may contain different invariant sub-groups. The invariant sub-group of maximum order is the one which comprises the greatest number of distinct substitutions.

Let us start with the symmetric group and consider its invariant sub-group of highest order (*i.e.*, the sub-group which involves the greatest possible number of distinct substitutions). The invariant sub-group in question turns out to be the alternating group. We then consider the invariant sub-group of highest order of the alternating group; and then the invariant sub-group of highest order of this sub-group; and so on. We thus obtain a succession of groups, each one of which is an invariant sub-group of maximum order of the preceding group. It is as though

* Let G_1 be a sub-group of G. Let S represent any substitution of G_1, and T any substitution of G. The inverse substitution of T is represented by T^{-1}. Suppose, then, that to any given permutation π we apply the substitution T^{-1}, obtaining thereby a new permutation π'. We then subject this new permutation to the substitution S and are thus led to a permutation π''. Finally, we perform the substitution T on π''. Our sequence of substitutions is represented by the product of substitutions $T^{-1} S T$. Since T^{-1}, S, and T are substitutions which are among those of the group G, it is certain that the product $T^{-1} S T$ is also a substitution belonging to this group.

Now, when S is any substitution of the sub-group G_1, and T any substitution of the group G, and when the substitution $T^{-1} S T$ always belongs to the sub-group G_1, then G_1 is said to be an invariant sub-group of G.

we had a collection of Chinese boxes, with each box (except the last one) containing another. The last member of this succesion of invariant sub-groups is represented by the identical substitution; for this substitution constitutes a group (though it is a very trivial illustration of one), and it also betrays the characteristics of an invariant sub-group.

To take a definite example, we shall consider the case of $n = 4$. The symmetric group contains $4! = 24$ distinct substitutions and hence is of order 24. The alternating group of order 12 is its invariant sub-group of maximum order. The alternating group in turn contains, as invariant sub-group of maximum order, a group of order 4. This sub-group contains an invariant sub-group of order 2; and the latter contains the invariant sub-group of order 1, represented by the identical substitution. The orders of our succession of groups are thus

$$24,\ 12,\ 4,\ 2,\ 1.$$

Galois considered the ratio of the order of each one of these groups to that of its successor. The ratios are given by the numbers.

$$2,\ 3,\ 2,\ 2.$$

He called this succession of numbers a *series of composition.* The present series arises when we start from the symmetric group for which $n = 4$. But the significance of a series of composition is the same whatever integral value may be ascribed to n.

For $n = 2, 3, 4, 5, \ldots n, \ldots$, the series of composition of the successive symmetric groups are the following:

(16)

$n = 2$	$2.$
$n = 3$	$2, 3.$
$n = 4$	$2, 3, 2, 2.$
$n > 4$	$2, \frac{n!}{2}.$

Sometimes a given group contains more than one invariant sub-group of maximum order. In this event, when we construct the series of composition, we are free to select any one of these sub-groups; and the series of composition may differ according to the choice we make. But it can be shown that the various series of composition will contain exactly the same numbers, and that the only difference that may arise consists in a change of the order in which these numbers appear in the series. Thus, all the series of composition of the same group contain

exactly the same numbers, so that these numbers are characteristic of the group considered.

If we revert to the series of composition written in the table (16), we see that the first three series contain only prime numbers. On the other hand, when n exceeds 4, the number $\frac{n!}{2}$, which now always appears in the series of composition, is composite. Hence we conclude that, when $n > 4$, the series of composition for the symmetric group do not contain prime numbers exclusively. The importance of this observation will be understood when we consider Galois' theory of equations.

Algebraic Equations—The general algebraic equation of the nth degree (where n is any positive integer) is obtained by equating to zero a polynomial of the nth degree in x. It is thus of the general form

$$(17) \qquad x^n + c_1 x^{n-1} + c_2 x^{n-2} + \ldots\ldots c_{n-1} x + c_n = 0,$$

where the coefficients c are constants of arbitrary value. According to a theorem known as d'Alembert's theorem, the equation (17) is satisfied when the variable x is credited with any one among n appropriate values

$$x_1, x_2, \ldots x_n.$$

These values are called the *roots* of the equation. D'Alembert did not give a rigorous demonstration of his theorem, but it has since been furnished by applying the theory of functions of a complex variable. The values of the roots depend on the values credited to the coefficients $c_1, c_2, \ldots c_n$ in the equation, and hence the roots are functions of these coefficients. To solve the equation means to obtain the explicit expressions of the roots in terms of the coefficients.

At the close of the sixteenth century the general equations of the first four degrees had been solved. The solutions of these equations (*i.e.*, the formulae defining the roots) were obtained in the form of explicit algebraic functions of the coefficients,* so that no irrationals other than radicals appeared in these solutions. Owing to this circumstance, the equations of interest were said to be soluble by radicals, or, equivalently, to be soluble algebraically. At the time it was confidently expected that

* By an explicit algebraic function of the coefficients c is meant a function which can be constructed by performing on these coefficients a finite number of algebraic operations (additions, subtractions, multiplications, divisions, and extractions of roots). Hence the only irrationals that can appear in an explicit algebraic function are radicals. When therefore the roots of an equation can be expressed by explicit algebraic functions of the coefficients, we are justified in saying that the equation can be solved by radicals.

further study would furnish similar algebraic solutions for the general equations of higher degree; and during the next two centuries numerous attempts were made to secure such solutions. But unexpected difficulties were encountered, and no progress was made. Some mathematicians, notably Lagrange, expressed doubts on the existence of algebraic solutions for the equations of higher degree. The arguments in support of these doubts are readily understood when we recall that algebraic expressions, or functions, represent a very special class of functions, so that there is no *a priori* reason why those functions of the coefficients which define the roots should necessarily be explicit algebraic expressions in all cases. Doubts, however, even when plausible, do not constitute mathematical proofs, and so the search for the elusive solutions continued.

Finally, in 1825, Abel put an end to further discussion by proving in a celebrated theorem that the general equation of the nth degree ($n > 4$) cannot be solved algebraically. Shortly after, Galois, by applying his theory of groups, extended Abel's results. Galois established a general theorem which provided a test for deciding under what conditions any algebraic equation (not necessarily an equation of the general type)* can be solved by radicals, *i.e.*, algebraically. The application of Galois' theorem to the general equation of degree n then furnished an immediate proof of Abel's theorem.

The theorems of Abel and Galois merely established the impossibility of obtaining *algebraic* solutions for the general equations of higher degree; they did not imply that non-algebraic solutions were unobtainable. Indeed, some fifty years after Galois' memoir, Hermite solved the general equation of the fifth degree, and Poincaré indicated the possibility of expressing the roots of the equations of still higher degree. These two masters utilized modular, and automorphic functions respectively, and their solutions were, of course, non-algebraic. In this chapter, however, we shall confine our attention to algebraic solutions and to Galois' investigations.

Before we examine Galois' procedure, let us inquire why mathematicians should have evinced any particular interest for the precise form in which the roots of algebraic equations could be expressed. It would seem that the sole aim of the mathematician should be to solve a given problem as best he can. We should expect him therefore to abstain from imposing on the solution some artificially restricted form, which might render the existence of the solution impossible. Why, then, should algebraic equations have been classed as soluble or as insoluble

* If particular numerical values are assigned to the coefficients of the general equation, the equation is said to be non-general, or particular.

according to whether their roots could or could not be expressed by algebraic functions of the coefficients? The answer to this question and to similar ones must be sought in the historical development of mathematics. Usually, what appears today as a restriction on the form of a solution was not regarded as such by the mathematicians who first attacked the problem of interest. In their eyes the so-called restriction appeared as a necessary requirement for the solution to be rigorous, and so the restriction was accepted without question. With the progressive development of mathematics, other forms of solution were constructed, just as rigorous as the earlier ones but of a more general kind. Only then did the supposedly necessary requirement of earlier days reveal itself as artificial, and only then did it assume the status of a restriction imposed voluntarily. Even after this discovery, however, problems in which the restriction was retained were still given serious attention by mathematicians; the only difference was that these problems were now viewed in their proper light.

Let us see how these considerations apply to the solution of algebraic equations. From the standpoint of the mathematicians of the sixteenth century, the only explicit expressions which could be accepted as mathematical were algebraic expressions and trigonometric functions. Accordingly, unless the roots of an equation could be defined by such expressions, the equation was held to be insoluble—solutions in the form given by Hermite would have been incomprehensible. Of course the naïve outlook of the earlier mathematicians was not shared by Abel and Galois. These mathematicians undoubtedly recognized that explicit algebraic expressions formed a very restricted class. Nevertheless the problem of solving algebraic equations by means of radicals was deemed by them of sufficient interest in itself to warrant investigation.

Many illustrations of the same kind may be given. Thus we mentioned that, not until the close of the eighteenth century, did mathematicians realize the arbitrariness of their procedure in restricting mathematical functions to be analytic. Yet, even after analytic functions were recognized as forming only a small sub-class among the continuous functions, mathematicians still sought to determine under what conditions the solutions of given differential equations would be analytic. Similarly, the Greek geometricians insisted on the exclusive use of the straight edge and compass in the demonstrations of their theorems—though in so doing they presumably did not realize that they were imposing a restriction which was by no means necessary. Nevertheless, even when the restrictive nature of the condition imposed by the Greeks

was understood, mathematicians still retained this restrictive condition in certain cases, and on this basis attempted to determine which of the geometrical problems were soluble or insoluble. For example, when we say that the squaring of the circle is impossible, what we really mean is that a square having the same area as a given circle cannot be constructed by means of the straight edge and compass. On the other hand, if we enlarge the field of permissible operations, the squaring of the circle may be performed, and the problem becomes soluble.

The Group of an Equation—Let us examine how Galois utilized his theory of groups in the study of algebraic equations. His first step was to connect a definite group with each equation. This group is called the *group of the equation*; its general significance will be understood from the following considerations.

Let $x_1, x_2, \ldots x_n$ represent the n roots of an equation of the nth degree,* and let $f(x_1, x_2, \ldots x_n)$ be some function of these roots. If we knew the precise expressions of the roots in terms of the coefficients c of the equation, and if we replaced the roots in $f(x_1, x_2 \ldots x_n)$ by these expressions, we should obtain the expression of the function f in terms of the coefficients.

Let us, then, consider all those rational functions † $f(x_1, x_2, \ldots x_n)$ of the roots $x_1, x_2, \ldots x_n$, which satisfy the following conditions: the coefficients they contain must be rational functions of the coefficients $c_1, c_2, \ldots c_n$ of the equation; furthermore, if in these functions the roots are replaced by their expressions in terms of the coefficients $c_1, c_2, \ldots c_n$, the functions must become rational functions of the coefficients $c_1, c_2 \ldots c_n$ and must have rational coefficients. These functions form a class which we shall refer to as the class A for short. Suppose now that we consider one of the functions of class A at random, and that we submit its n variables to all the substitutions of one of the

* The equation is assumed to be irreducible. This means that it is impossible to split the polynomial on the left hand side of the equation into a product of two or more polynomials the coefficients of which are rational functions (with rational coefficients) of the coefficients c of the equation. For instance, if the coefficients $c_1, c_2 \ldots c_n$ are integers and the equation is irreducible, it will be impossible to decompose the polynomial into a product of two or more polynomials the coefficients of which are rational numbers. Whenever an equation is irreducible, its roots are always unequal.

† Rational functions involve only the operations of addition, subtraction, multiplication, and division; and these operations may be repeated only a finite number of times.

groups of degree n. As an example let us suppose that we apply the substitutions of the symmetric group. These substitutions permute the positions of the variables $x_1, x_2, \ldots x_n$ in the function, and when account is taken of the numerical values of these variables (*i.e.*, the roots of the equation), some of the substitutions may be found to alter the numerical value of our function of class A. On the other hand, it may happen that the numerical value of the function remains unchanged under each one of the substitutions of this symmetric group. In such a case the function is said to belong to the symmetric group.

We first suppose that our function of class A belongs to the symmetric group. Obviously, then, the function will also remain unchanged in value under the substitutions of any one of the sub-groups of the symmetric group.

Next, let us assume that the numerical value of the function changes when the various substitutions of the symmetric group are applied. In this case we shall find that changes occur only for some of the substitutions, and that there will always be other substitutions of the symmetric group which do not change the value of the function. These particular substitutions that do not change the numerical value of the function always form a group, which is necessarily a sub-group of the symmetric group. Hence we conclude that there will always be a group of substitutions which leaves the numerical value of the function unchanged, and that this group is either the symmetric group or one of its sub-groups.

The considerations we have developed for one particular function of class A must now be extended to the whole aggregate of these functions. We are thus led to consider the group of highest possible order whose substitutions leave unchanged the numerical values of all the functions of class A corresponding to a given equation. This group may be the symmetric group or one of its sub-groups. It constitutes what Galois called the "group of the equation."

In short, all those rational functions of the roots $x_1, x_2, \ldots x_n$, which can be expressed rationally in terms of the coefficients of the equation, are functions which retain fixed numerical values under the substitutions of the group of the equation.

Conversely, all those rational functions of the roots, which retain fixed numerical values under the substitution of the group of the equation, are functions which can be expressed by rational functions of the coefficients $c_1, c_2, \ldots c_n$ of the equation.

We do not propose to enter into a discussion of the methods whereby the group of the equation can be determined in any particular case. Suffice it to say, this group may always be obtained.

Galois' Theorem—We wish to ascertain whether a given irreducible algebraic equation (general or non-general) can be solved by radicals. Galois' method consists in obtaining the group of the equation and then proceeding to construct the series of composition which starts from this group. Galois' theorem then states:

The equation will be soluble by radicals if, and only if, all the numbers appearing in the series of composition are prime.

Let us apply Galois' theorem to the general equation of the nth degree. We first determine the group of this kind of equation. We find that it is the symmetric group of degree n. Next we consider the series of composition. Now the table (16) shows that the series of composition for the symmetric group of degree n contains none but prime numbers only when $n < 5$, and that composite numbers occur when n exceeds 4. We conclude that the general equations of degrees $n = 1, 2, 3, 4$ can be solved by radicals and that those of higher degree cannot.

