Volume 2 *July, 1930* *Number 3*

REVIEWS

OF

MODERN PHYSICS

RELATIVITY AND ELECTRODYNAMICS

BY W. F. G. SWANN

BARTOL RESEARCH FOUNDATION

IN AN article of this length, it would be useless to attempt to deal in detail with the subjects covered by the title, and an apology may seem necessary for writing, in these days, a review of such time-worn fields. There are many well versed in electrodynamics and relativity, and experienced in the technique of their use, to whom this article will appear as no more than the most elementary and fragmentary of treatments. To these, its interest will be confined to a few variations in point of view here and there, and possibly as a reminder that more fundamental elements of philosophic interest lie behind some of the uses of these sciences than are sometimes realized by those who are able to appease their philosophic conscience in contemplation of the beauty of mathematical analysis. On the other hand, there are those who have not specialized in these fields, to whom electrodynamics has its clearest visualization in engineering problems, to whom there remain curious thoughts concerned with whether or not lines of force move, those to whom the application of electrodynamics to things like electrons assumes a more and more shadowy meaning until nothing but the ghost of a meaning remains in the subject as applied to such fields as that of wave mechanics. There will be those to whom relativity is a subject having to do with an attempt to correlate the Michelson and Morley experiment with common sense, and which by some mysterious means has flown on the wings of mathematical analysis to all domains of physical theory, nonchalantly hiding the story of how it got there, behind an imposing array of symbolism which they are challenged and required to dicipher before the answer to the story will be vouchsafed. It is mainly to all such as these that the present article is addressed; and, it is hoped that, while its perusal may still leave them with many problems of mathematical difficulty, it may remove from the horizon some matters of perplexity.

In any presentation of a branch of modern physics, two courses are open. The first is the historical. This has the disadvantage that, usually, it does not represent a sequence of logical developments. The ways in which conclusions are reached are founded frequently upon considerations of special cases, and sometimes are based upon experiments whose representatives have completely evaporated in the more general fields in which the conclusions are subsequently used. The alternative method is to take the results which have been stumbled upon in the historical development, review the path by which they have been reached, remove from it as much as possible of the obsolete debris with which it is encumbered, and try to construct some more satisfactory path by which the results might have been reached even to the extent of possible modifications in the starting points. This method runs the danger of appearing artificial. Many of the proofs which seemed fundamental to the old structure now play no part at all, and the whole vision may be apt to appear as a constellation in the sky instead of something standing on the rocks of the firm earth. However, an honest and unbiased facing of the situation in such a case will usually reveal that, in the last analysis, the older and apparently substantial concepts will not bear transportation to the realms in which the subject claims some of its most powerful uses. Those things which seem to endow the subject with such elements of tangibility when considered on an engineering scale dissolve in meaning, leaving only the fundamental content which, as a matter of fact, was the thing really of moment in the engineering problems where, however, it remained hidden in the glittering robes of a spurious reality. And so, to the critically minded, there appears a reality in the new artificiality and an artificiality in the old reality. The process of generalization is often not one in which the new forms a complication, or extension of the old, but rather one in which the old is a rather vague, somewhat incomplete, and illogical application of the principles of the new.

In what follows, it will be our purpose to dissect as far as possible the real content of the formulae, laws and concepts dealt with. When, as the result of a merciless stripping of irrelevant adornment from the laws, we have a spectacle impalatable to the intellectual taste, let the reader ask himself whether this cold remnant does not contain the whole essence of the laws in the sense that they are actually used. Our thoughts are accompanied by many visions which play no part in their essential structure. While we would strenuously deny that our picture of an all-pervading ether was like that of water, many of us supplement the cold statement of properties of medium with vague shadows of substantiality concerning which only the most dire intellectual torments would bring us to confess, even to ourselves, that we were semi-consciously thinking of the taste of the ether, of its boiling point and freezing point, of its latent heat of evaporation and so forth.

The Theory of Relativity

The so-called restricted theory of relativity arose, as is well known, in connection with the interpretation of experiments designed to detect what, in the language of thought of the day, was the earth's absolute motion in space. We may conveniently dispense here with the history of the matter, pass over one or two stages in the journey of emancipation from these rather troublesome experiments regarded as a starting point, and state that, in the hands of Einstein the content of the theory may be summed up in two statements as follows.

(1) If two observers A and B, moving relatively to each other with a velocity v, assign coordinates and time to some event recognizable by each of them, the values which they assign will not be the same, but will be such that if each of them uses his own measures to determine the velocity of the same set of particles shot out from a point, each will find for that velocity the same value, irrespective of the direction, provided that that value is a certain constant c associated with the velocity of light.

(2) The same relations between the above measurements of the two observers which serves to conceal from them any effect of their motion in the matter of their measurements of the speed of light, will also serve to conceal such effects as regards any tests in which they perform corresponding experiments in their respective systems.

These statements as they stand, are fraught with danger as to misinterpretation. No. 2 particularly is open to all sorts of misrepresentations, so it will be well to start by dissecting the statements in detail. Before doing this, however, I cannot resist making the remark that as one absorbs the philosophy of the theory of relativity, he soon comes to a point where he wonders what place these imaginary observers have in the scheme of things, and why the theory of relativity should have to do apparently so fundamentally with the velocity of light. In his suspicion of the fundamentality of these intermediaries he is justified. To dispense with them at this state, however, would be to depart too far from the historical development of the subject to give the mind of the reader the opportunity of attuning itself sufficiently closely to the habits of thought of those who have grown up with this development. He might fail to understand much that has been written on the subject or indeed to see any reason for its existence. For these reasons we shall defer the discussion of the more fundamental approach to the restricted theory of relativity, and shall for the present, follow the usual practise of talking of clocks and scales and the like.

Significance of Statement No. 1. Let us call our observers A and B, and let us suppose that each is supplied with scales by which he may measure distances, and clocks by which he may measure time; and suppose that the latter are distributed throughout all space, so that at any place at which the time is required for any event taking place, there is a clock conveniently by to record it. These clocks are simply to be regarded as symbolic; all we wish to suppose, in fact, is that there is some process by which each observer is able to arrive at what he calls the time of a recognizable event at any place. It may be partly by measurement and partly by calculation that he secures this end, as when an astronomer assigns the time of an eclipse allowing for the velocity of light in transit. We are not for the moment interested in any question of "correctness" of the measures. All we imply is that each observer has *some method* of assigning his times, and this fact is symbolized by his clocks. If B is moving relatively to A with velocity v, we are of course to suppose that as viewed by A, B's clocks and scales will be found travelling through space with the velocity v.

Suppose now that from the origin of A's system of coordinates light particles are emitted in all directions with a velocity which A measures as c. How does A measure this velocity? We may suppose that he is provided with

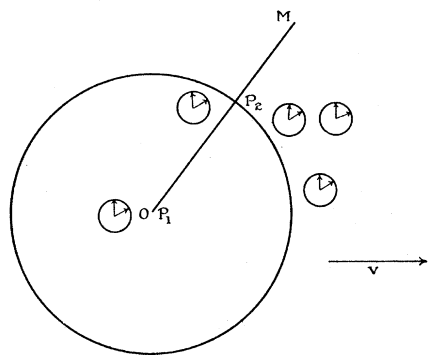

Fig. 1.

a scale OM, one of whose ends, of graduation P_1, is at his origin. He records the time t_1, as indicated by his clock at the origin when the particles are emitted. When a light particle passes some point reading P_2 on the scale the time recorded by one of his clocks located there is t_2. The amount $P_2 - P_1$

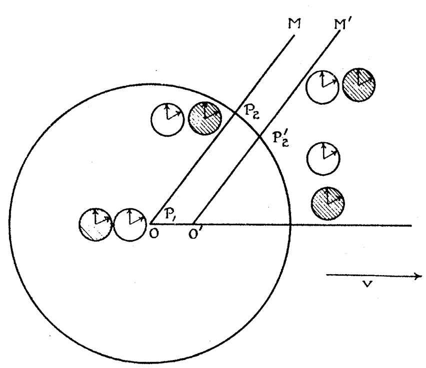

Fig. 2.

is then the value which he finds for the distance travelled by the light in the time $t_2 - t_1$ and we are supposing that the ratio is c *i.e.*

$$\frac{P_2 - P_1}{t_2 - t_1} = c.$$

Now let us superpose on Fig. 1, the apparatus owned by B and so obtain Fig. 2, where $O'M'$ is B's scale and the shaded clocks are B's clocks. Let us suppose that at the instant when the light signal was emitted from O the scales coincide, so that in the present picture, (Fig. 2) which is supposed to correspond to the instant t_2 as measured by A, the scales will be shown in the positions indicated.

Let us start by supposing that the clocks and scales of A and B read alike in the sense that if any clock in B be found coinciding with a clock of A it will be found to have the same reading, and if any scale of B be found by B coinciding with a scale of A, it will be found also to coincide division for division with that of A if one pair of divisions coincide. Now the time recorded by B for the emission of the light particles is t_1, the same as that recorded by A since the scales coincided initially, and the initial reading on B's scale will be P_1. When the light particles are on the circle shown in Fig. 2, A will, as we have stated, record P_2 for the position of one of them on his scale, and t_2 for the corresponding time. B's scale will have moved to the position shown in Fig. 2, but the time by A's clock when the light particle reaches the point we have designated as P_2' will obviously be t_2: and if we assume B's clocks to read the same as those of A, the time recorded by B for this coincidence of the light particle with the mark P_2' will also be t_2. Thus the difference of the scale readings recorded by B for the light particles journey in the time t_2-t_1 is $P_2'-P_1$ and the velocity which he will calculate for the light particle is

$$c' = \frac{P_2' - P_1}{t_2 - t_1} \cdot$$

Since we are supposing that one of A's scales if superposed on one of B's would coincide with it mark for mark, it is obvious that P_2' will be less than P_2, and hence the velocity c' measured by B will be different from the velocity c measured by A. If, however, B actually calculated from his measurements the same velocity c as is found by A, then either his scales or clocks or both record differently from those of A. Now Einstein set himself the problem of ascertaining how the readings of the scales and clocks of B must be related to those of A in order that both should find the same value c for the velocity of light. The answer to this question was as follows: Suppose one observer A adopts a rectangular system of coordinates designated x, y, z, and suppose t is the time which he records for a certain event at x, y, z. Then the corresponding values x', y', z', t' which the observer B will record for the event are such that

$$x' = \beta(x - vt) \tag{1}$$

$$y' = y \tag{2}$$

$$z' = z \tag{3}$$

$$t' = \beta\left(t - \frac{v}{c^2}x\right) \tag{4}$$

where

$$\beta^2 = \frac{1}{1 - v^2/c^2}$$

and where it is supposed that, in the frame of reference of A, B is moving along the axis of x with velocity v.