Binomial Equations—A special class of non-general equations, of any degree n that can be solved by radicals is illustrated by the so-called binomial equations. For instance

$$x^n - 1 = 0 \text{ (where } n \text{ is any positive integer)}, \tag{18}$$

is a binomial equation. Galois' theory proves that equations of this type are soluble by radicals. However, we need not utilize Galois' theory to establish this point; it was well known to earlier mathematicians, to Gauss in particular.

Binomial equations have an important application in geometry. A simple demonstration shows that if the lengths defined by the roots of the equation (18) can be constructed, it will be possible to divide a circumference into n equal arcs. Obviously, if this can be done, the regular polygon of n sides inscribed in the circle is obtained by drawing chords to the successive equal arcs of the circumference. In short, we see that if the roots of the equation (18) can be constructed, the regular polygon of n sides can also be constructed.

Now we have said that the roots of the equation (18) can be expressed by radicals. By this we mean that the expressions of the roots contain only the coefficients (here ± 1), integers, and radicals. But in order to construct these roots geometrically, *i.e.*, by ruler and compass, radicals other than square roots must be absent. Hence we conclude that the regular polygon of n sides can be constructed by the methods of geometry if, and only if, the roots of the equation (18) contain no radicals other than square roots in their expressions. The geometrical problem is thus

transformed into the algebraical problem of determining the values of n for which the roots of (18) will have expressions of the type just stated.

Gauss solved the problem. He first restricted his attention to prime numbers for n. He then proved that the roots of (18) will have expressions of the required form when, and only when, the prime number n can be represented by the formula

$$n = 2^k + 1, \tag{19}$$

where k is a positive integer. When in (19) we set $k = 1, 2, 3, 4, 5, 6, 7, 8, \ldots$, we obtain

$$n = 3, 5, 9, 17, 33, 65, 129, 257, \ldots.$$

Of the numbers written, only

$$n = 3, 5, 17, 257, \ldots. \tag{20}$$

are prime.

Thus, Gauss's result is that the regular polygons of n sides (n prime) which can be constructed geometrically are those for which n is a prime number expressible by the formula (19), and hence is one of the numbers of the list (20).* The value $n = 3$ refers to an equilateral triangle and is of course trivial. The value $n = 5$ signifies that the regular pentagon may be constructed—a result known prior to Gauss. But the value $n = 17$ and the higher values were unknown. We have mentioned Gauss's theorem because it illustrates a relationship between seemingly different departments of mathematics; namely, geometrical constructions with ruler and compass, and the abstract theory of equations.

Other Applications of the Theory of Groups—To Galois goes the credit of having introduced the group concept into mathematics. But Galois considered only those groups of substitutions which permute

* More generally a regular polygon of n sides, where n is any positive integer, may be constructed provided n be of form

$$n = 2^m p_1 p_2 p_3 p_4 \ldots, \tag{21}$$

where m is any positive integer or zero, and where p_1, p_2, p_3, p_4.... may have the value 1, or else are prime numbers, all of which are different and can be expressed by the formula (19). From formula (21) we obtain

$$n = 3, 4, 5, 6, 8, 10, 12, 15, 16, 17, 20, \ldots.$$

for the numbers of sides of the regular polygons that can be constructed.

objects, whereas the concept of group may be extended to operations generally. For example, the rotations which transform a regular polyhedron into itself constitute a group. We shall illustrate this statement in connection with a tetrahedron.

A tetrahedron can be rotated about any one of seven different axes in such a way that the edges of the rotated solid occupy the positions vacated by other edges. The tetrahedron thus appears to resume its initial position in space at the end of each rotation, and for this reason we say that the rotations transform the solid into itself. Twelve different rotations about the seven axes have this property of transforming the tetrahedron into itself. These twelve rotations (one of which is a zero rotation leaving the solid unmoved) form a group, called the tetrahedral group. The group property issues from the fact that a composition of rotations about any two of the axes is a rotation about one of the axes. If instead of operating with a tetrahedron we operate with an icosahedron (regular polyhedron having twenty sides), we obtain a group containing sixty different rotations. Klein established the connection between this icosahedral group and Hermite's solution of the algebraic equation of the fifth degree.

Galois' groups, as also the groups which we have just considered, involve finite numbers of substitutions, or operations. But a group may contain an infinite number of substitutions. For example, the substitutions

$$x' = \frac{ax + b}{cx + d}, \quad (ad - bc = 1)$$

where a, b, c, d are integers, are infinite in number, and they form a group. These substitutions, as we have seen, transform a modular function into itself, so that a modular function is a function which is reproduced under the substitutions of this group. Similarly, Poincaré's Fuchsian and Kleinian functions are transformed into themselves under the substitutions of groups which contain an infinite number of substitutions. Further examples of groups will be given in Chapter XX.

CHAPTER XV

MATHEMATICAL PRELIMINARIES (*Sequel*)

(Differential Equations)

Differential equations are probably the most important mathematical tools of theoretical physics. We shall examine only their most elementary properties. First, let us recall that when

$$y = f(x)$$

defines a continuous function y of a real variable x, the relationship between x and y can be represented graphically by means of a continuous curve (Cartesian method).

Derivatives—The rate of change of y when x varies is called the first derivative y' of the function. This first derivative usually varies in value when x is made to vary; hence it is itself a function of x. The value of the first derivative for any given value of x, corresponding to a point P on a curve $y = f(x)$, increases with the slope of the curve at P. It may be represented graphically by the trigonometric tangent of the angle θ contained between the Ox axis and the tangent AP to the curve at the point P (Figure 16), *i.e.*, by the ratio $\frac{BP}{AB}$. A continuous curve which, like Weierstrass's curve, has no definite tangent, or slope, at any point is thus equivalent to a continuous function having no first derivative.

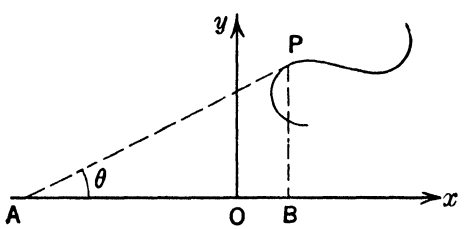


Fig. 16

The second derivative y'' of the function y is the first derivative of the first derivative; it represents the rate of change of the rate of change of y when x varies, or equivalently the rate of change of the slope of the curve as we proceed along the Ox axis. Obviously, if the curve is a straight line, so that the first derivative, or slope, remains unchanged, the second derivative will always vanish. Thus, the constant vanishing

of the second derivative indicates that our curve is some straight line. We may consider higher derivatives, *e.g.*, the third derivative y'''; it is the first derivative of the second derivative. More generally the nth derivative is the first derivative of the $(n-1)$th derivative. These elementary notions will suffice for the present.

Ordinary Differential Equations—Consider the function

$$y = \pm\sqrt{4 - x^2}. \tag{1}$$

If we plot x and y according to the Cartesian method, we obtain a circle of radius 2, with centre at the origin. Now at each point P of our circle there is a definite relationship between the slope of the circle and the position of this point P. Since the slope is determined by the first derivative y', there exists some definite relation holding, at all points of the circumference, between y' and the coordinates x and y of the point P. In the present example the relation is

$$yy' + x = 0. \tag{2}$$

An equation of this kind is called a *differential equation* because, in contradistinction to the equation (1), the relationship (2) involves a derivative.

Let us examine the comparative significance of the two equations (1) and (2). We know that (1) represents the particular circle mentioned above, and that (2) expresses a relationship holding for this same circle. We therefore inquire: Are the two equations equivalent, *i.e.*, do they merely express the same relations in different forms? If they do, we must assume that (2), like (1), defines the former circle and nothing else. We find, however, that the two equations are not equivalent, for it can be shown that (2) defines *all* circles having the centre as origin. In mathematical language, all circles

$$y = \pm\sqrt{R^2 - x^2}, \tag{3}$$

regardless of the value of their radius R, satisfy the differential equation (2). We conclude that our differential equation defines not one particular circle, but a whole class of circles. This important property of defining a class, and not a particular member of the class, is characteristic of all differential equations.

Suppose, then, we are given some differential equation; it expresses a more or less intricate relationship connecting x, y, and the derivatives of y. To say that the differential equation has solutions means that there

exist functions (or curves), $y = f(x)$, whose derivatives satisfy the relationship expressed by the differential equation. Our aim is to obtain these solution functions, as they are called. The problem of obtaining the solution functions of any given differential equation is called the problem of *integration*; it is one of the outstanding problems of analysis. To integrate a differential equation means to transform it in such a way that it is converted into a general mathematical expression which embodies all the solution functions. This general mathematical expression is called the *general solution,* or *general integral,* of the differential equation.*

Let us illustrate these considerations by reverting to the differential equation (2). All functions, or curves, that are solutions of this differential equation are expressed by the relation (3) in which any arbitrary fixed numerical value may be given to R. The function defined by (3) is thus the general solution of the differential equation. We obtain a particular solution by giving some specific value to R; for instance, the function (1) is a particular solution, which we obtain by setting $R = 2$. From a geometric standpoint the general solution represents all circles centred at the origin. A particular solution is any one of these circles.

Order of a Differential Equation—We may imagine as many differential equations as we wish: *e.g.*,

$$yy' + x = 0 \tag{4}$$

$$y'' = 0 \tag{5}$$

$$y'''(1 + y'^2) - 3\,y'y''^2 = 0 \tag{6}$$

are examples of differential equations. The order of a differential equation is by definition the same as the order of the highest derivative appearing in the equation. The equations written are thus of orders one, two, and three, respectively.

The equation (4) is the one just studied, and, as we have seen, it represents the class of all concentric circles that have the origin as centre; its general solution is given by (3), in which R is a constant of arbitrary value. The fact that one and only one arbitrary constant enters into the general solution shows that our aggregate of circles forms a simply infinite

* Cases may arise where the general solution does not comprise *all* the particular solutions, *e.g.*, when "singular" solutions are present.

family. The first order differential equation (4) thus defines a simply infinite family of curves, or functions.*

Let us now examine the equation (5). It expresses the vanishing of the second derivative; hence all functions that are solutions of (5) will be functions whose second derivatives always vanish. To interpret such functions geometrically, we recall that the vanishing of the second derivative implies *straightness.* Consequently, the differential equation (5) defines the class of all those continuous plane curves that have the attribute of straightness, *i.e.*, all the straight lines we may draw in the plane.†

Now all straight lines in the plane are expressed by the relation

$$y = ax + b, \tag{7}$$

in which any fixed numerical values may be given to a and b. We conclude that (7) represents the general solution of the differential equation (5). All particular solutions, *i.e.*, all particular straight lines, are obtained by giving in turn definite values to the constants a and b. Owing to the presence of two arbitrary constants a and b in the expression (7), the class of all straight lines in the plane forms a doubly infinite family; hence the second-order differential equation (5) defines a doubly infinite family of curves, or functions.

The differential equation (6) is of the third order; it represents the class of all circles in the plane regardless of their centres and radii. The general solution which defines all these circles will contain three arbitrary constants. The presence of three constants (the radius and the two coordinates of the centre) merely expresses the fact that the class of all circles in the plane is a triply infinite family.

Quite generally, it may be shown that if the differential equation is of order n, the general solution will contain n arbitrary constants. Consequently, all those curves, or functions, which are solutions of a differential equation of order n will form, in their aggregate, a n-ply infinite family. Conversely, the differential equation which defines a n-ply infinite family of curves will be an equation of the nth order. As an example, let us consider the differential equation whose general solution is the class of all

* A physical illustration of a simply infinite family of curves, defined by a first order differential equation, is afforded by the so-called magnetic spectrum. Iron filings are spread thinly on a sheet of paper placed over a magnet. The filings set themselves end to end, forming curves connecting the extremities of the magnet. These curves, which define the lines of force, constitute in their aggregate the general solution of a certain first-order differential equation.

† Vertical lines must be excepted, since for these $y' = \infty$.

conics in the plane. Since five independent data are required to determine a conic, the class of all conics is a quintuply infinite family of curves; hence the differential equation which defines this family will be of order five.

These considerations show that the higher the order of a differential equation, the greater the variety of functions that are its solutions. In particular, then, if we increase the order of a differential equation by differentiating it (*i.e.*, by taking its derivative), we shall obtain a new differential equation which will define a broader class of functions. An elementary example is afforded by the differential equations

$$y' = 0, \tag{7}$$

$$y'' = 0, \tag{8}$$

$$y''' = 0. \tag{9}$$

The second equation is obtained by differentiating the first; the third, by differentiating the second. It is therefore correct to say that the first entails the second and the third; and that the second entails the third. But the converse is not true, as is readily seen from the general solutions of these differential equations.

Thus the general solution of (7) is

$$y = c, \tag{10}$$

where c is an arbitrary constant. It represents all the horizontal straight lines of the plane.

The general solution of (8) is

$$y = bx + c, \tag{11}$$

in which b and c are arbitrary constants. It defines all straight lines in the plane and therefore defines a more extensive class of curves than does (10), which defines only horizontals.

The general solution of (9) is

$$y = ax^2 + bx + c, \tag{12}$$

where a, b, and c are arbitrary constants. The class of curves here defined comprises all straight lines and also a restricted class of parabolas. We conclude that each one of the general solutions written above is more inclusive than the preceding one. For example, by assigning the value zero to a in (12), we eliminate the parabolas and are left with (11), *i.e.*, with the class of all straight lines. Similarly, by assigning the value zero to a and to b, we obtain (10), *i.e.*, the class of all horizontals. In short,

all solutions of (10) are solutions of (12), but not all solutions of (12) are solutions of (10). We have thus verified that the differential equations (7), (8), and (9) define increasingly extensive classes of curves.

The theory of differential equations forms one of the most important chapters of analysis; and the problem of integration has ever been an outstanding one. Numerous methods of integration have been devised. Here also Cauchy's complex plane has proved useful. A fundamental problem has been to determine whether a given differential equation has solutions. Affirmative proof is called an *existence theorem.* Differential equations are intimately connected with the theory of functions (as is to be expected, since their solutions are functions). We mentioned that Poincaré, by constructing the automorphic functions, integrated an important category of differential equations. The significance of his discovery should now be clear: it is that the solutions of these differential equations can be expressed by automorphic functions.

Partial Differential Equations—The derivatives we have considered thus far are those of functions of a *single* variable x. We must now examine the derivatives of functions of several independent variables. Such functions are often encountered in mathematical physics. For instance, in a room with a curved ceiling, the height of the ceiling above a point P on the floor of the room varies with the position of P; the height is thus a function of P. If we define the point P by its Cartesian coordinates, x and y, and represent by z the elevation of the ceiling, then z is a function of both x and y. These two variables are the independent variables. We express this functional relationship by writing

$$z = f(x, y). \tag{13}$$

The concept of derivation also arises for a function of several variables. But it is more complicated because the derivation may be performed in different ways. Thus we may conceive of the rate of change of z when x is varied while y is kept fixed; or again, we may keep x fixed and allow y to vary. Such derivatives, represented respectively by

$$\frac{\partial z}{\partial x} \text{ and } \frac{\partial z}{\partial y} \tag{14}$$

are called the first partial derivatives of the function z with respect to x and with respect to y. These partial derivatives, like z itself, are functions of x and of y. Partial derivatives of higher order are obtained by taking the partial derivatives, with respect to x or to y, of the first partial

derivatives. In particular, if the latter derivatives are differentiated only once, we obtain the second order partial derivatives, viz.

$$\frac{\partial^2 z}{\partial x^2}, \frac{\partial^2 z}{\partial x \partial y}, \frac{\partial^2 z}{\partial y \partial x}, \frac{\partial^2 z}{\partial y^2}. \tag{15}$$

The two derivatives $\frac{\partial^2 z}{\partial x \partial y}$ and $\frac{\partial^2 z}{\partial y \partial x}$ are not necessarily the same, a fact which indicates that the order of derivation is not necessarily commutative. However, in theoretical physics the functions encountered are usually of such a kind that these two derivatives are identical. We have confined our discussion to derivatives of the first two orders, but derivatives of the third and higher orders may also be considered.

A partial differential equation is a relation connecting some of the partial derivatives of a function *inter se*; in addition, the relationship may comprise the function itself and the independent variables. The *order* of the partial differential equation is defined by the order of the highest partial derivative appearing in the equation. Explanation is made easier by considering a definite physical problem. Indeed it was for the purpose of understanding specific physical problems that partial differential equations were first studied; and we are following the method most acceptable to the human mind when we pursue the historical order. Let us suppose that we maintain the various points of the surface of some solid conductor at arbitrarily given temperatures. After a sufficient lapse of time the temperature distribution inside the solid becomes permanent. This permanent distribution will depend on the temperatures at which we keep the various portions of the surface (surface conditions), and also on the shape and size of the body—for instance, for a given body the temperature distribution will not be the same if we maintain the whole surface at the temperature of boiling water, or if we keep half the surface at this temperature and the other half at the temperature of melting ice. But though the precise temperature distribution must depend on the shape of the solid and on the surface conditions, there will always be a certain sameness in all temperature distributions: the temperature will always be finite and will always vary continuously from place to place; and so also will the gradient of the temperature taken in any direction; in addition, as Fourier has shown, the permanent distribution will always satisfy a second-order partial differential equation, called Laplace's equation. This equation, controlling as it does all permanent temperature distributions, represents the law of such distributions.

The mathematical significance of Laplace's equation is easily understood. Thus let us define the various points within the solid by their

Cartesian coordinates x, y, z; and let T denote the temperature at any point x, y, z. The value of T will usually vary from point to point, so that T will be a function of x, y, z. We may represent this function by $T(x, y, z)$. According to whether one temperature distribution or another is realized, the function $T(x, y, z)$ will be of one kind or another. The result obtained by Fourier from his speculations on heat conduction was that, regardless of what the precise temperature distribution might be, the function $T(x, y, z)$ would always have second-order derivatives satisfying the relation.

$$\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2} = 0. \tag{16}$$

This relation constitutes Laplace's equation.*

Inasmuch as the equation (16) is to control the temperature distributions that will arise in all cases, regardless of the solid's shape and of the surface conditions, we conclude that the restriction embodied in Laplace's equation is insufficient in itself to determine one particular function $T(x, y, z)$, or temperature distribution: the equation merely determines a class of possible functions. Such functions (the solution functions of Laplace's equation) are called *harmonic functions*. In other words, whatever precise temperature distribution, or precise function $T(x, y, z)$, occurs in a particular problem, this function will always be one of the solution functions of Laplace's equation (16). The characteristic class-feature of the ordinary differential equations, examined previously, must therefore be extended to partial differential equations. Thus, partial differential equations define whole classes of functions or possibilities, and not any one particular function.

Guided by what we have said of the ordinary differential equations, we might suppose that to integrate a partial differential equation means to obtain a general formula which comprises all the solution functions. This procedure, however, is not usually feasible with partial differential equations, and piecemeal methods must be resorted to. We therefore proceed as follows:

We specify that the solution functions of the differential equation must satisfy additional restrictions. By an appropriate choice of these restrictions we may narrow down the original class of all the solutions to a sub-class or even to one particular solution. If, then, the correspond-

* The same equation is satisfied by the distribution of the Newtonian potential throughout empty space (see note, page 216), and it was in this connection that the equation was studied by Laplace.

ing solution or solutions can be constructed, the partial differential equation is said to be integrated in a special case.

The problem of determining the restrictions that eliminate all but one particular solution is a problem which the pure mathematician is called upon to solve. No general rule can be given because partial differential equations which appear to be nearly the same often have widely different properties. In the particular case of Laplace's equation, the restrictions which serve to eliminate all but one of the solutions may be guessed at from physical considerations. Thus we have said that Laplace's equation expresses the general requirements which all permanent distributions of temperature must satisfy in a solid body; and we know from experience that a definite continuous temperature distribution arises throughout a solid of arbitrary shape when the surface points of the solid are maintained at fixed temperatures. These empirical results, transcribed into mathematical language, imply that one particular solution function of Laplace's equation will always exist within a closed volume provided the following restrictions be imposed:

1—the solution function is to be one-valued, continuous, and with first and second derivatives continuous.

2—the solution function is to assume given values over the surface of the closed volume.

But, of course, the mathematician cannot be satisfied with such physical promptings; what he demands is a rigorous mathematical demonstration. In the case of Laplace's equation, the mathematical demonstration has been obtained; hence we are certain that the restrictions listed above will single out one particular solution. The mere assurance that a solution function exists does not, however, furnish this solution; and we are still confronted with the task of constructing it. If the solution function satisfying given restrictions can be constructed, Laplace's equation is said to be integrated in the particular case considered.

The method actually followed in obtaining the solution function of Laplace's equation varies considerably with the shape of the volume. According to whether the volume is a parallelepiped, a sphere, a cylinder, or an ellipsoid, it is simpler to seek the solution in the form of a Fourier series, or in the form of a series of Laplace functions, or Bessel functions, or Lamé functions. If the shape is irregular, Fredholm's method of integral equations is usually applied. From this analysis we realize how the shape of the volume, which at first sight appears to be a minor consideration, may lead to different methods of approach and to totally different mathematical problems.

Suppose, then, we are dealing with a solid of prescribed shape and size whose surface points are maintained at given temperatures. If we can construct the solution function of Laplace's equation that corresponds to the conditions of our problem, we shall automatically obtain the temperature distribution which will persist within the solid. Thus, our ability to predict the temperature distribution depends on our ability to integrate Laplace's equation in the particular case considered. Besides controlling the temperature distribution in solids, Laplace's equation controls many other physical phenomena, *e.g.*, phenomena involving gravitation, electrostatics, elasticity, and the permanent flow of liquids. If, then, we can integrate Laplace's equation in the various cases that present themselves in the different problems, we shall be in a position to predict not only the temperature distribution in a given solid, but also the behavior of systems in which gravitation, electrostatics, etc., are involved.

The Importance of Integration—Let us suppose that a physical system is controlled by Laplace's equation. We have just seen that our knowledge of this fact does not in itself permit us to predict the behavior of the system, and that, unless Laplace's equation can be integrated in the particular case which corresponds to our physical problem, the exact behavior of the system cannot be foreseen. These general conclusions, which we have derived from a discussion of Laplace's equation, apply to all differential equations and to the physical processes they may control. Since many of the laws which regulate physical processes are expressed by differential equations (*e.g.*, Maxwell's laws of electromagnetics and Einstein's gravitational equations), we may well appreciate the importance of integration in theoretical physics. Indeed, the theoretical physicist, in his attempts to draw definite conclusions from his premises, is utterly dependent on his ability to integrate differential equations. The following illustration which concerns the theory of relativity will serve to clarify the situation further.

According to the general theory of relativity, free bodies, such as the planets, follow the geodesics of 4-dimensional space-time. In the neighborhood of matter, space-time becomes curved, so that the geodesics become curved lines and hence free bodies pursue curved paths. In order to predict the paths and motions of free bodies in any given situation—*e.g.*, to predict the paths and motions of the planets round the sun—we must therefore determine the curvature of space-time from place to place, and then obtain the geodesics. Now we cannot expect the theory of relativity to define this curvature in any precise way, for the curvature

will necessarily depend on the distribution of matter in the problem considered. What the theory of relativity furnishes is a system of relations which prescribes the connections which will always hold between the curvatures at neighboring space-time points. These relations are expressed by partial differential equations called Einstein's gravitational field equations. Einstein's equations, by the mere fact that they are of the differential type, do not define any specific space-time curvature from place to place; they merely furnish a whole class of possible curvatures. To obtain the exact curvature of space-time in any particular situation, we must integrate Einstein's gravitational equations and then restrict the solution so as to conform to the requirements of the situation contemplated. Let us suppose the solution has been found, so that we now know the curvature of space-time from place to place. We still have to determine the geodesics if we wish to predict the paths and motions of free bodies and of waves of light. The discovery of the geodesics requires further integrations, for these geodesics are themselves controlled by differential equations. In short, we shall be informed of the motions of free bodies in any particular situation only if we can perform several integrations, the first of which is the integration of the gravitational equations.

Einstein's gravitational equations, however, are so complicated that to the present day no one has been able to integrate them rigorously in the general case. Fortunately for the theory of relativity, the situation of major interest in astronomy happens to be one of the simplest imaginable. We refer to the situation which arises in the solar system, where a single spherical mass, the sun, may be considered alone, the masses of the planets being so small in comparison that we may neglect their effects on the space-time curvature. This circumstance simplifies considerably the problem of integration, for, in view of the symmetry of the conditions around the sun, we know that the solution we are seeking will be radially symmetric; all other solutions may therefore be ignored. By analogy with what we said of Laplace's equation, we must also stipulate the values of the solution over the bounding surface, here represented by spatial infinity. In the general theory of relativity the boundary values are furnished directly, thanks to Einstein's assumption that the space-time structure is flat at infinity.* The foregoing restrictions in the symmetry and in the boundary values being imposed on the solution functions, one definite solution is determined. Einstein himself was unsuc-

* The specification of conditions at spatial infinity was displeasing to Einstein; and one of his reasons for imagining the cylindrical universe was to render possible the determination of a solution without having to consider the conditions at infinity.

cessful in constructing the rigorous symmetric solution. But soon after, the gravitational equations were integrated by Schwarzschild in the special case considered, and the solution was constructed. The curvature of space-time round the sun being thus established, the geodesics were obtained by an easy integration and so the paths and motions of the planets and of waves of light were determined rigorously. It thus became possible to submit the theory to crucial observational tests.

The case of integrability just considered is one of the very few in which the integration of Einstein's equations has been performed; and, as we have said, this simplest of cases happens by sheer luck to be the one of major interest in astronomy. Other situations are also of interest, *e.g.*, when there are two gravitating masses of the same comparative importance, a situation realized in a double-star system. But here integration has proved impossible, and we cannot tell exactly how the stars of a double-star system should move according to the theory of relativity. The conclusions we have arrived at for Einstein's differential equations are general. Until the equations have been integrated, they are practically useless. Many of the problems which have beset the pure mathematician have arisen from a desire to assist the theoretical physicist in his integrations.

Differential Equations and Physical Laws—With the exception of the statistical laws, all the physical laws formulated prior to the quantum theory can be expressed by means of differential equations of one type or another. It is with these laws that we shall be concerned for the present. First let us understand what is meant by a physical law.

The name law is given in physics to a restriction which controls all the phenomena belonging to a class. For instance, the laws of mechanics control all mechanical phenomena, and the laws of electromagnetism control all electromagnetic ones. As a further illustration of a physical law, we may mention Boyle's law of gases. It states that, for a given mass of gas at constant temperature, the product of the volume and pressure will always have the same value. Here we are expressing Boyle's law in the so-called finite form. But there is another form in which Boyle's law and all similar physical laws can be expressed. This alternative form, which is called the differential form, was rendered possible by Newton's mathematical discoveries. It offers many advantages because it is far more general than the finite form; indeed, many of the physical laws can be expressed only in the differential form.

When a law expressed in the differential form deals with magnitudes which can be represented in the space-time frame, the law defines relation-

ships that must hold between one or more characteristic physical magnitudes at a current space-time point and these magnitudes at any space-time point in the immediate vicinity of the first one. In certain cases the phenomenon controlled by the law is static, so that the notion of time does not enter into account. Of this type is the permanent temperature-distribution in a solid. The law controlling this distribution (Laplace's equation) merely imposes a restriction on the manner in which the temperature may differ permanently from point to point. But in the majority of cases our concern is with processes of change. The laws regulating these processes then involve time.