No attempt is made to *explain* why there should be any such difference in the measures. Thus, older, more intuitive attempts to correlate the experimental facts, attempts which preceded the theory of relativity, sought some physical reason as to why the scales and clocks of an observer should change in their readings when their velocity was changed. For the moment at any rate, we shall disregard all thoughts of this kind, with their not too immediately obvious pitfalls in logical thinking, and we shall concentrate our attention on the state of affairs depicted by (1)–(4) as our starting point.

The reader may complain that we have said nothing about the way in which B adjusts his scales and clocks to warrant the belief in any such conclusion as was reached in (1)–(4). That is true. It is customary to start by making a definition of simultaneity founded on the idea that if in either system an observer sends a light signal from a point "a" at t_1, reflects it at "b" back to "a" where it arrives at t_2, the clock at "b" will be running synchronously with that at "a" if the time t_3 which it records for the arrival of the light signal is $t_3=\frac{1}{2}(t_1+t_2)$. This provides a logical method for setting the clocks, a method which is the same for each observer, and one which can easily be seen to be self-consistent and to perpetuate its consistency if the relations (1)–(4) are true. Moreover, we can leave to the reader such trivialities as to how the two observers decide that one of them is not measuring in meters and the other in yards. I have wished to suppress this discussion about the setting of the clocks, and the like, however, because it is just one of those elements which has really nothing particularly to do with the vital elements of the theory of relativity. It is something necessitated to provide a comfortable existence for these imaginary observers who do not exist. Perhaps it will be well to refrain from confusing matters too much by discussing this point at present, so we shall leave it now, with a promise to return to it later when the time is more ripe.

Significance of Statement No. 2. Now a primitive view of the way in which nature acts might suggest that if we measured the resistance of a wire when its length was parallel to what we suppose to be the line of the earth's motion and then turned it through 90°, the resistance might be changed, whereas there would be no change if the earth were at rest as we understood that term. Experiments of this and various analogous kinds have been tried and all failed to give any indication of a motion for the earth in the sense sought. In all such experiments, the experimenter sets up certain apparatus—he assigns coordinates x, y, z, to certain things. Possibly he may also produce certain velocities or accelerations of certain parts, but as a result of all his manipulations he finally observes something—the motion of a galvanometer spot for example. If we are thinking of two observers (possibly with different measures of space, x, y, z, and time t), and if we are supposing them to be performing identical experiments, we are to understand by that, that each does the same thing in his own system of coordinates. If one places a source of light at the point $x=2$, $y=4$, $z=6$, the other places a similar source of light at $x'=2$, $y'=4$, $z'=6$. If one arranges for the velocity of a particle to be $dx/dt=8$, $dy/dt=0$, $dz/dt=3$, the other arranges to have $dx'/dt'=8$, $dy'/dt'=0$, $dz'/dt'=3$. With this understanding as to the similarity of two experiments, two observers having relative motion will, on account of the Doppler effect observe different frequencies for a given spectrum line of a star. This fact constitutes no contradiction to the above statement, however; for here, the velocity of the star relative to each of the two observers is not the same. If each observer had his own star and if A's star moved relatively to him with the same velocity and direction as the other star moved relatively to B, the nature of the light observed by the two observers would be the same. Thus, one of the fundamental postulates of the restricted theory of relativity, in its historical aspect, at any rate, is that two observers moving relatively to each other with uniform velocity v obtain the same results in similar experiments. In such experiments as have been done, the hypothetical observers are at most really the same observer experimenting at two different times of the year when the direction of the earth's motion is not the same, and in the case of most of the experiments even this aspect is a generalization of the conclusions drawn for experiments performed at one time of the year,[1] but with the apparatus oriented first in one direction and then in a direction at 90° to the first direction.

It is not my intention to discuss here the comprehensiveness of the experimental evidence for the postulate, but simply to trace the consequence of the postulate. Accepting the postulate as true it implies that the laws of nature—the laws governing all phenomena—must be of the same form expressed in terms of x, y, z, t of equations (1)–(4) as in terms of x', y', z', t'. For, if this were not the case, it would, in general, be possible for our two hypothetical observers to obtain different results in similar experiments, similar in the sense above described.

It may perhaps be well to restate, in other words, the content of what has been written above. The laws of physics are expressed in terms of differential equations. Even an equation without any differential coefficients in it is a limiting case of a differential equation. It is a differential equation of zero order. Now when we proceed to seek the solution of any problem it is necessary not only to solve the differential equation, but also to specify the concrete values of certain things at $t=0$, for example. Thus, the Newtonian equations of dynamics are equations of the second order, and we must assign two things initially for each coordinate, the coordinate and the velocity (or momentum). If the equations had been of the third order, it would have been necessary to assign three things for each coordinate at $t=0$, the coordinate itself, the velocity, the acceleration; and so on. In general, we may say that the act of two observers doing the same experiment in two different systems is to be symbolized by their assigning the same numbers to the quantities which

[1] D. C. Miller, for example, has called attention to the fact that the original Michelson-Morley experiment is often quoted as having been performed at various times during the year, whereas it was actually performed over a very limited period.

the differential equations requires to have assigned in order to complete their solution. In fact, the act of doing the same experiment in the two systems is symbolized by the assignment of the same initial conditions in each.

It is manifest that if the differential equations for the two observers are the same, they will always get the same solutions for identical initial conditions. Or, in other words, they will always get the same *results* when they do the *same experiments*, understood in the above sense. Moreover it is not difficult to see, that, unless the differential equations are the same in the two systems,[2] the observers will not in general obtain the same results for the same initial conditions.

Let us enlarge upon the foregoing statement, and by way of illustration, suppose that the Newtonian law of gravitation were under test as to its fitness to conform to the requirements we have stated. It will suffice to consider the law as applied to two bodies a and b, in which case, it takes the form

$$m_a\frac{d^2x_a}{dt^2} = \frac{m_a m_b G(x_a - x_b)}{\left\{(x_a - x_b)^2 + (y_a - y_b)^2 + (z_a - z_b)^2\right\}^{3/2}} \tag{5}$$

$$m_a\frac{d^2y_a}{dt^2} = \frac{m_a m_b G(y_a - y_b)}{\left\{(x_a - x_b)^2 + (y_a - y_b)^2 + (z_a - z_b)^2\right\}^{3/2}} \tag{6}$$

$$m_a\frac{d^2z_a}{dt^2} = \frac{m_a m_b G(z_a - z_b)}{\left\{(x_a - x_b)^2 + (y_a - y_b)^2 + (z_a - z_b)^2\right\}^{3/2}}\,. \tag{7}$$

Now by means of the relations (1)–(4) it would be possible to express $x_a, x_b, y_a, y_b \cdots$ etc. in terms of $x_a', x_b', \cdots$, etc, and t', and to express d^2x_a/dt^2 etc. in terms of derivatives of the dashed spacial coordinates with respect to the dashed time coordinate. When all this was done, so that the transformed (5)–(7) had no longer any undashed letters in them, it might happen that the three equations contained as their consequence[3]

$$m_a\frac{d^2x_a'}{dt'^2} = \frac{m_a m_b G(x_a' - x_b')}{\left\{(x_a' - x_b')^2 + (y_a' - y_b')^2 + (z_a' - z_b')^2\right\}^{3/2}} \tag{8}$$

$$m_a\frac{d^2y_a'}{dt'^2} = \frac{m_a m_b G(y_a' - y_b')}{\left\{(x_a' - x_b')^2 + (y_a' - y_b')^2 + (z_a' - z_b')^2\right\}^{3/2}} \tag{9}$$

$$m_a\frac{d^2z_a'}{dt'^2} = \frac{m_a m_b G(z_a' - z_b')}{\left\{(x_a' - x_b')^2 + (y_a' - y_b')^2 + (z_a' - z_b')^2\right\}^{3/2}} \tag{10}$$

[2] If we could imagine a situation where the differential equations were not even of the same order, the very meaning of "doing the same experiment in two systems" would start to evaporate. Thus, for example, if the equations in B's system were of the third order, while those in A's were of the second, we might *define* the doing in B of an experiment similar to that done in A as the assignment of the same initial quantities in B as in A; but, under these conditions there would be an ambiguity in the results obtained in B, depending upon the unspecified values of the additional initial condition necessitated by the higher order of the equations in B.

[3] We do not imply that (5) might be expected to go into (8), (6) into (9), or (7) into (10); but simply that these three equations which we get by the transformation might be shown to contain (8)–(10) as their combined consequence.

Now such would not be the case in the instance cited, but if it had been,[3] the equations would have remained invariant under the transformation given by (1)–(4), and they would be consistent with the restricted theory of relativity. If our two observers A and B each in his own coordinates observed not the same pair of planets, but different pairs; and if it so happened that, for example, the quantities

$$\frac{dx_a}{dt},\ \frac{dy_a}{dt},\ \frac{dz_a}{dt},\ \frac{dx_b}{dt},\ \frac{dy_b}{dt}\cdot\ \frac{dz_b}{dt}$$

were found to have the values 3, 6, 10, 4, 8, 2, respectively; and if x_a, y_a, z_a, x_b, y_b, z_b, had the values 2, 9, 7, 4, 2, 8, at $t=10$, and if B observed that for his planets

$$\frac{dx_a'}{dt'},\ \frac{dy_a'}{dt'},\ \frac{dz_a'}{dt'},\ \frac{dx_b'}{dt'},\ \frac{dy_b'}{dt'},\ \frac{dz_b'}{dt'},\ x_a',\ y_a',\ z_a',\ x_b',\ y_b',\ z_b',$$

had, respectively, the values 3, 6, 10, 4, 8, 2, 2, 9, 7, 4, 2, 8, at $t'=10$, then the equations (8)–(10) being of the same form as (5)–(7) would insure that the subsequent history of B's two planets would go on in the same manner in the dashed system as A's would go in in the undashed system.