Let us, then, consider some physical system which is subjected to change as time passes, and let us suppose that the law controlling the evolution of the system is known. The law expresses a permanent relation between the state of the system at any current instant and its state at the immediately following instant. The restriction which this relation imposes is not sufficient, however, to determine a definite course for the physical system; it leaves open the possibility of an infinite number of different courses. But if we supplement the law by specifying the initial state of the system, one definite course becomes inevitable (at least this is so if we accept the doctrine of rigorous causality). As for the meaning of a "state", it varies from one system to another. Quite generally, the state of a given physical system may be defined as that complex of data which, under the assumption of rigorous causality, suffices to determine the history of the system uniquely.

Further insight into the nature of physical laws is obtained when we examine the characteristics of the physical processes they control. Classical physicists assumed that all physical processes were continuous. Even in the collision of two billiard balls, where the courses of the balls appeared to change abruptly, the changes were claimed to be continuous though extremely rapid. The discovery of quantum phenomena has since led physicists to question the doctrine of the continuity of change. But for the present we shall not consider the quantum revolution and shall restrict our attention to the classical doctrine. In a continuous process, the sequence of states assumed by a physical system forms a continuous succession; by this we mean that the change in the state of the system, over an infinitesimal period of time, is itself infinitesimal. The laws controlling such processes will, therefore, exhibit the features of continuity.

From this analysis we see that a physical law (involving the passage of time) expresses a permanent relationship between the present state of a continuously evolving system and the state of the system at the follow-

ing instant. More concisely, we may say that the law establishes a permanent relationship between the antecedent and the consequent; it defines the connection between the successive links in the causal chain.

An example of a physical law is afforded by Newton's second law of motion. According to this law the product of the mass and acceleration of a body is at all instants equal to the force acting on the body. Since the acceleration is the time-rate of change of the velocity, Newton's law is seen to define the necessary relation which must connect the velocity of the body at a current space-time point with the velocity at a point in the immediate vicinity of the first.

Next, we consider the mathematical expression of physical laws. The mathematical relations that will express a law must necessarily satisfy the following requirements: they must be susceptible of defining connections between magnitudes at neighboring space-time points; they must be sufficiently general to leave open a whole class of possibilities; and they must connote the continuity of change. All these requirements are satisfied by differential equations. Thus in a differential equation the derivatives establish a connection between the values of a magnitude at neighboring points. Furthermore, the presence of these derivatives implies that the feature of continuity is satisfied. Finally the feature of generality is assured by the fact that the general solution of a differential equation embodies a whole class of possibilities. We conclude that differential equations furnish the mathematical expression of the physical laws. Let us verify this conclusion in the particular case of Newton's second law of motion.

To take a simple illustration we shall assume that Newton's law is applied to a point-mass moving along a straight line under the action of a force. We denote by x the distance of the point-mass, at any instant, from a point O taken as origin on the straight line. At any instant t, the point-mass occupies some position on the line, and hence this position x is a function of the time t. Since the first derivative x' of x with respect to t is the time-rate of change of x when t varies, x' represents the velocity of the point-mass. Consequently, this derivative, or the velocity it measures, has one value or another according to the instant t considered. Similarly the second derivative x'', defining as it does the rate of change of the velocity at any instant, represents the acceleration of the point-mass at this instant. We shall assume that the force varies with the position x of the body, and we may therefore represent it by some function $F(x)$. Newton's law of motion is then seen to be represented by

$$mx'' = F(x), \tag{17}$$

i.e., by a second-order differential equation of the simpler kind studied at the beginning of the chapter (for x is a function of only one independent variable, namely, the time t).

In this discussion we have supposed that the point-mass is moving along a straight line; but more generally it may be moving in space, and in this event we should have three differential equations similar to (17): one equation for each of the three coordinates x, y, z of the point-mass. If, instead of a single point-mass, we were dealing with a solid body of finite size or with several point-masses, Newton's law would be expressed by a larger number of differential equations, but these equations would still be of the same type as the equation (17). We therefore conclude that the equations of dynamics are second-order differential equations.

As we shall now see, the equation (17) does not prescribe any definite motion for the point-mass; instead, it defines a whole class of possible motions, the exact motion being determined only when the initial state of the point-mass is specified. To verify these assertions, we shall suppose for the sake of simplicity that the force acting on the point-mass is a constant force, so that it may be represented by a constant K. The differential equation (17) thus becomes

$$mx'' = K. \tag{18}$$

We wish now to determine the motion of the body or, what comes to the same, its position x at any instant t. We acquire this information by integrating the differential equation, thereby obtaining its general solution. The general solution is

$$x = \frac{K}{2m} t^2 + at + b, \tag{19}$$

in which a and b are arbitrary constants. By giving different sets of values to the constants a and b, and hence by considering different particular solutions of the differential equation (18), we obtain different motions for our point-mass. Each one of the possible motions is thus expressed by a particular solution of the differential equation, or mechanical law, (18). In the present case the differential equation admits a doubly-infinite family of particular solutions, so that it is compatible with a doubly infinite class of motions.

To obtain the particular motion of the point-mass in any given problem, we must specify the numerical values of the arbitrary constants a and b in the general solution (19). We may easily verify that these values are determined by the initial position of the point-mass and by its

initial velocity. Calling x_o and v_o the initial position and the initial velocity respectively at time $t = 0$, we obtain

$$a = v_o \text{ and } b = x_o; \tag{20}$$

and our particular solution which defines the actual motion is

$$x = \frac{K}{2m} t^2 + v_o t + x_o. \tag{21}$$

Since the initial position and the initial velocity determine the motion, these two items of information, considered jointly, constitute what we have called the "initial state of the point-mass." Thus, the form of the differential equation that expresses the mechanical law has enabled us to establish the nature of the items of information which define a state of the mechanical system. More generally, the state of any system whose history is controlled by a given differential equation, or law, is defined by the aggregate of those items of information which suffice to single out a particular solution from the general solution of the differential equation.

We have just seen that the equations of mechanics (for point-masses) are second-order differential equations of the ordinary type that were discussed at the beginning of the chapter. But in the majority of cases physical laws are expressed by partial differential equations. Even in mechanics these latter equations are often met with. Let us, then, examine under what conditions the two different kinds of differential equations will occur in mechanics.

The position of a point-mass or of an extended solid is defined by magnitudes which vary solely with time. Consequently the differential equations which control the motion will involve functions of a single independent variable (time). We conclude that differential equations of the simpler type must be expected when we are dealing with the mechanics of point-masses or of solid bodies. Partial differential equations in mechanics occur in connection with the motions of continuous deformable media. A simple illustration is found in the vibrations of strings.

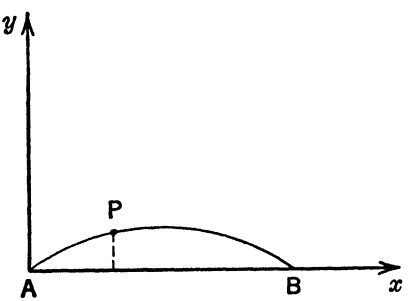


FIG. 17

Consider a vibrating string fixed at its two extremities A and B. Let P be any point of the string. We call x and y the coordinates of this point. As time passes, the elevation of the point P varies, and hence y is a function of the time t. But at any

instant the elevation y also depends on the position of the point P along the string; hence y is also a function of x. Thus, y is a continuous function of both x and t. In order to determine how the string will vibrate, we must obtain the value of y for all values of x (over the length of the string) and for all values of the time t. The vibrations are caused by the restoring force developed by the tension of the string. When to this force we apply Newton's second law of motion, we obtain a differential equation involving the derivatives of the function y with respect to both variables x and t. We thus have a partial differential equation (called d'Alembert's equation), *viz.*,

$$a^2 \frac{\partial^2 y}{\partial x^2} = \frac{\partial^2 y}{\partial t^2}, \tag{22}$$

where a is a constant connected with the tension of the string. Incidentally, this is the equation which Bernoulli solved by expressing the solution function $y(x, t)$ as a Fourier series. We refer the reader to page 134, where the importance of Bernoulli's solution is discussed.

Similarly the equation of wave disturbances in a homogeneous elastic medium is given by d'Alembert's partial differential equation

$$\frac{\partial^2 \theta}{\partial x^2} + \frac{\partial^2 \theta}{\partial y^2} + \frac{\partial^2 \theta}{\partial z^2} = \frac{1}{V^2} \frac{\partial^2 \theta}{\partial t^2}, \tag{23}$$

where θ is the dilatation and V the velocity of propagation.

To summarize: All differential equations in which time is one of the independent variables express the continuity of change. The differential equations of the simpler type are associated with the motions of point-masses and rigid bodies; they furnish the proper mathematical instrument to be used if we adhere to an atomistic conception of Nature. Partial differential equations arise when we are dealing with continuous, deformable media. Interesting conclusions may be drawn from these statements. For instance, an elastic jelly is assumed to be formed, in the final analysis, of tiny molecules acting on one another and separated by empty spaces. From a macroscopic standpoint, however, the jelly appears to be a continuous substance. If we view the jelly as continuous, the equations that regulate the propagations of disturbances through its substance will be partial differential equations. If, on the other hand, we suppose that the jelly is atomistically constituted, we must consider the motions of each one of the individual corpuscles. With each corpuscle are then associated three differential equations of the ordinary type (we say three because of the three coordinates x, y, z defining the position of each corpuscle). The physical analogy between a continuous jelly and one atom-

istically constituted leads us to surmise that a partial differential equation may in certain cases be equivalent to an infinitely large number of differential equations of the simpler type. This assumption is indeed correct, but of course it must be proved mathematically and cannot be justified by mere physical intuition.

The conclusion to be derived from this example is important. It shows that even if a medium be discrete, provided it be composed of an enormous number of corpuscles in each tiny volume, we are perfectly justified from the standpoint of macroscopic treatment in viewing it as continuous, not solely on physical grounds but also from the standpoint of mathematical treatment. We are therefore justified in introducing partial differential equations in mechanics, even if we adhere to a corpuscular structure of matter. This fact simplifies considerably the mathematical investigations, for though a partial differential equation is more complicated than a differential equation of the simpler type, it is easier to manipulate a single partial differential equation than to treat the aggregate of an enormous number of the simpler equations.

The examples given thus far are drawn from mechanics, but quite generally we shall meet with partial differential equations whenever we are considering laws that regulate the continuous distribution of a magnitude through space. In the majority of cases the magnitude at a fixed point will also vary with time; and as a result the partial differential equation will also involve time-derivatives. Equations (22) and (23) are illustrations of partial differential equations involving time. Fourier's equation for the flow of heat is likewise of this type. It is expressed by

$$\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2} = a \frac{\partial T}{\partial t}, \tag{24}$$

where T is the temperature throughout the volume, and a is a magnitude characterizing the solid. Laplace's equation, which regulates the permanent distribution of temperature, is a restricted form of (24). Thus, if the temperature distribution is assumed permanent and hence does not vary with time at any fixed point of the volume, the derivative $\frac{\partial T}{\partial t}$ always vanishes. In this event, (24) becomes Laplace's equation (16).

Quite generally, the equations (or laws) of the field theories are partial differential equations. Maxwell's equations of electromagnetics and Einstein's gravitational equations are examples. The reason field theories lead to partial differential equations is that they deal with the continuous distribution of magnitudes throughout a continuum, such as space or 4-dimensional space-time. Indeed all the laws of the general

theory of relativity are partial differential equations, and in this theory we are confronted with the simpler equations only when we wish to determine the motions of individual point-masses, such as the planets.

There are other physical laws, antedating the quantum theory, which cannot be expressed by means of differential equations of either of the two types. Laws of this kind have been revealed in the study of elasticity. Experiment shows that the evolution of an elastic system is not always determined solely by its initial state but also by all previous states: a kind of ancestral influence must be appealed to. Volterra, to whom the mathematical study of these elastic phenomena is due, has proposed calling the laws controlling them "hereditary laws." Such laws cannot be expressed by the differential equations mentioned in the present chapter; they require a new mathematical instrument, called the *integro-differential equation,* the study of which we owe chiefly to Volterra. As has, however, been remarked by Painlevé, the reason the knowledge of the initial state alone is insufficient to determine the evolution of the elastic system is that we do not define the word "state" with sufficient completeness. If, instead of viewing a state as defined by a small number of magnitudes, we assumed that it was defined by an appropriate choice of an infinite number of magnitudes, we should be justified in saying that the initial state does determine the evolution of the elastic system. Painlevé's attitude seems to be justified on mathematical grounds, for an integro-differential equation is equivalent to an infinite number of ordinary differential equations; and, if we replace the integro-differential equation by an infinite aggregate of ordinary ones, we may quite well claim that the evolution of the system is determined by the initial state alone. However, we must remember that the initial state now requires the enumeration of an infinite number of magnitudes. The physical significance of Painlevé's observation appears to be that, if we view the elastic body atomistically and therefore consider an extremely large number of ordinary differential equations regulating the motions of an extremely large number of corpuscles, we may dispense with the hereditary feature which experiment seems to suggest.

In the earlier days of the quantum theory, when the sudden jumps characteristic of quantum processes were first discovered, physicists were of the opinion that differential equations, because of their association with continuous processes, could no longer be applicable in theoretical physics. Other types of equations, known as difference-equations, were suggested as appropriate. These equations connect states that differ by finite, instead of by infinitesimal, amounts. In the more recent form of the quantum theory, where rigorous laws are no longer held to govern the

courses of individual processes, differential equations are also useless. However, if we consider statistical results and not individual processes, continuity and rigorous determinism reappear. A continuous fluid of a fictitious nature, sometimes called "the probability fluid," is introduced in wave mechanics; and the motions of this fluid in space or in hyperspace (multidimensional space) are regulated by rigorous laws. In view of the fluid's continuity throughout space and time, partial differential equations express the laws. Schrödinger's wave equation of the hydrogen atom is of this type. As for the differential equations of the ordinary kind, they never arise in wave mechanics. The reason is clear: it is that the motions of individual particles are no longer assumed to be submitted to rigorous laws.

CHAPTER XVI

THE CONTROVERSIES ON THE NATURE OF MATHEMATICS

INASMUCH as mathematics is practically the only instrument that can be applied in the construction of physical theories, the nature of mathematical reasoning is a subject of interest to the physicist. Various definitions of mathematics have been proposed by different writers, but the consensus of opinion is that none of them is entirely satisfactory. The fact is that one cannot compress into a few words the definition of so vast and varied a subject. One of the definitions suggested is due to Benjamin Pierce. It runs: "Mathematics is the science which draws necessary conclusions." Although Pierce's aphorism is not patently incorrect, it may convey a misleading impression to the unwary.

In the first place, a conclusion can be termed necessary only in virtue of rules of reasoning which we may see fit to adopt. A proper definition of mathematics should therefore throw some light on the nature and origin of these rules. Pierce presumably supposes that they are the rules of logic, so that his definition is tantamount to identifying mathematics with logic. But we believe we are correct in saying that none of the leading mathematicians of the present day would subscribe to this identification. Finally, if mathematics is truly but the science which draws necessary conclusions, why is it that every man who has devoted his time to mathematics, and who presumably has a clear mind, has not revealed himself as a creative mathematician? We shall first seek an answer to this question.

There are many universities, and in each of these universities there are many men who deliver courses in elementary and in advanced mathematics. Many of them have written textbooks, which in view of their excellence have been translated into other languages. We may presume therefore that these men have a considerable knowledge of their special provinces of mathematics. And yet we may count on our fingers the number of creative mathematicians living today who have evolved new methods and established new theorems instead of merely commenting on the work of others. This is strange if, as Benjamin Pierce appears to believe, mathematics is nothing but the drawing of necessary conclusions. Also, how does it happen that an eminent logician like Bertrand Russell, who is

constantly talking about mathematics, should never have furnished a single theorem, whereas Poincaré, whose hostility to the identification of mathematics and logic is well known, should have been one of the most prolific creators of all time? Are we to suppose that Mr. Russell is unable to draw necessary conclusions? We do not believe that this supposition would be warranted. In our opinion the answer to our question must be sought in a totally different direction.

Some hint is secured when we examine a mathematical demonstration for the first time. We read it over, understand it step by step, and are perfectly convinced of the validity of its proof. But suppose now we are required to repeat the demonstration, even a few hours later. We often find this impossible, and we excuse ourselves by saying that we do not recall the demonstration. Yet, if a mathematical demonstration involves nothing more than perceiving that p implies q, we may well wonder what part memory has to play in the matter. If A is greater than B and B is greater than C, we do not have to appeal to our memory to assert that A is greater than C. Anyone who wishes to test what we have just said can easily do so without any particular knowledge of mathematics. He has but to revert to Euclid's proof of the unending sequence of prime numbers or of the irrationality of $\sqrt{2}$. Both of these proofs are mentioned in Professor Dantzig's admirable book, "Number," which will long serve as a model of semipopular exposition in mathematical philosophy.*

This leads us to inquire: What do we have to recall in a mathematical demonstration in order to repeat it? On reverting to the demonstration which we have forgotten, we usually find that our memory has failed at a point where some expedient known as a mathematical artifice has been introduced. To take a simple example, Euclid's proof of the unending sequence of prime numbers starts with the introduction of factorials. Euclid shows that if n is any prime number, a greater prime number always exists. He establishes this theorem by considering the numbers defined by $n!$ and by $n! + 1$. Regardless of the sequel of Euclid's demonstration, his initial introduction of factorials is an artifice and is certainly not dictated by any logical necessity. Should the reader forget this first step, all the logic in the world would not enable him to reproduce the demonstration. Euclid does not tell us how he came to think of factorials. The idea may have come to him suddenly, like a flash, or it may have arisen as the result of prolonged reflection. These are the idiosyncrasies of the creator; they pertain to psychology, not to logic. Such

* T. Dantzig. Number. Macmillan, New York, 1930; pp. 48 and 102.

artifices are very frequent in the writings of the masters. In many cases they are devised because some seemingly insuperable difficulty blocks the way. An attempt is then made to transform the problem into an equivalent but easier one. Riemann, for instance, sought to solve a certain problem in the theory of functions when the area under consideration was that of an arbitrary continuous closed curve. But he did not solve the problem directly; he transformed it into a simpler one by transforming the arbitrary closed curve into a circle. With the circle, the solution is relatively easy and hence the original problem was solved. If we forget that Riemann made a transformation at this point, or even if remembering that a transformation was made we forget its precise nature, we shall probably be unable to reproduce the proof even though we once understood it. Mathematical transformations are often extremely ingenious, and on coming across one for the first time we are tempted to exclaim: "How did anyone ever think of that?" We cannot suppose that the great masters experimented with all conceivable transformations before hitting on the appropriate one; human life is too short to permit this procedure. Nor can we suppose that sheer luck invariably guided these men to the correct choice. What then is the secret of their success? Some call it genius; others intuition; but whatever it be, let us not be confused into thinking that it is merely an instance of "*p* implies *q*." If it were, we should have to admit that, except for a handful of creators, all the present-day mathematicians are logical imbeciles.

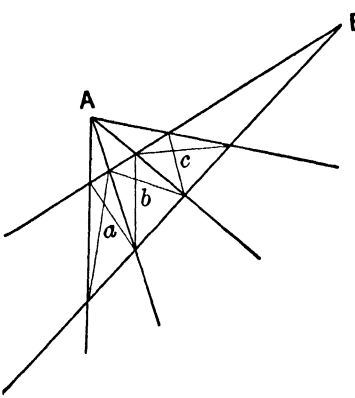


FIG. 18

A simple example of an ingenious transformation in elementary geometry illustrates the thoughts we have been expressing. The example is well known; it occurs in connection with the following problem.

Four straight lines issuing from a point *A* are intersected by two straight lines issuing from a point *B*. We draw the diagonals in the three quadrilateral figures thus formed, and we thereby obtain the three points of intersection marked *a*, *b*, and *c*. We are required to prove that these three points lie on the same straight line, *i.e.*, are collinear (see Figure 18).

This theorem may be proved by analytical geometry or by straightforward geometry, but in either case the proof would be laborious. However, there is a transformation which, if we are lucky enough to think

of it, transforms our problem into an equivalent one of the utmost simplicity. The artifice consists in projecting the figure in an appropriate way onto another plane. Thus suppose the sheet of paper on which our figure is traced is set in a vertical position, with the line AB placed along a horizontal. Consider now a point-source of light S situated at the same elevation as the horizontal line AB. This point-source of light casts a shadow of the drawing on a horizontal sheet placed below the figure. The shadows of the straight lines will be straight lines, and as a result the four straight lines issuing from A will have as shadows four straight lines issuing from the shadow of A. But the shadow of A is at infinity; hence the four straight lines issuing from A will have as shadows four straight lines meeting at infinity; and this implies that the four shadow lines will be parallel. The same argument applies to the two lines issuing from B. Altogether, the shadow of our drawing will consist of four parallel straight lines intersected by two other parallel straight lines (see Figure 19).

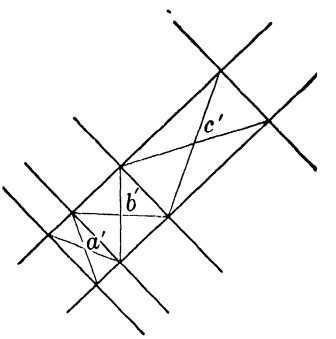


FIG. 19

Since the shadow of a straight line is a straight line, and vice versa, we see that the three points, a, b, c, will be collinear if their shadows, a', b', c', lie on a straight line. We shall therefore have solved our original problem if we can establish the collinearity of the points, a', b', c'. Thus our original problem is transformed into a new one. Now the new problem, in contradistinction to the original one, is absurdly simple, for the points, a', b', c', obviously lie on a straight line. We conclude that a, b, c are collinear, and so our original problem is solved. Thanks to an ingenious transformation the solution has been obtained without difficulty.

Although we have expressed the artifice in physical language so that the non-mathematical reader may follow it more easily, the demonstration can be made perfectly rigorous. In short, by means of an artifice we have transformed a difficult problem into a simple one. In our opinion the student who would think of this simple artifice would prove to have greater mathematical ability than the one who might solve the same problem by more laborious but less ingenious methods.

These examples and many others would seem to indicate that the conception of the creative mathematician, as implied in Pierce's definition, is radically wrong. Between the creative mathematician and the logician there is a vast difference. The logician must be able to concentrate and

to think clearly. So also must the mathematician; but, in addition, the creative mathematician needs imagination, or genius, or whatever we wish to call it. He must be an artist, not solely a worker with a one-track mind. Imagination to the logician, if not harmful, would at any rate be excess baggage. No imagination is required to realize under what conditions "*p* implies *q*." Possibly more than a coincidence is responsible for the well-known fact that those writers who have been most insistent in identifying mathematics with logic are usually found in that larger group of men who have failed to create. To the intrusion of extra-logical influences in mathematics is presumably due the marked differences in the presentations of the same subject by different mathematicians. The trained student may usually detect very easily by the form of the presentation whether a demonstration is due, say, to Riemann or to Weierstrass. A similar situation arises of course in poetry and in music; but there it is so familiar that it scarcely warrants mention.

We may summarize this discussion by saying that mathematics involves two steps: firstly, the application of rules; secondly, *the ability to combine the rules.* The second step, however, cannot itself be expressed by rules.

Poincaré, discussing the various qualifications required by a creative mathematician, writes:

"Others will commit to memory the details one after another; they will be able to understand mathematics and sometimes to apply it, but they will be unable to create. Still others will have, in a more or less marked degree, that special form of intuition of which I have just spoken, and then they will not only be in a position to understand mathematics, even though their memory be none too good, but they may also become creators and seek to invent, with more or less success, according to the degree of development of their intuition.

"What indeed is mathematical invention? It does not consist in setting up new combinations with mathematical beings that are already known. For anyone could do this, and the combinations that might thus be formed would be infinite in number and the majority of them without the slightest interest. To invent consists precisely in constructing not useless combinations, but useful ones, and these form an infinitely small minority. To invent is to discern, to select.

"Sterile combinations will not even suggest themselves to the mind of the inventor. In the field of his consciousness will appear only those combinations that are really useful, some few of which he will reject but which nevertheless have something in common with the useful ones." *

*The quotations from Poincaré in this chapter are translations of passages appearing in his four popular books: *La Science et l'Hypothèse, La Valeur de la Science, Science et Methode,* and *Dernières Pensées.*

Some people will contend that those characteristics of a mathematical demonstration that we have been stressing are irrelevant to the nature of mathematics itself, and that a confusion is being made between the qualifications of the creator and the subject matter he is manipulating. This contention may be defended, but the point is that if the man who is a mere logician cannot develop into a creative mathematician, we must conclude that the methods of logic alone are insufficient to construct mathematics. Under these conditions how can the subject matter of mathematics be exhausted by logic alone? If the bare rules that are applied in mathematics are to be viewed as constituting the subject matter, regardless of whether or not we be capable of applying these rules and obtaining significant results, then according to the same criterion we should be justified in saying that music is nothing but the rules of counterpoint and harmony. We may grant, however, that a study of the qualifications of the creative mathematician is not the only way of obtaining an understanding of the nature of mathematics. More direct means of approach can be resorted to. These we shall now consider, and in the course of the chapter we shall summarize the conclusions to which the most competent authorities have been led.

The creative mathematicians of international fame who have investigated the foundations of mathematics are, in the historical sequence of their writings, Poincaré, Peano, Hilbert, Brouwer, and Weyl. To these we should add Borel, Hadamard, Lebesgue, and Baire, all of whom have expressed valuable opinions. Many of the great masters are not included in this list, Volterra and Picard for example. This is because their discoveries have been confined to the more advanced regions, and, so far as we know, they have expressed no definite opinions on the foundations. The fact is that mathematics offers a vast field of unsolved problems and unexplored territory; the foundations do not constitute the sole sphere of interesting research.

We are particularly fortunate to find Poincaré and Hilbert among those who have discussed the foundations; for Poincaré is universally recognized as one of the greatest mathematical minds of all time, and Hilbert is one of the greatest among the living. Yet in spite of their unquestioned competence, Poincaré and Hilbert are at loggerheads; and the same is true of all the other masters mentioned: no two of them agree on all points. This situation creates a mathematical scandal, for we have been taught to view mathematics as a science in which a rigorous demonstration should always put an end to any dispute. However, we must remember that the foundations of mathematics, and mathematics proper,

are not the same thing. To stress the difference, Hilbert refers to the former as "metamathematics," a very appropriate name.

If we neglect minor differences of opinion, we may class the leading thinkers into two groups: the "formalists" and the "intuitionists." Between these two schools, mathematical thought is divided. The appellation intuitionist is likely to confuse the layman. He may be tempted to suppose that the mathematicians of the intuitionist school propose to rely on "hunches" in mathematics instead of appealing to rigorous mathematical demonstrations. It is scarcely necessary to say that no such absurdity is intended. As a matter of fact quite the reverse is true: the intuitionist is a rigorist. We shall resume the thread of this discussion when we examine the points of disagreement between the two schools. For the present we shall confine our attention to the results which appear to have been accepted unanimously. One of these is that mathematics and logic are not identical. This was always Poincaré contention, as the following passage shows:

"If all the propositions which mathematicians enunciate can be deducted one from the other by the rules of formal logic, why is not mathematics reduced to an immense tautology? The syllogism can teach us nothing that is essentially new, and, if everything is to spring from the principle of identity, everything should be capable of being reduced to it. Shall we then admit that the theorems which fill so many volumes are nothing but devious ways of saying that A is A?"