Of course, it is not to be supposed that if A and B observe the *same* pair of planets they will see motions of the same form. If one of these planets is going around the other in a curve which to A appears a circle, it will not appear a circle to B. It has no right to. It is obeying the same general law for A and B, the law as expressed by (5)–(7) or (8)–(10), but the initial conditions—the assigned positions and velocities at $t=10$ (using the symbolism of our example above) in the undashed system of coordinates, are not the same as the assigned values at $t'=10$ in the dashed system of coordinates, for the same planet.

There is perhaps a little more importance to be attached to the foregoing statement than might be apparent at first sight; for, there is here involved the rather delicate question of what we mean by a law of nature. Thus it might so happen that our observer A found that each of his planets described a circle around the other. Yet if he stated his law in this form, completing his specification as regards the size of the circle, etc., he would not make a statement which was invariant under the transformation (1)–(4), even though equations (5)–(7) had reverted to the type (8)–(10) under the transformation. It might be perfectly true that, to A, the planets did describe the circles to which we have referred. You will say, of course, we had no right to expect the forms of the orbits to be invariant, it is the differential equations (5)–(7) which are the things whose invariance is under examination, and these are the things which express the laws of nature. This is quite true, but it is as well that we should formulate for our own clarity the statement of precisely what we are to mean by the laws of nature in a discussion of this kind. Thus if we should take the circular orbits which we supposed viewed by our observer A, and differentiate the coordinates once with respect to the time we should arrive at differential equations—and why are these not capable of being regarded as an expression of the law of nature? The answer to the question is that there would be left in the equations so formed constants which were peculiar to the problem we were considering. In the totality of all the problems to which our equations would be supposed to apply we should have to recognize the necessity of assigning different values to these constants for each problem. The equations (5)–(7) have no constants of this type. For the totality of all of the problems to which we consider them as applying the constants G and m_a are the same. When our equations have been so far robbed of special features (usually by reducing them to a sufficiently high order) as to have in them no constants which are to be supposed different for different problems, they may be said to stand as expressions of *laws* in the same sense implied when the word "law" is used in the theory of relativity.

The invariance we have discussed so far in speaking of equations (5)–(7) is that pertaining to equations in which there occur, in addition to constants, no quantities other than the coordinates and times pertaining to certain identifiable points. We have occasion, however, to speak of invariance in the case of such equations as the electromagnetic equations, where quantities such as the electric and magnetic fields $\boldsymbol{E}$ and $\boldsymbol{H}$ occur. What has been written above as to the significance of corresponding experiments in the two systems still applies in the sense that to perform like experiments in the two systems we must assign in those systems the same numbers to the corresponding quantities whose numerical specification is necessary to provide a solution of the problem with the aid of the differential equations. There remains, however, something to be said as to the significance of the quantities concerned in relation to some fundamental set of definitions, when the quantities in question are electric and magnetic fields, potential differences and the like. Thus, when I say "Let a planet occupy the position $x=3$, $y=6$, $z=7$, at $t=2$ and let its velocities be $\dot{x}=5$, $\dot{y}=9$, $\dot{z}=11$," I say something immediately interpretable in terms of the coordinates whose transformation is expressed by (1)–(4). When, however, I say, "Consider an electron in an electromagnetic field, $E_x=5$, $E_y=7$, $E_z=10$, $H_x=5$, $H_y=2$, $H_z=12$," it is, in general, necessary to know the definitions of these quantities in terms of something ultimately related to a moving point (as when we define the fields in terms of the motion of an electron) before I can attach a meaning to the transformation.

Thus, to anticipate for a moment the section on electrodynamics, it will be recalled that the electromagnetic equations may be written

$$\frac{1}{c}\left(\rho u+\frac{\partial E}{\partial t}\right)=\text{curl } \boldsymbol{H} \tag{11}$$

$$\rho=\text{div } \boldsymbol{E} \tag{12}$$

$$-\frac{1}{c}\frac{\partial \boldsymbol{H}}{\partial t}=\text{curl } \boldsymbol{E} \tag{13}$$

$$0 = \operatorname{div} \boldsymbol{H} \tag{14}$$

where ρ is the electric density, $\boldsymbol{u}$ its velocity, $\boldsymbol{E}$ and $\boldsymbol{H}$ the electric and magnetic fields, and c the same constant as occurs in (1)–(4).

Now it is an algebraical fact that if (11)–(14) be subjected to the transformation (1)–(4), they assume the form:

$$\frac{1}{c}\left(\rho'\boldsymbol{u}' + \frac{\partial \boldsymbol{E}'}{\partial t'}\right) = \operatorname{curl}' \boldsymbol{H}' \tag{15}$$

$$\rho' = \operatorname{div}' \boldsymbol{E}' \tag{16}$$

$$-\frac{1}{c}\frac{\partial \boldsymbol{H}'}{\partial t'} = \operatorname{curl}' \boldsymbol{E}' \tag{17}$$

$$0 = \operatorname{div}' \boldsymbol{H}' \tag{18}$$

where by the dashed symbols we understand as follows:

$$(\boldsymbol{u}_x', \boldsymbol{u}_y', \boldsymbol{u}_z') = \left(\frac{\boldsymbol{u}_x - v}{(1 - \boldsymbol{u}_x v/c^2)}, \frac{\boldsymbol{u}_y}{\beta(1 - \boldsymbol{u}_x v/c^2)}, \frac{\boldsymbol{u}_z}{\beta(1 - \boldsymbol{u}_x v/c^2)}\right) \tag{19}$$

$$\boldsymbol{E}_x' = \boldsymbol{E}_x, \quad \boldsymbol{E}_y' = \beta\left(\boldsymbol{E}_y - \frac{v\boldsymbol{H}_z}{c}\right), \quad \boldsymbol{E}_z' = \beta\left(\boldsymbol{E}_z + \frac{v\boldsymbol{H}_x}{c}\right) \tag{20}$$

$$\boldsymbol{H}_x' = \boldsymbol{H}_x, \quad \boldsymbol{H}_y' = \beta\left(\boldsymbol{H}_y + \frac{v}{c}\boldsymbol{E}_z\right), \quad \boldsymbol{H}_z' = \beta\left(\boldsymbol{H}_z - \frac{v}{c}\boldsymbol{E}_x\right) \tag{21}$$

$$\rho' = \beta\rho(1 - \boldsymbol{u}_x v/c^2). \tag{22}$$

Before we can say that the equations (11)–(14) are "invariant" under the transformation, however, it is necessary to insure that $\boldsymbol{u}'$, ρ', $\boldsymbol{E}'$, $\boldsymbol{H}'$, are the quantities which the observer in the dashed system really would recognize as velocity, density, electric field and magnetic field respectively. As regards $\boldsymbol{u}'$, the desired conclusion follows very easily, since (1)–(4) enables us to compare velocities in the two systems. In order to say anything about $\boldsymbol{E}'$ and $\boldsymbol{H}'$ it is necessary to appeal to the definitions of these quantities. We shall deal with this matter more fully later; and, it will suffice now to say as follows:

If $\boldsymbol{E}$ and $\boldsymbol{H}$ are defined in terms of the motion of some identifiable entity such as an electron, it is possible to write down the meaning of $\boldsymbol{E}'_y$, for example, in terms of the motion of something in the dashed system. We can then transform this to the undashed system so that $\boldsymbol{E}_y'$ in (20) becomes expressed in terms of velocities, accelerations, etc. of an electron in the undashed system. The relation now assumes the form.

Something about the motion of an electron in terms of the undashed coordinates $= \beta\left(\boldsymbol{E}_y - \dfrac{v\boldsymbol{H}z}{c}\right).$

But this is a statement about the motion of the electron in relation to the fields $\boldsymbol{E}$ and $\boldsymbol{H}$. If it is true, it provides for $\boldsymbol{E}_y'$ being the quantity which the dashed observer recognizes in terms of his own definitions, as the y component of the electric field. If it is not true the significance of E_y' as an electric field and with it the meaning of the invariance, falls to the ground. Similar remarks apply, of course, to the other quantities E_x', E_z', H_x', etc.

The Mathematical Procedure for Securing Invariance. Now as we have already remarked, equations (5)–(7) would not turn out to be invariant under the transformation (1)–(4); and, if, without examining the vital elements controlling the situation, anyone should try to modify these to a form which was invariant (without departing appreciably from the numerical story which they tell, since that story is known to be approximately correct), he would probably invite a confusion of symbols sufficient to occupy his efforts for many a long day, and he would conclude that any practical efforts along this line would be hopeless. In order that we may present the possibilities in a more favorable light, we shall make a temporary digression from the theory of relativity, and consider the situation which we really have in prerelativity physics, where something of the same kind of difficulty is really existent, but has become smoothed over in our thinking because the way has been carefully prepared for us so that some of us may never have realized the difficulties which we might have encountered had things been a little different from what they are.