Let us now consider an expression of opinion from the opposing school. Hilbert writes:

"Arithmetic is indeed designated as a part of logic, and it is customary in founding arithmetic to presuppose the traditional fundamental principles of logic. But, on attentive consideration, we become aware that, in the usual exposition of the laws of logic, certain fundamental concepts of arithmetic are already utilized, for example the concept of aggregate and, in part, also the concept of number. We thus fall into circularity, and, therefore, to avoid paradoxes a simultaneous development of the laws of logic and arithmetic is requisite."

Thus, the formalist views logic and arithmetic as complementary, and not as identical.

Another point on which agreement has been arrived at (though only in part) concerns the impossibility of defining number, integers in particular. We say that the agreement is only partial because, though both schools agree that a direct definition of number in terms of logical concepts is impossible, the formalists claim that an indirect definition, based

on a postulate system, is feasible. The following passage of Poincaré illustrates the attitude of the intuitionists:

"The definitions that have been suggested for 'number' are numerous and varied. I cannot even attempt to give the names of their authors. We need not be surprised at the plethora of definitions suggested. For if one of these definitions were satisfying, no further one would be required. If each new philosopher who has been interested in this matter has found it necessary to invent some new definition, it is because he was dissatisfied with those of his predecessors, and the reason he was dissatisfied was that he regarded the earlier definitions as circular."

Similar opinions have been expressed by the other intuitionists. The attitude of the formalists is much the same. Hilbert, for instance, discusses Frege's definition of integers* and also several others; but he rejects them all as unsatisfactory.

Though both schools of mathematicians recognize that a direct definition of number cannot be given, the formalists contend that an indirect definition (by the method of axiomatics) is possible. Peano's celebrated postulate system, which yields an indirect definition of integers, is a case in point. The intuitionists deny even this more modest claim. Hadamard, who occupies an intermediary position between the more extreme members of the two schools, states in his preface to a book by Gonseth on the foundations of mathematics:

"The notion of integer appears to Poincaré as a notion which is inherent in the mind and without which it would be impossible to think. Others have since implicitly upheld the contrary opinion, for they have tried to define integers. May I be permitted to be of Poincaré's opinion? And when those who state their definitions proceed to number their postulates, I am not so sure that the circularity which thus appears is purely on the surface. This criticism, however, does not prevent me from regarding the axiomatics of the integer as an unquestionable and enormous progress, provided we do not expect more than it, or than anything else, can yield."

Axiomatics—The new science of axiomatics, developed chiefly by Peano and by Hilbert of the formalist school, constitutes the one bright spot in these obscure controversies on the nature of mathematical reasoning. Although axiomatics was conceived for the purpose of clarifying the foundations, it has proved itself of use in genuine mathematics. The germ of axiomatics can be traced to the system of axioms and postulates with which Euclid prefaced his geometrical discoveries. The ancient

* Frege defines the number "two," for instance, as the class of all couples.

Egyptians and the earlier Greeks presumably discovered many of the geometrical theorems (such as those of Thales and Pythagoras) by direct measurement with rods, no rigorous demonstrations being given. But Euclid introduced rigor by deducing his geometrical theorems from clearly stated assumptions. These initial assumptions were embodied in his axioms and postulates.

Today we know that the assumptions explicitly stated by Euclid were insufficient to serve as a basis for Euclidean geometry. Yet Euclid constructed his geometry. We are thus faced with a paradox, for if the assumptions were insufficient, how could a geometry be constructed? The answer is clear when we find that Euclid, in addition to the assumptions which he stated explicitly, utilized others unconsciously in the course of his demonstrations. For example, in one of his theorems Euclid seeks to establish the equality of two triangles. To do so, he imagines that one of the triangles is placed on the other, and he then shows that the two triangles must necessarily coincide and hence be equal. Yet none of the axioms or postulates explicitly stated by Euclid entitled him to operate in this way; and in proceeding as he did he was unconsciously assuming that his triangles could be displaced without changing in shape and size, an assumption to which he was undoubtedly guided by the behavior of the rigid bodies observed in his daily activities. Incidentally, we have here a first indication of the empirical origin of Euclid's assumptions. We shall revert to this matter later.

Summarizing Euclid's idea, we see that it consists in stating a certain body of assumptions and then in deriving from these assumptions their necessary consequences by logical means. Euclid's execution of this idea was defective because, as we have said, he inadvertently omitted to state all the assumptions which he subsequently utilized in his demonstrations. Nevertheless the idea in itself has since been proved valid, and it forms the essence of the new science of "axiomatics."

A point of nomenclature must here be mentioned. Euclid made a distinction (which was none too clear) between axioms and postulates. Today this distinction is no longer recognized, and all the assumptions on which a doctrine is founded receive the same name. Hilbert uses the word "axiom" whereas English-speaking writers prefer the word "postulate." We shall adopt the latter.

The method of axiomatics was first applied with rigor by Peano, when he gave a system of postulates serving to define the integers of arithmetic. Hilbert pursued the same method for Euclidean geometry. He introduced 21 postulates which defined certain relations among points, straight

lines, and planes. The aggregate of these postulates is called "Hilbert's postulate system." Hilbert then showed that, by applying the logical principles to his postulate system, he could deduce all the theorems of Euclidean geometry without having recourse to any assumption not explicitly stated in the postulate system. As may be gathered, Hilbert's postulates comprise Euclid's axioms and postulates (though in a modified form) and also those assumptions which Euclid made use of unconsciously in his demonstrations.

An example of one of Euclid's tacit assumptions, which Hilbert states explicitly in the form of a postulate, is Hilbert's postulate of continuity, also called the postulate of Archimedes. The significance of this postulate is as follows: If A and B are two points on a straight line, and if from A we mark off towards B equal intervals one after the other along the line, we shall eventually reach points beyond the point B. The postulate of Archimedes becomes intuitive when we use a rigid rod to mark off the equal intervals. And since Euclid was guided unconsciously by the properties of rigid bodies, we may well understand why he took this postulate for granted and did not state it explicitly. We have here a further example of the empirical origin of Euclidean geometry.

Other typical postulates of Hilbert, taken at random, are listed below:

(1) "Two distinct points, A and B, always determine a straight line."
(2) "Three distinct points, A, B, and C, that are not on a straight line always determine a plane."
(3) "Of three points on a straight line, one and only one is between the other two."
(4) "If a segment AB is congruent (equal) to a segment $A'B'$ and to a segment $A''B''$, then $A'B'$ is congruent to $A''B''$." *

These samples suffice for our purpose.

Hilbert's postulates in their aggregate express the relations which the points, straight lines, and angles of Euclidean geometry are assumed to satisfy. But points, straight lines, planes, and angles are not the only entities which satisfy these relations; and with a little ingenuity many others may be found. A particularly simple illustration can be given in the case of the Hilbert postulates which apply to plane geometry; these differ but slightly from the former postulates which correspond to space geometry. The illustration follows:

By a point in the plane we shall mean any point of the plane, with the exception of one particular point O chosen arbitrarily. By a straight line in the plane we shall mean any circle passing through the point O.

* Hilbert, Foundations of Geometry, Chicago.

The distance between two points, A and B, on one of our new straight lines (now circles) is defined by

$$\cot O\omega B - \cot O\omega A,$$

where $O\omega B$ and $O\omega A$ are the angles (in the ordinary sense) obtained by joining O, A, and B to the centre ω of the circle. Finally, the angle between two straight lines (now circles passing through O) is defined by the angle (in the ordinary sense) made by the two circles. It can then be verified that all the relations satisfied by the points and straight lines of plane Euclidean geometry will also be satisfied by our new "points" and "straight lines" (and vice versa). We may therefore view plane Euclidean geometry and the model just defined as two different models, or so-called "concrete representations," both of which satisfy Hilbert's postulates for plane Euclidean geometry.

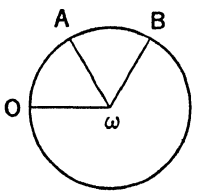


FIG. 20

At first sight, it seems absurd to call the circle traced in the figure a straight line. However, our hesitation to do so brings out precisely one of the aspects that Hilbert is stressing, for Hilbert requests us to refrain from assigning any *a priori* properties to the points and straight lines that he mentions in his postulates. Points and straight lines are to be viewed as representing mere undefined terms, which we may conveniently refer to as a's and b's. Under these conditions Hilbert's postulates, say for Euclidean geometry in space, may be rewritten by replacing everywhere in the postulates the terms point, line, and plane by the symbols a, b, and c. If in these postulates, which now express relations connecting the a's, b's, and c's, we replace these a's, b's, and c's by the words points, lines, and planes, we obtain Euclidean geometry; and if we replace them by other appropriate words representing other entities whose mutual relations happen to be the same as those of the points, straight lines, and planes of Euclidean geometry, a new model satisfying the same postulates is obtained. Superficially, the original Euclidean model may differ considerably in appearance from the new one; yet, beneath the surface, both models have the same form,* or logical structure, since in both of them the entities which are substituted for Hilbert's a's, b's, and c's exhibit the same relationships, expressed by Hilbert's postulates.

Hilbert gives an important example of a group of elements which exhibit the same relationships as do points, straight lines, and planes. The new elements are represented by triads of numbers and by algebraic

* Technically, "are isomorphous."

expressions connecting these numbers. He was led to this example when, in place of points, straight lines, and planes, he took their Cartesian representations in terms of numbers x, y, z. The fact that the new elements, here numerical, satisfy the Hilbertian postulates merely proves the equivalence of straightforward geometry and of the Cartesian method of analytical geometry. In other words, it proves the equivalence, as to logical structure, of the geometric continuum and of the arithmetic continuum. From considerations such as the foregoing, we see that there is no particular reason to associate Hilbert's system of postulates with the familiar model of Euclidean geometry rather than with any other model having the same logical form. What is characteristic of a postulate system is the form, the network of relations, and not the subject matter, or content; the subject matter is contingent.

These conclusions of Hilbert's have a familiar ring, for long before Hilbert constructed his postulate system mathematicians realized that the concern of mathematics lies in relationships, and not in subject matter. An example is afforded by Laplace's equation. Laplace's celebrated equation expresses a relationship connecting the derivatives of a function of three variables x, y, z. The relationship happens to be realized in the physical world by the Newtonian potential distributed through space; and it was in this connection that Laplace came upon the equation. But as Fourier subsequently showed, exactly the same equation, or relationship, holds for the distribution of the temperature in a solid when a condition of heat equilibrium is attained. It also holds for the velocity potential in the case of an incompressible fluid in a permanent state of flow. The equation itself and its peculiarities transcend the particular verifier which happens to fit it. We have here an additional illustration of the fact that mathematics deals with relations, and not with content. And so we may wonder why such prominence has been given to Hilbert's conclusion: namely, that from his postulate system we can construct Euclidean geometry even when we have no idea of what is meant by a point, a straight line, and a plane.

To obtain the answer to this question we must recall that Euclid deduced his theorems by utilizing, in addition to his axioms and postulates, certain tacit assumptions which were suggested by the properties of rigid bodies. If, then, we were to base our judgment solely on Euclid's demonstrations, we might suppose that Euclidean geometry could not be constructed without a prior knowledge of its subject matter, *i.e.*, the points, lines, and planes with which the study of rigid bodies makes us familiar. According to this view, geometry in contradistinction to other mathematical doctrines would appear to be inseparable from its content. The

significance of Hilbert's investigations has been to show that this belief would be erroneous, and that, by reasoning on *a*'s, *b*'s, and *c*'s devoid of all content but satisfying well-defined relations, we could obtain Euclidean geometry. To Hilbert therefore goes the credit of having shown that in geometry, as in other parts of mathematics, the specific content is of no particular importance.

Einstein, commenting on Hilbert's discoveries, contrasts the older and the newer views by saying:

"The older interpretation:—Every one knows what a straight line is, and what a point is . . .

"The more modern interpretation:—Geometry treats of entities which are denoted by the words straight line, point, etc. These entities do not take for granted any knowledge or intuition whatever, but they presuppose only the validity of the axioms (postulates), such as the one stated above, which are to be taken in a purely formal sense, *i.e.*, as void of all content of intuition and experience. . . ." *

The Mathematical Utility of Postulate Systems—Axiomatics has been of use not only in clarifying obscure points in the foundations, but also in leading to constructive discoveries. We may readily understand how this situation has arisen. Take Hilbert's postulate system for example. As it stands, it defines indirectly the model of Euclidean geometry or any equivalent model (in the sense previously stated). Suppose now we withdraw one of the postulates and replace it by a different one. We thus obtain a new postulate system, for which the former Euclidean model is no longer valid. But we may find that the new postulate system is verified by some new geometry, or model; and in this event the method of axiomatics will have entailed the discovery of some novel geometrical doctrine. For instance, Hilbert showed that when the postulate of parallels is replaced by an appropriately selected one, the new postulate system is verified by Lobatchewski's geometry. Since, however, Lobatchewski's geometry was already known to Hilbert, nothing new was discovered thereby. But a totally new kind of geometry was secured when Hilbert denied the postulate of Archimedes. Thus, by juggling the postulates, we may obtain geometries, and more generally models, which might never have been thought of. It is in its capacity to furnish new doctrines that axiomatics reveals its constructive side.

A point that may seem strange is that axiomatics should have come to the fore as a constructive branch of mathematics. In this connection we must remember that Riemann's great contribution to the develop-

* Einstein. Sidelights on Relativity. E. P. Dutton and Co., 1920; p. 29.

ment of non-Euclidean geometry consisted precisely in his supplanting the narrow axiomatic method of dissecting postulates (as practised by Bolyai and Lobatchewski) by the more powerful methods of analysis. In this way Riemann generalized the various possible types of non-Euclidean geometry; from these generalizations the theory of relativity draws its material. Weyl's new geometry is also an example of a doctrine discovered by following Riemann's method rather than the postulate one. Thanks to Hilbert, however, the axiomatic method is also seen to have its advantages.