Let me recall the well known-equations

$$4\pi i_x = \frac{\partial H_z}{\partial y} - \frac{\partial H_y}{\partial z} \tag{23}$$

$$4\pi i_y = \frac{\partial H_x}{\partial z} - \frac{\partial H_z}{\partial x} \tag{24}$$

$$4\pi i_z = \frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} \tag{25}$$

which relate the current density i to the magnetic field H in the case of a system of steady currents. Suppose that we should wish to transform these equations to a system of axes which were inclined to the axes in terms of which (23)–(25) are expressed. If $i_x', i_y', i_z', H_x', H_y', H_z'$, represent the corresponding components of our quantities referred to the new axes, it is obviously possible to express i_x, i_y, i_z, $\partial H_z/\partial y$, $\partial H_y/\partial z$, $\cdots$ etc. in terms of quantities with dashed letters. Thus for example, if l_1, m_1, n_1, are the directional cosines of the new x axis with respect to the old axes, l_2, m_2, n_2 and l_3, m_3, n_3 the directional cosines of the new y and z axes, respectively, with respect to the old,

$$i_x = i_x'l_1 + i_y'l_2 + i_z'l_3 \quad \text{etc.}$$

When we had completed our transformation, we should be greatly surprised if it should fail to result that the equations which we finally obtained contained as their consequence

$$4\pi i_x' = \frac{\partial H_z'}{\partial y'} - \frac{\partial H_y'}{\partial z'} \tag{26}$$

$$4\pi i_y' = \frac{\partial H_x'}{\partial z'} - \frac{\partial H_z'}{\partial x'} \tag{27}$$

$$4\pi i_z' = \frac{\partial H_y'}{\partial x'} - \frac{\partial H_x}{\partial y'}. \tag{28}$$

For if this consequence did not result we should say that the laws of nature depend upon the directions in which we choose our axes in space. Such a situation is not on the face of it necessarily absurd, for it might have been a fact that the universe had been constructed in a sort of huge crystal, with different properties in different directions. In fact, in the last analysis, such a state of affairs would constitute exactly what we should imply by the general statement that the universe had been constructed on a crystalline basis of this kind. The universe might have been so constructed, and, in limiting ourselves to laws which *do* retain their form unchanged when the axes are rotated, we implicitly postulate that it has *not* been so constructed.

But why is it that the law (23)–(25) *does* remain unchanged under the transformation? It is by no means an obvious conclusion. Anyone unfamiliar with the process, who tries the transformation for the first time would probably get into a hopeless muddle. And, to show how near things came to being different in this respect from what they are, it will suffice to remark that, had the minus signs on the right hand side of equations (23)–(25) been plus signs, the invariance of form under transformation would not have followed. Yet, the equations would have appeared in every way as natural and symmetrical as before. Why is it that the equations we happen to have chosen in physics do obey this criterion of invariance? It is because the quantities which we have equated are components of what we call vectors. We know that if we want to transform the coordinates of a point x, y, z to an inclined set of axes in which the directional cosines of the old x, y, z axes with respect to the new axes, are, respectively, l_x, m_x, n_x, l_y, m_y, n_y, l_z, m_z, n_z, the equations of transformation are:

$$x' = xl_x + yl_y + zl_z \tag{29}$$

$$y' = xm_x + ym_y + zm_z \tag{30}$$

$$z' = xn_x + yn_y + zn_z \tag{31}$$

Suppose we have three numbers P, Q, R, which are associated with the three axes, x, y, z, and three numbers P', Q', R', which are associated with x', y', z', in the same manner as P, Q, R, are associated with x, y, z. Then if P', Q', R', are related to P, Q, R, by the transformation

$$P' = Pl_x + Ql_y + R_z \tag{32}$$

$$Q' = Pm_x + Qm_y + Rm_z \tag{33}$$

$$R' = Pn_x + Qn_y + Rn_z \tag{34}$$

the three quantities P, Q, R, may be said to correspond to the components of a vector. This definition is as we shall see, far more fundamental than the statement that a vector is a directed quantity, which, of itself means nothing unless one is able to say what he means by a directed quantity. If one speaks of a velocity he may think there is no ambiguity in his meaning as to its direction. But suppose he speaks of an angular momentum. We know that an angular momentum may be regarded as a vector, its direction being perpendicular to the plane of rotation. Now this is the one direction in which there isn't any motion at all. What then gives it the characteristic of being the line to represent the vector?

The answer is as follows: let us consider the case of a particle of unit mass moving in a closed orbit in a plane, and consider three quantities $(v_y z - v_z y)$, $(v_z x - v_x z)$, $(v_x y - v_y x)$, which we ordinarily regard as the components of an angular momentum, but let us for the moment regard them simply as algebraical expressions. The notation is sufficiently obvious to render explanation unnecessary. Calling the above three quantities A, B, C, for shortness, it can readily be shown that they are respectively proportional to l, m, n, the directional cosines of a line which is perpendicular to $\boldsymbol{v}$ and to the radius vector $\boldsymbol{r}$. Moreover, it is easy to see that the factor of proportionality is $(A^2+B^2+C^2)^{\frac{1}{2}} = |\boldsymbol{v}|\,|\boldsymbol{r}| \sin \Theta$ where $|\boldsymbol{v}|$ is the resulting velocity, $|\boldsymbol{r}|$ is the distance of the unit mass from the origin and Θ is the angle between $\boldsymbol{v}$ and the radius vector $\boldsymbol{r}$. In other words $(A^2+B^2+C^2)^{\frac{1}{2}}$ is what we call the resultant angular momentum, M.

Hence, since quantities like A, B, C, are always equal to M times the directional cosines of a fixed line whatever the set of axes chosen, A, B, C transforms according to (29)–(31) which are the equations of transformation of a line. Hence A, B, C, form the components of a vector and the direction of that vector is that determined by l, m, n. In other words, it is perpendicular to $\boldsymbol{v}$ and to the radius vector $\boldsymbol{r}$. Moreover, since also the sum of the squares of A, B, and C gives the square of the angular momentum, A, B, and C are called the components of a quantity, the angular momentum, whose direction is perpendicular to $\boldsymbol{v}$ and to the radius vector $\boldsymbol{r}$.

It is only in such a sense as that outlined above that there is any meaning to the statements that the angular momentum is a vector whose components are A, B, and C, and whose direction is perpendicular to $\boldsymbol{v}$ and to the radius vector $\boldsymbol{r}$.

Thus, a vector is a set of three quantities which transform as the components of a line transform under a rotation of axes.

Suppose now that we have two vectors whose components are P, Q, R, and X, Y, Z, respectively, and suppose that

$$P = X \tag{35}$$

$$Q = Y \tag{36}$$

$$R = Z \tag{37}$$

then, because each of these sets of three quantities constitute a vector,

$$P' = Pl_x + Ql_y + Rl_z \tag{38}$$

$$Q' = Pm_x + Qm_y + Rm_z \tag{39}$$

$$R' = Pn_x + Qn_y + Rn_z \tag{40}$$

$$X' = Xl_x + Yl_y + Zl_z \tag{41}$$

$$Y' = Xm_x + Ym_y + Zm_z \tag{42}$$

$$Z' = Xn_x + Yn_y + Zn_z \tag{43}$$

Now, because $P = X$, $Q = Y$, and $R = Z$, equations (38) to (43) show us that

$$X' = P' \tag{44}$$

$$Y' = Q' \tag{45}$$

$$Z' = R'. \tag{46}$$

In other words, equations (35)–(37) have reverted to the same type under the transformation.

We see that all that was necessary for this invariance to follow was that P, Q, R, and X, Y, Z, should transform in the same way. It was not even necessary that that way should be the same as the way in which the axes transformed. As a matter of fact the result (44)–(46) would have followed even though the l's, m's, and n's had not been constant but had been functions of x, y, and z, so that they were no longer directional cosines. All that would have been necessary would have been that the l's, m's, and n's of (41)–(43) should be the same as the corresponding l's, m's, and n's of (38)–(40).

The spirit of the idea is even more general than this. For, suppose we have any number of quantities P, Q, R, S, J, and another set of quantities X, Y, Z, U, T, such that

$$P = X, Q = Y, R = Z, S = U, J = T. \tag{47}$$

We must suppose that each of these quantities is defined in some way in the x, y, z system of coordinates. Suppose we consider another system of coordinates x', y', z' related to the first set by

$$\left.\begin{aligned} x' &= f_1(x, y, z) \\ y' &= f_2(x, y, z) \\ z' &= f_3(x, y, z) \end{aligned}\right\} \tag{48}$$

and define quantities P', Q', R', S', J', etc., X', Y', Z', U', T', etc. in the x', y', z' system of coordinates in the same way that the undashed quantities are defined in the x, y, z system of coordinates, and suppose that, when this is done, it can be shown that P', Q', R', S', J' are related to P, Q, R, S, J in the same way as X', Y', Z', U', T' are related to X, Y, Z, U, T. Suppose in fact

$$\left.\begin{aligned} P' &= a_{11}P + a_{12}Q + a_{13}R + a_{14}S + a_{15}J \\ Q' &= a_{21}P + a_{22}Q + a_{23}R + a_{24}S + a_{25}J \\ R' &= a_{31}P + a_{32}Q + a_{33}R + a_{34}S + a_{35}J \\ S' &= a_{41}P + a_{42}Q + a_{43}R + a_{44}S + a_{45}J \\ J' &= a_{51}P + a_{52}Q + a_{53}R + a_{54}S + a_{55}J \end{aligned}\right\} \tag{49}$$

and

$$\left.\begin{aligned} X' &= a_{11}X + a_{12}Y + a_{13}Z + a_{14}U + a_{15}T \\ Y' &= a_{21}X + a_{22}Y + a_{23}Z + a_{24}U + a_{25}T \\ Z' &= a_{31}X + a_{32}Y + a_{33}Z + a_{34}U + a_{35}T \\ U' &= a_{41}X + a_{42}Y + a_{43}Z + a_{44}U + a_{45}T \\ T' &= a_{51}X + a_{52}Y + a_{53}Z + a_{54}U + a_{55}T \end{aligned}\right\} \tag{50}$$

Then it is obvious that because of (47) we shall have

$$P' = X', Q' = Y', R' = Z', S' = U', J' = T'. \tag{51}$$

In other words the equation (47) will remain invariant under the transformation. It is immediately obvious that the result is not limited to three variables x, y, z. We might have had any number of variables $x_1, x_2, x_3, x_4, \cdots$ etc., so that (48) took the form

$$\left.\begin{aligned} x_1' &= f_1(x_1, x_2, x_3, x_4, \cdots) \\ x_2' &= f_2(x_1, x_2, x_3, x_4, \cdots) \\ x_3' &= f_3(x_1, x_2, x_3, x_4, \cdots) \\ x_4' &= f_4(x_1, x_2, x_3, x_4, \cdots) \end{aligned}\right\} \tag{52}$$

etc.

and the a's of (49) and (50) could have been any functions of x_1, x_2, x_3, x_4, etc.

Already we have passed beyond the stage necessary for the mere purpose of the restricted theory of relativity, but the more general conclusions we have reached will find their place in the general theory. The net result of our conclusions is that if two sets of quantities $P, Q, R \cdots$ etc.; $X, Y, Z, \cdots$ etc. transform in the same way, under a certain transformation of the coordinates a statement of equality between the two sets will be true in the transformed system of coordinates. As a special case of all this we have that where there are as many of the quantities $P, Q, R, \cdots$ etc., as there are variables $x_1, x_2, x_3, \cdots$ etc., then, of course, it is certainly sufficient that $P, Q, R, \cdots$ etc. shall transform in the same way as $x_1, x_2, x_3, \cdots$ etc. respectively, and that $X, Y, Z, \cdots$ etc., shall transform in the same way as $x_1, x_2, x_3, \cdots$ etc., in order that such relations as $P=X$, $Q=Y$, $\cdots$ etc., shall be invariant.

Now, we have seen that in the case of the Einstein transformation $x'=\beta(x-vt)$, $y'=y$, $z'=z$, $t'=\beta(t-vx/c^2)$.

As a matter of fact, it is customary for reasons of symmetry to change from the variable t to $u=ict$ where $i=\sqrt{-1}$, so that

$$x' = \beta\left(x + i\frac{v}{c}u\right), \quad y' = y, \quad z' = z, \quad u' = \beta\left(u - \frac{ivx}{c}\right). \tag{53}$$

Any set of quantities X, Y, Z, U, which transforms according to the law

$$X' = \beta\left(X + \frac{ivU}{c}\right), \quad Y' = Y, \quad Z' = Z, \quad U' = \beta\left(U - \frac{ivX}{c}\right)$$

constitute what is called a 4-vector. It is obvious from what we have said above that any equations between two 4-vectors will remain invariant under the restricted theory of relativity.

The restricted theory also makes use of sets of six quantities which transform in a certain way which is characteristic of all such sets. Such sets of six quantities are called 6-vectors. We shall not find it necessary to speak of them here. Thus provided that we decide to express our laws of physics as relations between 4-vectors, we shall quarantee our invariance from the start. But how are we going to be lucky enough to find such suitable 4-vectors in terms of which to express our laws? What was the corresponding problem in classical physics? How was it that we succeeded in talking in terms of vectors? It was because we *defined* the quantities which formed our fundamental concepts in such a way that they should be vectors. Thus the three components of a length, x, y, z constitute a vector (the most fundamental vector since it is the thing in terms of which the vector was defined). A velocity is a vector because it is simply a length divided by a time. An acceleration is a vector because it is a velocity divided by a time. A force is a vector because it is defined as proportional to an acceleration. To say that a force is a vector without specifying the sense in which it is defined is to talk nonsense.

And so, it turns out in the restricted theory of relativity that the reason we encounter such things as 4-vectors, naturally, is that we define the fundamental things which form the basis of our discussion in such a way that they shall be 4-vectors, or of course 6-vectors or scalars.

Thus, to take an illustration: if we consider any particle at any time t x_1, y_1, z_1, ict_1 constitute a 4-vector. If we consider the same particle, or any other particle at another place, x_2, y_2, z_2, ict_2 constitute a 4-vector, and since it is easy to see that the difference of two 4-vectors constitute a 4-vector, x_1-x_2, y_1-y_2, z_1-z_2, $ic(t_1-t_2)$. This is true even if the two particles are quite unrelated, just as those quantities obtained by taking the difference of the coordinates of two quite unrelated points in three dimensions constitute an ordinary vector. It is, of course, also true as a special case, that if x_1-x_2, y_1-y_2, z_1-z_2, represent the distances moved by a particle in the time t_1-t_2, then x_1-x_2, y_1-y_2, z_1-z_2, $ic(t_1-t_2)$ constitute a 4-vector. In the case where the differences are infinitesimal dx, dy, dz, $icdt$ still constitute a 4-vector. Now it is easy to show by making the substitution $dx'=\beta(dx-vdt)$, $dy'=dy$, $dz'=dz$, $dt'=\beta(dt-(v/c^2)dx)$ that

$$dx'^2+dy'^2+dz'^2-c^2\,dt'^2=dx^2+dy^2+dz^2-c^2dt^2.$$

In other words,

$$dx^2+dy^2+dz^2-c^2dt^2 \text{ is an invariant,}$$

so that

$$dt\left[1-\frac{(\dot{x}^2+\dot{y}^2+\dot{z}^2)}{c^2}\right]^{1/2} \text{ is an invariant.}$$

Calling $\dot{x}^2+\dot{y}^2+\dot{z}^2=w^2$, we have that $dt(1-w^2/c^2)^{1/2}$ is an invariant.
It is customary to write $k\equiv(1-w^2/c^2)^{-\frac{1}{2}}$, so that k/dt is an invariant. Now,

multiplying the components of a 4-vector by an invariant, yields, of course, another invariant, so that

$$k\frac{dx}{dt},\ k\frac{dy}{dt},\ k\frac{dz}{dt},\ kic \text{ is a 4–vector,}$$

i.e., kw_x, kw_y, kw_z, kic is a 4-vector. Since k is very nearly equal to unity unless w is comparable with c, the first three components of this 4-vector are very nearly equal to the x, y, and z components of the velocity. The whole 4-vector is called the velocity 4-vector. Again, by taking the difference in the values of the velocity 4-vector at successive instants, we have $d(kw_x)$, $d(kw_y)$, $d(kw_z)$ $d(kic)$, is a 4-vector. Multiplying by the invariant k/dt, we obtain

$$k\frac{d}{dt}(kw_x),\ k\frac{d}{dt}(kw_y),\ k\frac{d}{dt}(kw_z),\ k\frac{d}{dt}(kic)$$

constituting a 4-vector. Since k is very nearly equal to unity, unless w is comparable with c, the first three components of this 4-vector are very nearly the same thing as the components of the acceleration. However, they are quantities well fitted to the use of the theory of relativity, because they are components of a 4-vector. Thus, if there should be other particles present, and we should build up some other 4-vector, P_x, P_y, P_z, icP_t depending upon the position of this first particle in relation to the others, and possibly upon the velocities of the particles, and if we should write down the relations

$$k\frac{d}{dt}(kw_x) = P_x$$

$$k\frac{d}{dt}(kw_y) = P_y$$

$$k\frac{d}{dt}(kw_z) = P_z$$

$$k\frac{d}{dt}(kic) = icP_t$$

the results might be wrong, but they would at least be invariant under the transformation of the restricted theory of relativity.[4]

4 If, in ordinary vector algebra, we equate two vectors A_x, A_y, A_z, and B_x, B_y, B_z, it is necessary for consistency that $A_x B_x + A_y B_y + A_z B_z = |A|\,|B|$. It is necessary that the vectors shall be parallel to each other. So here the consistency of the relations demands a restriction on P_x, P_y, P_z, P_t, and it can easily be shown that this relation is

$$P_xw_x + P_yw_y + P_2w_2 - c^2P_t = 0.$$

Thus, when P_x, P_y, P_z, are assigned, a restriction is placed on P_t. In fact,

$$P_t = \frac{P_xw_x + P_yw_y + P_zw_z}{c^2}.$$

The quantity kw, or $w/(1-w^2/c^2)^{\frac{1}{2}}$ will be recognized as the relativity momentum of a particle of unit rest mass. However, to pursue further the details of relativity particle dynamics at this stage would take us into the realms of a detailed treatise on the theory of relativity, and would be out of place in a general survey of fundamentals. It will suffice to say that in seeking invariant relations to correspond to those of prerelativity physics, our procedure is usually to seek relations between 4-vectors (and sometimes between six vectors) which degenerate when the velocities are small to the old expressions familiar in prerelativity physics.

The Use of the Restricted Principle of Relativity. One is well justified in inquiring as to the ultimate use of the restricted theory of relativity. The answer is this: Suppose we have certain laws such, for example, as the Newtonian law of gravitation, which are very nearly correct, and we desire to formulate them in a more exact form, particularly to take account of what might happen in the case of velocities much greater than those we happen to have encountered in our initial formulation of the law. What have we to guide us? At first sight very little. We may place a very great restriction upon our possibilities by adjusting such modified laws as are proposed to be of a form invariant under relativity. Relativity cannot discover laws of physics for us, but it can act as a sort of supreme court to pass upon such laws as we propose; and, in this act it serves a useful function. Sometimes, the restriction which it imposes, when taken in conjunction with other required restrictions is so drastic as to leave only one final possibility open.

Then another use is illustrated by such a case as the following: Suppose a source fixed with respect to our axes is emitting monochromatic light of a certain frequency. By transferring to another system of axes moving with a velocity v in reference to the first set, it is possible to ascertain how the frequency of the light, wave-length, etc., are affected by a relative motion between the source and the observer. In fact, if we have the laws of any phenomenon given in a case in which one of the objects entering into the production of the phenomenon or indeed merely some identifiable point, has zero velocity, we can always, by relativity, find the corresponding law when the same object or identifiable point, is in motion. Of course, it must be noted that if bodies other than the one referred to are in motion, and if their velocities are involved in the statement of the law for the first case, relativity will have the power to pass upon that expression of the law as an invariant phenomenon or otherwise in that case also.