The possibility of replacing Hilbert's postulates by others and of obtaining verifiers for the new postulate systems has clarified the significance of these postulates. As late as the early nineteenth century, it was permissible to suppose that Euclid's axioms, postulates, and tacit assumptions, though suggested by experience, were imposed by the requirements of the human mind, *i.e.*, were *a priori* in the Kantian sense. Had this view been justified, no geometrical doctrine which rejected the Euclidean postulates and assumptions could have been conceived. But when Lobatchewski and Riemann rejected some of Euclid's postulates and succeeded nevertheless in constructing geometries, mathematicians realized that the rejected postulates, at any rate, could not be *a priori*. The status of the other postulates and of the tacit assumptions, all of which were retained in the new geometries, was, however, undisclosed; they could still be regarded as *a priori*. Finally, when Hilbert, after grouping all Euclid's postulates and tacit assumptions into one class, modified his postulates one by one and then showed that doctrines verifying the new postulate systems could be constructed, mathematicians were constrained to reject the Kantian thesis not merely for some of Euclid's postulates, but for all of them and also for all his tacit assumptions. Thus, as a result of Hilbert's investigations, there was no longer a place for the *a priori* in the concepts that are at the base of Euclidean geometry. From a philosophic point of view, this is probably the most interesting contribution of axiomatics. We shall see, however, that the exclusion of all *a priori* notions from the postulate systems does not necessarily imply that no *a priori* notions enter into the construction of a geometrical doctrine. The fact is that the postulates are not the only elements that must be examined; we must also consider the rules to which these postulates are submitted.

The Compatibility of Postulate Systems—Thus far we have dealt only with the bright side of axiomatics. We must now mention its limitations. One of the aims of mathematicians has been to furnish

so-called *existence theorems.* The purpose of such theorems is to ensure that the mathematical solution we are seeking really exists, for to search for something that does not exist would be a waste of time. But before making more precise the meaning of mathematical "existence," let us examine a non-mathematical illustration.

The metaphysician Meinong is reported to have maintained that since we can say truly "The round square does not exist," there must be such an object as the round square, although it must be a non-existent object. Mr. Russell, who mentions this example, confesses himself puzzled and comments on it by saying:

"The present writer was at first not exempt from this kind of reasoning, but discovered in 1905 how to escape from it by means of the theory of 'descriptions,' from which it appears that the round square is not mentioned when we say, 'The round square does not exist.' "

We shall leave Mr. Russell with his theory of "descriptions," presumably to be found in his "Principia." What we shall be concerned with here is the different meanings that may be attributed to the same word. In ordinary conversation, the two verbs "to exist" and "to be" are regarded as synonymous. For instance, we say that "a unicorn does not exist," or that "there is no such animal as a unicorn." Here both verbs imply that no unicorns can be found, *i.e.*, that they have no empirical existence. In mathematical discussion, the verbs "to exist" and "to be" are still regarded as synonymous, but they now imply "absence of any logical contradiction." According to this understanding we are equally well justified in saying: "The round square does not exist," or "There is no such thing as a round square." The two propositions are equivalent.

Now, if we revert to Meinong's puzzle, we find that "to exist" and "to be" are no longer regarded as synonymous; the former verb is taken to imply "absence of contradiction," whereas the latter verb is applied to anything that may play the part of the subject in a proposition. There is therefore no puzzle to Meinong's example when due attention is paid to the meaning of words determined implicitly by the context. The puzzle arises only if the reader is careless enough to overlook the fact that the verbs "to exist" and "to be" are not credited by Meinong with the same meaning. If Meinong's example is taken seriously, it denotes, in our opinion, an inability to think clearly. If viewed as a kind of a pun, it is of the same type as the following one: "Where does the light go when it goes out." Here also the verb "to go" is not used in the same sense in the two cases. The moral is that whatever meaning we may attribute

to a word and to its synonyms, we must be consistent and avoid introducing surreptitiously different meanings, at least in the same sentence. Symbolism has sometimes been suggested as a safeguard against confusions of this sort; but symbols merely replace words, and if we are unable to differentiate the various meanings of the same word, we shall be no better off with symbols.

Let us pass to a more important subject. When a mathematician tells us that a certain problem has a solution, or that a solution exists, he means that no logical contradiction prevents us from assuming the existence of the solution. An existence theorem is a rigorous proof that the solution exists, even though we may be unable to discover it.

One of the most celebrated problems in mathematics is an existence theorem. It is known as Dirichlet's problem, and over it even the great Riemann stumbled. Briefly, the problem is to determine whether or not there will always exist a solution of Laplace's equation, which satisfies certain restrictions and certain boundary conditions. In other words does the aggregate of relations expressed by Laplace's equation and by the restrictive conditions form a consistent whole. Riemann, in a celebrated demonstration, thought he had established the existence of a solution and hence the consistency of the aggregate of relations. But his demonstration involved two unwarranted assumptions. One was pointed out by Weierstrass and another has been uncovered more recently by Hadamard. A physical interpretation of Dirichlet's problem was indicated in the last chapter in connection with the temperature distribution throughout a solid.

Let us return to Hilbert's postulates. When Hilbert constructed his postulate system, one of his aims was to establish the consistency of Euclidean geometry, *i.e.*, to prove that, however far the consequences of the Euclidean theorems were extended, no contradiction would ever arise. Now we have seen that Euclidean geometry furnishes a model which satisfies Hilbert's postulates. Hence we conclude that the consistency of Euclidean geometry will be established if we can show that the relations embodied in Hilbert's postulate system are compatible. If the postulates are not compatible, then Euclidean geometry and also all the other models which verify the postulate system will be inconsistent and will have no greater claim to mathematical existence than have Meinong's round squares. (The problem of proving the compatibility of Hilbert's postulates is thus seen to have very much the same significance as has Dirichlet's existence theorem.) We therefore request the **axiomatician** to furnish us with the required existence theorem for Hilbert's postulates.

Now, quite generally the compatibility of postulates could be tested if the postulate system yielded only a finite number of consequences, for in this case we might pursue the deductions to their end and ascertain whether or not a contradiction occurred. If no contradiction occurred, our postulates would be compatible. We may observe, however, that this method of establishing the compatibility of a system of postulates would be unsatisfactory, since it would constitute an empirical testing and not a general demonstration. Besides, in the case of Hilbert's postulates the course just outlined is impossible because the consequences deducible from the postulates are infinite in number and so cannot be exhausted. Hilbert circumvents this difficulty by reversing the problem. He notes that his system of postulates will be proved consistent if we can establish the existence of a model verifying the postulate system, the word *existence* here implying the absence of any inner inconsistency. He then maintains that the model based on numbers * satisfies this requirement; in other words, he accepts the inner consistency of the arithmetical continuum. Such is the argument advanced by Hilbert in support of the consistency of his postulate system. But the trouble is that we are by no means certain of the consistency of the arithmetical continuum, *i.e.*, of the number system. Indeed, Brouwer and Weyl question it seriously. The net result is that only by an act of faith can we accept the compatibility of Hilbert's postulates † and of any of the models that are thought to verify them. Logic alone is helpless. The significance of Hadamard's reservations on the value of axiomatics, quoted on a previous page, now becomes apparent.

All mathematicians have acknowledged the debt their science owes to Hilbert; and aside from the difficulty we have just discussed, disagreements occur only on minor details. Some of these we wish now to examine. Suppose we start with a postulate system such as Hilbert's. We then deduce the various theorems of Euclidean geometry by purely logical means, without being guided by any extra-logical intuition of what constitutes a point, a straight line, and a plane. These elements may be regarded as vague a's, b's, and c's, serving solely as substrata to the relationships expressed in the postulates. Hence it would seem that no extra-logical elements are present in our mathematical reasoning. We

---

* See page 197.

† The impossibility of establishing the consistency of a postulate system entails the impossiblity of proving that the postulates are independent. For, to prove that two postulates A and B are independent, we must show that the denial of A still leaves us with a consistent postulate system.

must note, however, that these arguments have no bearing on the conclusions we arrived at previously; namely, that mathematics does not consist solely in the rules; that it also embodies the science, or art, of combining these rules; and that this science cannot itself be compressed into a set of rules.

But even if we dismiss these objections, we may still inquire: Is it true that Euclidean geometry can be obtained merely by applying the rules of logic to the Hilbert postulates? Poincaré answers the question in the negative. He observes that in some of Hilbert's demonstrations the "principle of mathematical induction" is utilized, and he maintains that this principle is the expression of some extra-logical intuition of the human mind. Poincaré therefore concludes that geometry cannot be derived from a set of postulates by purely logical means. The validity of Poincaré's conclusion obviously depends on the precise status of the principle of mathematical induction. This principle is of constant use in mathematics; it dates back at least to Euclid, who applied it in his proof of the unending sequence of prime numbers. The principle cannot be a rule of logic, for we may quite well construct a mathematics in which the principle is denied. Let us, then, inquire whether the principle is a mere postulate, which we are free to accept or to reject. Hilbert himself does not regard it as such, since he does not mention it in his list of postulates. Peano, on the other hand, accepts it as a postulate, stating it explicitly as one of his postulates for the definition of integers. Peano is then able to prove the consistency of his postulate system. Poincaré agrees with Peano in the contention that a set of postulates must be proved consistent before the postulate system can be credited with any real significance, but he claims that Peano's proof of this consistency is circular and hence worthless. The fact is that Peano utilizes the principle of induction in a dual capacity: first as a postulate, and then as a rule which he applies when he seeks to establish the compatibility of his postulates. Circularity, according to Poincaré, can be avoided only if we remove the principle of induction from Peano's postulate system and class it with the rules to which the postulates may be submitted. But then, since the principle of induction is not a principle of logic, Poincaré is led to the conclusion we mentioned previously: the principle of induction expresses some extra-logical intuition of the human mind, and hence logic alone is insufficient to construct a mathematics.

The uncertain status of the principle of mathematical induction is, however, not the only argument that can be brought against the identification of mathematics and logic. As we have seen, the methods of axiomatics

require that we start from some postulate system and that we then deduce its logical consequences; axiomatics, as such, throws no light on the motives which prompt us to adopt one postulate system rather than another. But the trouble is that the task of the mathematician does not begin when the postulate system is given; it starts far earlier, in the construction of the system itself. No one, we may presume, imagines that Hilbert's 21 postulates suddenly fell into his lap, or that they were revealed to him through some inner vision, or were adopted by caprice. Hilbert's postulate system, like every other such system, was constructed according to the same routine. The method consists in starting from some previously established doctrine in the consistency of which we have at least a moral certitude. We then pick the doctrine to pieces, examine every step in the demonstrations, and state explicitly all the assumptions utilized. These assumptions are then listed and our postulate system is obtained; care is usually given to retain only a minimum of assumptions and to discard those that are redundant. It was by following this course that Hilbert constructed his postulate system. Euclidean geometry was before his eyes and all he was required to do was to dissect it, analyze every step taken by Euclid, and fill in the omissions.

The reason why constructors of postulate systems usually start from some accepted doctrine which seems consistent is that the majority of postulate systems involve an infinite number of consequences, and postulate systems of this sort cannot be proved consistent. Of course, once a postulate system is constructed, mathematical fancy may come into play; we may change some of the postulates and obtain thereby new doctrines of which we might never have thought. Nevertheless for the reasons stated, the initial postulate system, whence the others are derived, is usually moulded on some previously recognized doctrine.

These conclusions have important philosophical consequences, for we see that, in our attempts to analyze the nature of mathematics, we cannot merely start from a given postulate system and deduce its logical consequences. We must seek the origin of the postulate system, and we are thereby referred to the preestablished doctrine on which the postulate system was moulded. In the case of the Hilbert system, we are referred to Euclidean geometry. No modern mathematician would deny the empirical origin of Euclidean geometry. Indeed the very words "geometry," "cubit," "foot," show conclusively how geometry originated, and so does the importance the Greeks attributed to geometrical constructions that can be performed with the straight edge and compass. An examina-

tion of Hilbert's postulates also leads to the same conclusion. The empirical origin of the ancient's belief in the postulate of Archimedes has already been mentioned; let us give another illustration. Consider the following postulate of Hilbert. In ordinary language, it asserts:

> If A, B, and C are three points in a plane and are not in a straight line, and if a straight line intersecting the segment AB is drawn in the plane, this straight line must intersect the segment AC or the segment BC.

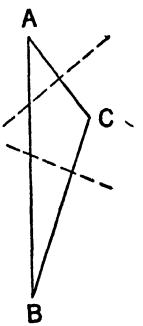


FIG. 21

The postulate appears so obvious that we cannot imagine the possibility of its being denied. Yet, as Hilbert's investigations have shown, the postulate may quite well be rejected. Hence its implicit acceptance by the ancients must be attributed to the fact that it is verified empirically when lines are traced in a plane. The chronological priority of Euclidean over non-Euclidean geometry has not been touched upon in the previous discussion, but we may justify it on empirical grounds. Thus the savage and the child naturally assume that stones and sticks retain the same volumes or lengths when displaced. Consequently, when primitive people sought to divide their fields into equal parts, they invariably utilized rods or ropes in their measurements. Now all measurements which the geometricians antedating Euclid could have performed would have shown that, within the limits of experimental error, similar triangles could exist. It would therefore have been gratuitous for Euclid to have constructed a geometry in which this elementary result of measurement was denied. These considerations suggest that even had non-Euclidean geometry been imagined by the ancients, it would in all probability have been discarded as absurd.

The work of Peano and of Hilbert made it at one time the fashion to seek systems of postulates for every known mathematical theory. The construction of these postulate systems requires ingenuity and is in many respects analogous to the construction of a complicated cross-word puzzle. Unless the mathematical theory which we propose to set into postulate form pertains to an advanced branch of mathematics, the construction of a postulate system makes but little demands on our mathematical knowledge. So many different postulate systems have been devised, frequently for the same doctrine, that it would be impracticable to mention the names of all their authors. But what is important is not the construction of one more postulate system; it is the fact first demonstrated by Peano and by Hilbert that postulate systems may be constructed.