The Fundamental Content of the Restricted Theory: Removal of Certain Irrelevant Concepts. If we strip from the theory all the concepts incidental to its historical development; the fundamental outstanding dogma

Incidentally, we may remark that the subscript t in P_t is not intended to carry any special significance as relating P_t to the time. It is simply intended to denote the fourth component of the 4-vector, and is used because the fourth component of the 4-vector x, y, z, ict, is the t component.

which remains is that the laws of nature—the differential equations—shall be invariant under the transformation (1)–(4). *Once we have written down some proposed laws*, the test of this conclusion is a matter for pen and paper, and not for experiment. There is no need to form the concept of v as the velocity of some hypothetical observer who exists apart from ourselves. We can even take our proposed law and our transformation (1)–(4) to a mathematician and ask him to test its invariance without telling him what v is, or what c is. In order to deceive him in case he has read about the theory of relativity we can change v to another letter a, so that he will have no suspicion of its being a velocity, and his test will be the same as if he had read all about the classical visionary observers with their clumsy clocks and scales. We need not even tell him that x, y, z refer to space, and t to time. The test is purely an algebraical one. All of the fictitious observers, with their scales and clocks have evaporated, and there remains nothing in the theory of relativity but this question of the algebraical form of the laws in terms of x, y, z and t. But you may now begin to be a little doubtful as to how x, y, z and t are to be measured in order that the theory shall be true. Let there be any one way of measuring x, y, z and t for which the laws, when expressed in terms of those measures do conform to relativity. Then it is obvious that if a demon should take it into his head to alter the things by which you measured x, y, z, t, in such a way that you got different values from before, the laws which you would build up would, in general, not obey relativity. Or conversely, the demon may work the other way. It may be that with your clocks and scales or whatever you use, the laws do not conform to relativity, and the demon may make such adjustment in those instruments while you are making the measurements that the laws when chosen to agree with the numbers you finally write in your notebook do conform to the transformation. When we contemplate the rather arbitrary methods we do adopt for measuring space and time, we are lead to wonder whether it would not be well to state that the content of the theory of relativity is to the effect that there exists *some* way of assigning the numbers x, y, z, t in experiments such that the laws when adjusted to be consistent with measurements taken in that way shall be of a form invariant under the transformation (1)–(4). If we do this, any numbers which we might obtain and which did not lead to consistency with invariant laws are to be regarded as calling for some kind of correction, just as inequalities of temperature of the scales and clocks call for correction. Once we get one set of "corrected" measures, however, there will be an infinity of other measures obtainable by all possible transformations of the type (1)–(4), for which the laws will also be true. Now, I do not wish to raise here the question of whether we could always "correct" the measures in such a manner that in terms of the corrected measures, any law we might like to write down would become true, I wish simply to state that insofar as its working aspects are concerned the theory of relativity is only of value provided that it can successfully demand invariance of laws when expressed in terms of the measurements as we make them. On the other hand, starting with laws expressed initially in some definite set of coordinates, it is to be distinctly understood that mathematical invariance of those laws under the Lorentzian transformation by no means implies that the coordinates for the transformed system would have any relation to actual measures of an observer travelling with a velocity v associated with transformation to that system. It is true that the principle of restricted relativity owed its formulation to a belief that the coordinates associated with the various systems corresponded to the actual measures; but, once formulated, the working content of the theory, involving as it does the mathematical invariance of the laws, is independent of this hypothesis.

It is usually assumed that if the Michelson-Morley experiment gives a positive result the theory of restricted relativity must be abandoned. I do not think that this conclusion can be substantiated. Let us imagine first a system of measurements such that physical laws when expressed in terms of them are mathematically invariant. Let us suppose these measurements to be made in a system, which, in the language of prerelativity days, would be regarded as fixed in space. Suppose that our earth is moving in relation to this "fixed" system with a velocity of about 18 miles per second, as the astronomers of old used to think to be the case, and let us suppose that some demon prevents those alterations of clocks and scales (alterations in relation to those of the fixed system) which the historical aspects of the theory of relativity would demand as a result of the velocity. Then all of our measurements will be a little different from what they would otherwise be; and, if we should test our laws by using the "erroneous" measure we should find them wrong. However, these errors will, in general, be beyond the limits of accuracy of our measurements. For since accurate experiments designed to test any such theory as the electromagnetic theory, for example, depend ultimately, in the last analysis, upon measurements of the distances between two points at a fixed time, or upon the difference between two times at a fixed point,[5] and since it is easy to show that such measurements will only be in error as the result of the supposed operation of the demon to the order v^2/c^2, it will be necessary to measure to this accuracy if the theories in question are to be found wrong. Thus such a great scheme of mathematical regularity as the electromagnetic theory, for example, lying as it were, with one part in 10^8 of our measurements will be discovered by us, although strictly speaking, we ought not to discover it. Then, we shall find the corresponding equations invariant under the transformation (1)–(4). This is not a matter of experiment, but one of pen and paper. Then we shall probably formulate the theory of relativity. We shall set out to test the way in which the mass of an electron varies with the velocity. We ought to find it different from that predicted by relativity, but we should have to measure the one part in 10^8 to discover the difference. Only when we do some experiments such as the Michelson and Morley experiment, which is capable of detecting one part in 10^8 shall we find the discrepancy. In the face of such a discrepancy, however,

[5] This is not true of "ideal," but impracticable experiments. The whole situation is discussed in detail in the writer's paper "Relativity and Aether Drift" in Phys. Rev. **35**, 336–346 (1930).

the valuable part of the theory of relativity still remains, the part which asserts that *there is a system* of measurement such that the laws when expressed in terms of it are invariant under the transformation (1)–(4); and, if there is one system, there is of course an infinity of others obtainable by all possible transformations of the type (1)–(4). The fact that our particular system of measurement may be as slightly different from this, as it would be if, in the languages of 30 years ago, the clocks and scales were not altered by motion, is a circumstance of little practical interest. The practical working feature of the principle of relativity, the feature which has to do with the forms of the laws in a *properly chosen* system of measurement, might still remain.

But it is natural to inquire, why should the laws of nature be invariant under this particular type of transformation? The answer to this question is to be sought in the general theory of relativity where the special theory finds a much more logical birthplace than in that old region populated by imaginary observers travelling with different speeds and measuring the velocity of light.

The General Theory of Relativity. It is not possible in an article of this length to attempt even an elementary account of the general theory in all its essential features, but I wish to refer to the subject for the purpose of emphasizing one or two features which are perhaps not always fully grasped. For one who is interested in relativity primarily in relation to electrodynamics, it may be advisable to omit the reading of this section.

The restricted theory of relativity made its appearance in 1905. In 1915, Einstein brought out his "General Theory of Relativity." This caused so much disturbance in the scientific world that the noise thereof got out into the popular world, which began therefore to hear for the first time of the *restricted* theory of relativity. Now, since the restricted theory of relativity has associated with it some of those more appealing paradoxes which delight the lay mind, and since the "man in the street" was quite unable to comprehend the general theory, we found a state of affairs in which the layman was talking about things more characteristic of the restricted theory under the impression that he was talking about the general theory.

To illustrate the matters of which we must now speak, let us consider a scaffolding extending up into the sky, with three sets of rods, mutually perpendicular to each other, and suppose that there are on these rods marks which are equally spaced, in the sense which you understand the phrase, and are labelled, 1, 2, 3, 4, etc. Suppose further that I am provided with a clock on whose face there are also marks which again are numbered 1, 2, 3, etc., giving me what I would call equal intervals of time. Suppose a sun and a planet move about in this scaffolding, and suppose I consider the Newtonian law as expressed by equations

$$m_a f_a = \frac{G m_a m_b (x_b - x_a)}{[(x_b - x_a)^2 + (y_b - y_a)^2 + (z_b - z_a)^2]^{3/2}}$$

$$m_a g_a = \frac{G m_a m_b (y_b - y_a)}{[(x_b - x_a)^2 + (y_b - y_a)^2 + (z_b - z_a)^2]^{3/2}}$$

$$m_a h_a = \frac{G m_a m_b (z_b - z_a)}{[(x_b - x_a)^2 + (y_b - y_a)^2 + (z_b - z_a)^2]^{3/2}}$$

as a candidate for the expression of the facts of the case. What this means is, of course, perfectly obvious. It means that I take a very large number of readings of the position of the planet a and the sun b, for example, and the corresponding times, so that I am able to express the rate at which the position of the planet changes (its velocity), and the rate at which the velocity changes (the acceleration), and that I proceed to test whether the component accelerations are related to the corresponding space measures by this formula. I am not particularly concerned with discussing what correction I make on the time to allow for the time taken for the light from the planets to reach me. I simply suppose that I have decided upon some way of reading my clock, and possibly correcting its reading, for the purpose of obtaining a number t_a which I can associate with each value of x_a, y_a, z_a, which I measure for a, and a number t_b which I can associate with each value of x_b, y_b, z_b; and having written my data down on paper I proceed to test the law represented by the above equations. Suppose that I find that the law holds. Then that may well be the end of the matter. But, it will be a most astonishing thing if this does happen *exactly* for the reason which I shall try to make clear.

Suppose that I make each rod of my framework of scaffolding double, and that, leaving one frame which I shall call A as it is, I solder the points of intersection of the other, which I shall call B, and then distort the frame in an arbitrary way, leaving the same numbers attached to the points of intersection of the rods. Suppose further that I do a similar thing as regards the clock. In other words, suppose that I duplicate my original clock and leaving the old one as belonging to the system A, I modify the other one, which shall belong to B, in such a manner that it goes highly irregularly, and suppose that I modify my way of using its readings to assign times to the instants when objects coincide with certain intersections on the frames. Suppose I do this in a perfectly arbitary manner so that all that is left of the connection between the new time readings and the old is the fact that for every time reading that I assigned with my A clock as corresponding to the coincidence of a body with a point of intersection of my A frame, there corresponds one, and only one, time as recorded by my B clock for that same coincidence. I shall even go farther and imagine, that as time (recorded by the A clock if you wish) increases, the twisting of the B frame and the modification of the A clock alter. Moreover, I shall suppose that the B clocks and the method of calculation used in conjunction with them are continually changing, so that even the *relation* between the time readings recorded by the B and A clocks for coincidences of bodies with a given point of either frame changes with the time. Suppose I now use the numbers on the B frame to test the law given by the three above equations, which we supposed correct when the A

system was used. It is hardly to be expected that I shall find the law to hold now. But you will say, "of course not, the numbers are all wrong." Let us inquire a little further into this accusation against the numbers, however.

Suppose that when I asked you to imagine that the framework B became distorted I had asked you yourselves to take part in that distortion in every detail. Suppose, in fact, everything took part in this distortion. Then you would be totally unconscious of the change. For, every intersection of two or more lines would be preserved in that distortion in the sense that, if there were a certain identifiable molecule at the point of intersection in the old frame, that same molecule would be found at the corresponding point of intersection in the distorted frame. There might be a scale on which the numbers 1 and 2 and 3 were not separated by anything like equal distances as I viewed them, I who had not taken part in the distortion. As you viewed them, however, they would appear quite satisfactory; for your retina would be distorted, too, and the images of the number 1, 2, and 3 would fall on the same retinal cells as they fell on before the distortion, so that you will be quite satisfied with the picture. It is true that the A rods would appear crooked to you and the numbering on them would appear wrong. You might think that you could discover the inequality of distances between the marks on the B rods referred to above by using a movable scale which just fitted between the marks 1 and 2, and which you might suppose would fail to fit the space between 2 and 3. This supposition would be wrong, however, for, in terms of my standards of measurement, the rod would change in length as you moved it, and in such a way as to preserve the illusion. You would appear to me curiously deformed; but you could not test the crookedness of your backs because the straight edges you used would be crooked, too. You would not even look crooked to each other, because the distortion of the rays of light, and of the retinas of your eyes would be such as to bring the images of the various parts of your neighbors on to the same retinal cells as they fell on before the transformation. You might change your form as you moved about, but you would not know it for similar reasons. The same sort of thing would happen as regards time. You could not discover the irregularity in your clocks because any other clock you used to test it would suffer a like irregularity. You could not time it by your pulse beat because that would go irregularly, too. In fact, to you, everything associated with that crooked measuring system would appear correct. The most humiliating thing from my standpoint is that to you I should appear deformed and possessed of a highly irregular pulse. I should appear to alter my dimensions as I moved about; and, in fact, there would be absolutely nothing which I could do to prove that my system was undeformed and yours was deformed, that you could not do to prove that yours was undeformed and mine deformed. This being so, how do you know what the situation may be at this moment? If I miss out the introduction to this discussion, the picture of the deformation of the framework, etc., and simply start the story where we find ourselves with the difference of opinion as to which scheme of measures is right and which is wrong, you will readily see that there is no way of settling the question. In fact, there is no question to be settled. There is no meaning to one being right and the other wrong. If you think there is, state it, and I will show how the advocate of either system will be able to prove in terms of your criterion that his is the right system. In fact, both systems are on a par with each other; and, since, as you will recall, the difference which I supposed to exist between the labelling of the two systems and the differences between the clocks was perfectly arbitrary, for it arose from an arbitrary distortion of the framework and arbitrary changes in the true measurements, all systems must be regarded as on an equal footing. Now if you can only eliminate from your consciousness the idea that one of these systems is right and all the others wrong, and you certainly ought to be willing to do this unless you can say what you mean by one being right, you will be in a position to admit that if we write down any law of such a kind that it is true when one measuring system (one method of attaching numbers to the events) is adopted and is untrue when another system is adopted, that law must have in it some properties of this system of measurement we have chosen. You will reply, "Of course it has. That is the object of physics, to choose a measuring system and then find laws expressed in terms of it." But if, as I have tried to emphasize to be the case, the system of measurement which we have chosen is a perfectly arbitrary one, should we expect to be able to express in terms of it, and in a simple form, any intrinsic regularity which there may be in Nature? Of course, we certainly can write down laws which are very approximately correct under certain conditions; but have we any right to expect simplicity in the particular set of measures we have chosen if the laws when transformed to any other set of measures, which are on equal footing with ours, assume another form; and, in general, a form which is almost infinitely complex?

It has become customary for the relativist to adopt the standpoint that laws which are not invariant in the sense of the general theory must have in them certain features which depend simply on the system of coordinates chosen; are therefore arbitrary, and can have no place in nature. It has been customary to maintain the stand that the exact laws of nature must conform, at least, to the criterion that they are invariant under a general transformation of coordinates.

It has been truly remarked that, in all physical measurements, the only things which are ultimately observed are coincidences of two or more marks. We say that the deflection of the spot of light was 3 cm; but, what was actually observed was that a certain dark line on the spot coincided with another dark line on the scale. Suppose that, taking all the numbers, x_1, y_1, z_1, t_1, x_2, y_2, z_2, t_2, . . . etc., which we have allotted to the various points in our experiments, including the numbers which have been allotted to the scales themselves it is obvious that if we should renumber all the points (both in space and time) in such a manner as to produce a one to one correspondence between the old system of numbering and the new, the coincidence between the old system of numbering and the new, the coincidence of two events (i.e. position and time) would be symbolized just as well by the equality of the four numbers on the new system as by the four numbers in the old system. Thus, if we intend to imply in our laws no more than is contained in our observations there is no more reason for using one set of numbers than the other.

It is perhaps worthwhile pointing out that since our actual measures *do* make use of properties of the universe in their realization, the numbers which we assign to them do not have quite that arbitrary character which has been implied in the foregoing. The twisting of our scales, and the alterations of our clocks, etc., might well be criticized on the basis that a logical pursuance of the procedure would demand that the phenomena we were examining—the astronomical motions in the case cited—should also be made to suffer the same distortion, as otherwise we might destroy, unfairly, a sort of sympathetic understanding between the motions of the heavenly bodies and the ways of the scales and clocks of our actual experience. Thus, for example, it is a fact that the moon's disk is related to our rods in the sense that it has a simple shape—a circle—when measured with the rod. Again, the planetary orbits are nearly circles as measured by the rod, and so on.

We shall not attempt any further justification or criticism of the principle of invariance but shall now proceed to discuss its meaning in the form in which we meet it in the development of the subject.

The Significance of General Invariance. As a basis to which to pin our remarks, we shall take the general relativity law of planetary motion. For the immediate purpose in view it is unnecessary to enter into the question of how the law was reached.[6] It will suffice simply to take the law as it stands and discuss its significance. The law is as follows: A planet moves according to the differential equations,

$$\frac{d^2 x_\alpha}{ds^2} + \frac{1}{2} \sum_{\mu=1}^{4} \sum_{\nu=1}^{4} \sum_{\lambda=1}^{4} g^{\alpha\lambda} \left(\frac{\partial g_{\mu\lambda}}{\partial x_\nu} + \frac{\partial g_{\nu\lambda}}{\partial x_\mu} - \frac{\partial g_{\mu\nu}}{\partial x_\lambda} \right) \frac{dx_\mu}{ds} \frac{dx_\nu}{ds} = 0 \qquad (54)$$

where

(1) x_1, x_2, x_3, x_4, correspond to the ordinary x, y, z, ict for the planet, so that (54) represents 4 parametric equations for x, y, z, ict, in terms of the arbitrary parameter s, which of course becomes eliminated in the process of getting x, y, z, as functions of t after the equations are solved.

[6] Our avowed purpose is to avoid encumbering our comments on the significance of the general theory with the question of how the invariant laws are arrived at. For the purposes of the discussion it is sufficient to accept (54) and (55) as having fallen from heaven. However, for the benefit of those unfamiliar with the subject, it may be well to add the following: (54) is the equation of a Geodesic. It is the analytical equivalent of stating that the particles move in such a way that

$$\delta \int ds = 0 \qquad (54.1)$$

where ds is give ı by (56). The invariance of (54) is guaranteed by the fact that $\delta \int ds = 0$ is the condition for a maximum or minimum of the integral, and (54.1) is an invariant relation. Even if the particle moved in one system of coordinates in a straight line with constant velocity, *i.e.*, if it moved so that its spacial coordinates changed linearly with the time, we could disguise this simple motion into all sorts of complicated motions by expressing it in terms of other

(2) The sixteen[7] $g_{\mu\nu}$'s are functions of x_1, x_2, x_3, x_4, concerning which, all we say is that they are solutions of a certain definite set of 16 partial differential equations in x_1, x_2, x_3, x_4, which equations could be written down here in full if there were space, but which are symbolized by writing

$$G_{\mu\nu} = 0 . \tag{55}$$

(3) The quantity $g^{\alpha\lambda}$ is the minor of the term $g_{\alpha\lambda}$ in the determinant formed from the sixteen g's.

To the foregoing we must add a very important remark. If we make an arbitrary point to point transformation of coordinates,

$$\begin{aligned} x_1 &= f_1(x_1', x_2', x_3', x_4') \\ x_2 &= f_2(x_1', x_2', x_3', x_4') \\ x_3 &= f_3(x_1', x_2', y_3', x_4') \\ x_4 &= f_4(x_1', x_2', x_3', x_4') \end{aligned}$$

(54) and (55) come back to exactly the same form in dashed coordinates and in dashed g's with this important understanding—that the *dashed g's are related to the undashed g's by the transformation corresponding to that of a covariant tensor of the second rank.* Now in case the reader is unfamiliar with this language, it will form the analytical equivalent of it to say that the dashed g's may be obtained from the undashed g's in the following manner: Construct the quantity ds^2, defined as

$$ds^2 \equiv \sum_{\mu}\sum_{\nu} g_{\mu\nu} dx_\mu dx_\nu \tag{56}$$

where the g's are of course functions of dx_1, dx_2, dx_3, dx_4. Now make the transformation. It is easy to see that ds^2 in the new coordinates assumes the form of a homogeneous quadratic function in $dx_\mu' dx_\nu'$; and the coefficient of $dx_\mu' dx_\nu'$ is the quantity $g_{\mu\nu}'$ referred to above.

coordinates related to the first set by some general transformation. We might even be lead to wonder whether the complicated motions of the heavens could not all be represented as straight lines with constant velocity in some suitably chosen system of coordinates. If such were the case, the law of astronomy would be that a particle moved according to (54) where the g's were the sort of g's which could have come from the g's 1,1,1,1, by some transformation or other. The condition that these sorts of g's must satisfy can be expressed as a set of differential equations governing the g's, which differential equations are usually written

$$B^{\tau}_{\mu\nu\sigma} = 0 .$$

The notation need not detain us. It will suffice to say that it can be shown that in all the possibilities available under such a situation, there are none which will give a motion approximating to the planetary motions. The next question is whether there are any other differential equations which the g's could satisfy, differential equations which would deny that they could have come from the g's 1, 1, 1, 1, by transformation, but which would permit solutions which when substituted in (54) made it fit the facts. The answer is that there is such a set, and (55) is the set in question.