We may summarize the major contributions of axiomatics in the following statements:

1. Axiomatics has been of inestimable value in mathematics, both from an analytical and from a constructive standpoint.
2. It has shown, more clearly than ever before, that mathematics consists in systems of relations and that the specific content is unimportant.
3. It has shown that logic, by itself, cannot establish the consistency of mathematics.
4. It has shown that, in a study of the nature of mathematics we cannot be satisfied with axiomatics alone but must go behind the postulates and discover their origin. In this way we are led to trace the historical development of a doctrine; and here psychological considerations cannot be ignored.

The foregoing conclusions appear to express the consensus of opinion of the leading authorities; no serious disagreements occur thus far.

Intuitionists and Formalists—In our discussion of the foundations, we have restricted our attention as much as possible to points on which only minor differences of opinion separate the various masters. But there are many other aspects of the subject on which no agreement has been arrived at or presumably ever will be. To elucidate the matter, an examination of the tendencies of the two opposing schools is necessary.

We mentioned that the appellation "intuitionist" does not refer to a mathematician who wishes to substitute hunches for rigorous demonstrations; but that, quite the reverse, an intuitionist is a "rigorist." His principal characteristic is that he rejects as invalid or insufficient some of the definitions and proofs which the formalist accepts. He claims that if a conclusion is established by means of what he regards as an unsatisfying proof, and that if this conclusion must nevertheless be accepted for mathematics to be possible, then the best course is to recognize that the conclusion cannot be established by logical means and admit that it is forced upon us by intuition. It is in this sense that the word "intuitionist" is used. The intuitionist accuses the formalist of circularity and of lack of rigor. The formalist would presumably retort by telling the intuitionist that he is narrow-minded, old-fashioned, and that the ideal of rigor to which he wishes to restrict mathematical proofs is gratuitous and certainly unattainable.

The net result of these divergences is that the formalists accept the proofs of the intuitionists, whereas, the latter reject many of the proofs of the former. For this reason the intuitionists are on the attacking side,

the formalists on the defensive. A few quotations will clarify these points. Poincaré discusses a proof proffered by Zermelo, one of the formalists. In this proof Zermelo sets out to show that the continuum (*e.g.,* all the points of space) can be transformed into a well-ordered set. The latter technical expression means that the points may be placed one after another in such a way that each point has a definite successor. In connection with this demonstration, Poincaré imagines a typical argument between a formalist and an intuitionist, whom he refers to under the equivalent names of "Cantorian" and "Pragmatist." He points out that a Cantorian will be won over by the rigor, whether real or apparent, of the demonstration. And then follows the discussion:

"*Pragmatist*—You say that you can transform the space continuum into a well-ordered aggregate. All right then, do so.
Cantorian—It would require too much time.
Pragmatist—Then show us at least how a man having the necessary time at his disposal and sufficient patience could effect the transformation.
Cantorian—This cannot be done because the number of operations to be effected is infinite; it is even greater than Aleph-zero.*
Pragmatist—Can you at least express in a finite number of words the rule you would have to follow to secure the ordering?
Cantorian—No, I cannot.
"And the Pragmatist will conclude that the theorem is meaningless, or incorrect, or in any event unproved."

The following passage, due to Weyl, is characteristic of the same rigoristic attitude:

"We must learn a new modesty. We have stormed the heavens but have succeeded only in building fog upon fog, a mist which will not support anybody who earnestly desires to stand upon it."

As may be imagined, it is more especially in problems involving the infinite (whether the infinitely great or the infinitely small) that the disputes arise. Cantor's theory of infinite aggregates was thus the starting point of the controversies. When Cantor published his epochal papers, many mathematicians refused to follow him. But, as Poincaré remarks, owing to the great beauty of Cantor's discoveries the majority of mathematicians were won over, even though they felt they were treading on dangerous ground. Then, in 1897, Burali Forti obtained a result which flatly contradicted one of Cantor's theorems. An inner inconsistency was thus revealed in the Cantorian theory, and a general cry of "I told you so" was raised by those who had ever felt but lukewarm towards the

* Aleph-zero represents infinity as exemplified in the infinite number of integers

new ideas. Since Burali Forti's initial discovery, a number of other antinomies, all of them of the same general pattern, were unearthed in rapid succession by Russell, Richard, König, and a host of other logicians. Today, this search for antinomies has about the same interest as searching for asteroids in astronomy. The important point is that such antinomies are inherent in Cantor's theory of aggregates.

Desperate attempts to save all that was possible of Cantor's theories were made by his supporters, the formalists. And it is in this field of mathematics that the controversies have been most acute: the formalists on one side, the intuitionists on the other.

Thus we have Poincaré writing: "I am here speaking of genuine mathematics, not of Cantorism."

Turning to Hilbert, we read:

"From the paradise created for us by Cantor, no one will drive us out."

With the passing of Poincaré, the formalists lost their most powerful opponent, and for a time discussion subsided. The attack was renewed at a later date by Brouwer and by Weyl, and the sphere of the controversy was extended to cover practically the whole field of mathematics. This may be seen in the following passage from Weyl's writings:

"The antinomies of the theory of aggregates are usually viewed as mere skirmishes, which relate only to the extreme frontiers of the mathematical empire, and which in no wise threaten the safety and the security of the empire itself.

"The explanations that are given for these disturbances in competent circles (with a view of contesting them or of weakening their significance) do not appear to be dictated by any solid conviction. Rather are they attempts at autosuggestion, not altogether sincere, such as are met with in politics and in philosophy. But in point of fact, an attentive and sincere examination of the matter can only lead to the conviction that these irregularities in the frontier regions of mathematics must be interpreted as symptoms. In this way we may bring to light the hidden evil, concealed by the seemingly perfect interplay of the mechanism in the central regions; this hidden evil is the inconsistency and lack of solidity of the foundations on which the whole empire is built. I know of only two attempts which go to the root of the evil. One is that of Brouwer. . . ."

The second attempt to which Weyl then refers is the one he makes himself in his book: "Das Kontinuum."

In view of these attacks of Brouwer and Weyl, Hilbert returned to the defense, temporarily interrupting his researches in the advanced parts.

a final effort he sought to establish the perfect consistency of mathe-

matics. But the intuitionists were neither convinced nor silenced, and the controversy rages on. According to Weyl, the concept of irrational number must be sacrificed or at least profoundly modified. Brouwer also contends that in dealing with infinite collections the principles of Aristotelian logic cannot be applied as they stand, and that the principle of Excluded Middle in particular must be discarded.

Assuming for the sake of argument that this view (which is not shared by the more moderate intuitionists) is valid, let us see where it will lead us. We recall that, in Hilbert's postulates, a distinction is made between the postulates themselves and the rules of logic, which are applied to them. The postulates may be varied according to the doctrine we wish to construct; but the laws of logic applied to all postulate systems are the same. If we follow Brouwer and decide that the law of excluded middle applies to finite, but not to infinite, collections, we are in effect treating this logical principle as a postulate. Many will feel that this course is the more satisfactory one to adopt, for it forces us to the conclusion that the laws of logic (or, at least, the principle of excluded middle) are not *a priori* forms of the understanding but arise from our commonplace experience with finite collections. Small wonder then that when we deal with infinite collections we fumble, since we have nothing in the world of our experience to guide us.

In favor of Brouwer's stand is the fact that results which are valid when we deal with finite quantities often cease to be true when infinite magnitudes and aggregates are considered. For instance, the properties of the sum of a finite number of terms cannot be extended uncritically to infinite series. Similarly the properties of finite numbers differ from those of infinite ones. In view of these examples and of many others which could be given, what right have we to assert that the passage to the infinite cannot possibly affect the validity of the commonly accepted rules of logic?

Against Brouwer's stand there is the objection that the antinomies do not only arise in connection with infinite aggregates. In this connection we have the well-known example of Epiminedes, who told his listeners that all Cretans were liars. Since, however, Epiminedes was himself a Cretan, the precise significance of his statement constituted a puzzle for the ancients: the logical principle of excluded middle seemed to be invalid. As another example of the same kind, Mr. Russell mentions the amusing case of a barber who shaves everybody in the town except those who shave themselves. And he then asks us whether the barber shaves himself. Whatever answer we give appears to be wrong. Here also the principle of excluded middle cannot be applied; and let us observe

that in neither of the two preceding examples are infinite aggregates involved. The intuitionists dismiss this objection to Brouwer's views by claiming with Poincaré that, for finite aggregates, the antinomies are mere toys. To quote Poincaré:

"The antinomies to which certain logicians have been led arose because they were unable to avoid circularity in some of their arguments. This has occurred when they considered finite collections, but it has happened more often when they professed to argue on infinite collections. In the former case they might easily have avoided the trap into which they fell; more exactly, they themselves set the trap into which it pleased them to fall, and indeed they were very careful not to fall outside the trap. In short, in this case the antinomies are mere toys. Very different are those that are generated by the concept of infinity; it often happens that here we fall without intending to, and even when we are forewarned, we cannot feel safe."

In our opinion, the most illuminating method of making headway in this labyrinth of assertions and counter-assertions is not to defend the thesis of any particular school, but to examine in an unbiased spirit typical demonstrations of the formalists, and to concentrate on the points where disagreements start. In this way at least we shall be in a position to understand clearly the points of disagreement. Poincaré followed this method in his last paper on the subject; and as far as we know he is the only writer who has done so.

In a chapter called "Les mathématiques et la logique" which appears in "Dernières Pensées," Poincaré gives a wealth of illustrations, one of which we have already referred to in connection with Zermelo's proof. Here is another. The formalist sees no objection in defining X in the following way:

(a) X (object to be defined) has such and such a relation with *all* the members of the genus G.

(b) X is a member of G.

Poincaré then notes that, according to the intuitionists, the foregoing definition of X is circular and therefore unacceptable. Poincaré proceeds to elucidate the objections of the intuitionists by pointing out that, in their opinion, definitions of the previous type can never be proved consistent, and that as a result it is impossible to say whether X has a mathematical existence or is a "round square." He notes that the antinomies usually issue from these kinds of definition.

Another controversy between the intuitionists and formalists arises from definitions which cannot be expressed by a finite number of words; these the intuitionists reject as unfit for mathematical use. A particular

example will clarify the nature of the point at issue. Consider the infinite series

$$1+\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\frac{1}{16}+\ldots\ldots$$

This series contains an infinite number of terms, which it would of course be impossible to write out since the terms are inexhaustible in number. Nevertheless the intuitionists accept a series of this sort as being susceptible of mathematical study. Their contention is that the law of formation of the series can be expressed by a finite number of words, *i.e.*, by the statement that each term has one-half the numerical value of the one immediately preceding it. But the intuitionists refuse to consider a series whose law of formation could be expressed only by an infinite number of words. Note that the distinction is *theoretical* and not *practical,* for the intuitionists would accept a series defined by means of, say, a quadrillion words; and from the practical standpoint, a definition containing a quadrillion words is just as impossible to state as one containing an infinity. But whereas a quadrillion words may be uttered *in theory,* the intuitionists claim that an infinite number cannot, for the simple reason that infinity cannot be exhausted. The result is that we can never complete the definition, even in theory, and for this reason we cannot determine whether a definition involving an infinite number of words has any meaning.

As may be gathered from these illustrations, the root of the disagreement lies beyond logic and mathematics in obscure psychological differences pertaining to the various kinds of minds. And so it would seem that a formalist will remain a formalist from the cradle to the grave, and that the same holds true for an intuitionist. There is not much hope of the one ever convincing the other, so that nothing is to be gained by perpetuating endless controversies. Yet in a sense these controversies have not been entirely futile, for they have convinced mathematicians that agreement is impossible—and this in itself is a form of agreement.

Poincaré ends his last paper on these fruitless controversies with a note of resignation. He writes:

"At all times there have been opposing tendencies in philosophy, and it does not seem that these differences are nearing any settlement. The reason is presumably that men have different minds, and that these minds cannot be changed. There is therefore no hope of expecting any agreement between the Pragmatists and the Cantorians. Men do not agree because they do not speak the same language and because some language will never be learned.

"And yet they usually agree on mathematical questions; but this is precisely thanks to what I have called verifications. These latter are the judges of last appeal, and to their verdicts everyone submits. But where these verifications are lacking, mathematicians are as helpless as mere philosophers. When it is a matter of deciding whether a theorem which has no means of being verified may nevertheless have a meaning, who can act as judge, since a verification is impossible? The sole recourse would be to force one's adversary into a contradiction. But this has been tried without success.

"Many antinomies have been recorded, and yet disagreement has persisted; no one has been convinced. The fact is that we can always evade a contradiction by introducing a reservation: I mean a *distinguo*."

From the standpoint of the theoretical physicist, the study of the nature of mathematics would have been of interest had it led to definite conclusions; for the whole structure of theoretical physics is based on a mathematical scheme. And unless mathematics be reliable, what hope can we have of applying it with success to the physical world? In some respects this difficulty is mitigated in current theoretical physics because mathematics is here applied only for the establishment of probabilities and pseudo-accuracy is less objectionable in the calculation of probabilities than it was in the days when rigorous laws were thought to exist. It might even be said that modern physics is witnessing the same crisis that we have been discussing in mathematics: the quantum theorists occupy the position of the intuitionists while Einstein and Planck occupy that of the formalists.

Thus the quantum theorists reject rigorous determinism for microscopic processes, though they grant its validity on the more commonplace levels of ordinary physical experience. Note the resemblance of this attitude to that of Brouwer, who rejects the law of excluded middle for infinite aggregates, while granting its validity for the finite aggregates with which our ordinary experience makes us familiar. Also recall the insistence of the quantum theorists on accepting only those notions and results which, in theory at least, are susceptible of being observed. Is not this attitude very similar to that of the intuitionists, who insist on definitions that can be expressed by means of a finite number of words? The following quotation from Heisenberg, exhibiting his philosophy in physics, is very similar to one we might credit to an intuitionist in mathematics. He writes:

"In this connection, one should particularly remember that human language permits the construction of sentences which do not obey any consequences and which therefore have no content at all—in spite of the fact that these sentences produce some kind of picture in our imagination;