[7] Owing to the fact that in practical cases $g_{\mu\nu} = g_{\nu\mu}$ the number of the g's reduces to 10. However, we do not wish to encumber the exposition with this fact at the present time.

In other words, the relation between the dashed and the undashed g's is that implied in

$$\sum_{\mu\nu} g_{\mu\nu} dx_\mu dx_\nu = \sum_{\mu\nu} g_{\mu\nu}' dx_\mu' dx_\nu'. \tag{57}$$

Now let us elaborate a little on the foregoing. In ordinary dynamics, we have equations of motion analogous in a sense to (54), but of the form, for example

$$\frac{d^2 x_\alpha}{ds^2} - \text{Force} = 0. \tag{58}$$

Our business is to solve the four equations, assign the initial values of the x_α and the dx_α/dt, and the problem is finished. Equation (54) is not as definite as (58), however. Before we can solve it we have to assign not only initial values of the x_α and dx_α/dt, but we have to assign the g's which occur in the parenthesis. For these we look to equation (55), and all this tells us is that we can have any g's we like, so long as they satisfy (55). In fact, these two equations (54) and (55) together are much more general—they say much less than an equation like (58). That is the penalty we have to pay on their behalf for their luxurious heritage of invariance. Of course, when we start to solve (55) we again need to *assign* certain values of the g's, certain boundary values as they are called. In practise we do not usually do this. What we do is to get *some* set of g's which satisfy (55) and which when substituted in (54) make that equation give final results approximately like the Newtonian equation symbolized by (58). When we have done this we can state that our ultimate solution of the problem has been built up as a special case of a set of equations having the characteristic of invariance. Indeed, there would be nothing *surprising* if we had found that we could have reproduced exactly our old Newtonian equation as such a special case. The result would be very instructive to our thinking if we could. For in that case, one person could look at our final equations and say they are not invariant at all. They are not invariant even for the simple transformation of the restricted theory let alone the general transformation of the general theory. And we should say to him, "You are looking at invariance in the wrong way. You must not boil the parenthesis of (54) down to an exact expression in the coordinates and then ask for invariance. The equations are not as invariant as all that. The parenthesis will not appear in the same form in the coordinates on transformation. It will appear in the same form in the derivatives of the g's with the understanding that the dashed g's are related to the undashed g's according to the principle stated above, and implied in (57)."

Let us view the matter from another angle. Suppose I should consider two beings A and B who start to study an identical physical problem—an actual planetary motion in two different sets of coordinates.

A will, or should, assign his boundary values to the g's and solve (55). For the moment I am not interested in saying how he is to assign the boundary values. Next he will substitute for the g's in (54). He will then assign the necessary initial (or boundary) values for the appropriate quantities concerned in (54) and proceed to his solution. Now B will start with equations like (54) and (55) with everything dashed; but, in order that he shall get the same solution as A[8] it will be necessary that first he shall assign not the same but corresponding boundary values in (54), corresponding in the sense that one is obtained from the other by the transformation which takes us from A's coordinates to B's. This is not all, however. For, before B gets his equation (54) he must solve (55) and here he must assign the boundary values to the dashed g's so that the boundary values of the dashed g's are related to the boundary values of the undashed g's assigned by A, by the law implied in (57). But how is B to know what boundary values A assigned? There is nothing simple which he can measure to get his boundary g's as he could measure the position and velocity of the planet to assign boundary values to (54). The answer to the conundrum is that really neither A nor B have any immediate way of assigning boundary values to the g's. They must leave them arbitrary—at least, if they are going to be respectable mathematicians and avoid the guessing process referred to above. They must solve (55) with arbitrary boundary g's. They must let these arbitrary quantities get into (54). They must assign the boundary values peculiar to (54) by direct observation, and solve (54). The solutions will then still be indefinite to the extent of the arbitrariness of the boundary values of the g's still remaining in that solution. The whole content of the truth of the theory now lies in the possibility that A shall be able to assign such boundary values as will make the solution of (54) fit the facts. He will have to take a lot of points on the orbit, determine the boundary values in question and then see if his solution fits all the other points on the orbit. It is obvious that in general there will only be one set of boundary values (if any) which will do this. Moreover, we may feel quite certain that if B does the same thing, the boundary values of the dashed g's assigned by B will be related to those assigned by A by the relation implied in (57), for that must follow as a mathematical necessity.

Perhaps an apology is needed for this long digression on the significance of invariance which is usually dismissed in so few words. I tender such an apology to those whose conscience permits them to accept it.

The Relation between the Restricted Theory and the General Theory. One very important point remains. We have interested ourselves in the quantity

$$ds^2 \equiv \sum_\mu \sum_\nu g_{\mu\nu} dx_\mu dx_\nu \tag{59}$$

Now, if we consider the quantity

$$dx_1^2 + dx_2^2 + dx_3^2 + dx_4^2 \tag{60}$$

and make an arbitary transformation of coordinates it assumes the form (59) as is very well known and easily seen. However, even with an *arbitrary* transformation, we cannot produce from (60) all possible expressions like (59).

[8] The same in the sense that the motion calculated by B will be exactly the same as that obtained by superposing the transformation of coordinates on A's calculation.

We cannot make the g's any function of x_1, x_2, x_3, x_4 we like by an arbitrary transformation. Now there is a set of coordinates (we call it an Euclidean set) in which far away from what we call matter a material particle moves in a straight line with constant velocity. In other words this particle obeys the law

$$\frac{d^2x_\alpha}{ds^2} = 0$$

which is a particular case of (54), when the parenthesis is zero, as it would be if the g's were constants. It is consequently customary to limit our interest to those g's which we get as solutions of (55) and which at infinite distance[9] from matter assume such functions of the coordinates that it is possible to find a transformation which will transform them to the g's, 1, 1, 1, 1, characteristic of (60). How if we are in a set of coordinates in which the g's are 1, 1, 1, 1, i.e., in which

$$ds^2 = dx_1^2 + dx_2^2 + dx_3^2 + dx_4^2$$

we know that a *Lorentzian transformation of the restricted theory of relativity*, i.e., the transformation (1)–(4) will leave those g's equal to 1, 1, 1, 1. This is the fundamental characteristic of the Lorentzian transformation. One of the best known ways of determining the transformation is to determine it as the transformation which leaves $dx^2+dy^2+dz^2-c^2dt^2$ invariant. Hence for this transformation, the g's remain 1, 1, 1, 1, i.e. they certainly remain constants. Hence if there be one set of coordinates in which the g's disappear from the equation owing to their being constants whose derivatives consequently are zero, then they will disappear in any other set of coordinates obtained from the first by a Lorentzian transformation. Hence, if there be any set of equations (containing derivatives of the g's) which are invariant in the sense of the general theory of relativity, then the equations obtained by putting the derivatives of the g's equal to zero will be invariant in the more drastic sense of the restricted theory of relativity. It would thus seem that the restricted theory of relativity finds a more logical origin as a special case of the general theory, an origin which emphasizes its bearing on the mathematical forms of the expressions, than it finds in arguments having to do with different observers moving relatively to each other and performing certain hypothetical experiments.

The Significance of the Velocity of Light in the General Theory. The customary method of approach to the general theory, is a sort of generalization of the restricted theory. The velocity of light c, which, again in the customary approach, makes its appearance through the activities of hypothetical observers crawls into the general theory by a sort of back door, which leaves one with the sort of feeling of surprise in the end that it should

[9] The significance of infinity is ambiguous in an arbitrary unknown set of coordinates. All that we ought to say is that it is possible to assign a set of coordinates such that the g's when expressed in terms of them assume forms transformable to 1, 1, 1, 1 at infinite values of those coordinates.

be found in the general theory at all. By removing the artificial and unnecessary steps peculiar to the historical development, it is possible to place these matters on a more satisfactory basis. The problem of particle motion will provide a sufficiently general basis to illustrate the idea.

Starting with no thoughts of light in our minds, let us proceed from (54) and (55). As is well known, and as may be verified from all the classical treatises on the subject, one solution of (55) for the g's in terms of r, θ, ϕ, makes the ds^2 of (56) take the form

$$ds^2 = -\gamma^{-1}dr^2 - r^2 d\theta^2 - r^2 \sin^2\theta d\phi^2 + c^2\gamma dt^2 \tag{61}$$

where $\gamma \equiv 1 - 2m/r$, m, and c being merely constants. Formally, there is no reason to recognize r, θ, ϕ as polar coordinates, although the particular solutions of (55) which have been chosen have been chosen with the idea that ds^2 comes out of a form closely resembling that which we have been accustomed to speak of as the line element in the polar coordinates. I do not wish to pursue this matter because it would lead us into fields interesting in themselves, but calculated to obscure the points really at issue. Since any transformation of coordinates will give from the g's of (61) another set of g's which are also solutions of (55), it is perfectly obvious that we could omit the c^2 in (61), and indeed could modify the expression by writing $r = ar'$, $\theta = b\theta'$, $\phi = k\phi'$ if we wished, and still have g's which were solutions of (55). But in order that we may have an ultimate solution for the planetary motion it is necessary that we shall not take *any* solutions of (55), but the *right* ones. In including c^2 in (61) we have anticipated to some extent the needs of the final solution. It is now a question of mere algebra, a question which need not detain us, and which will be found elaborated in any of the standard books, to show that with the g's specified in (61), equation (54) may be shown to result in an orbit which is in a plane of constant θ (taken arbitrarily as $\pi/2$), and in which the motion is governed by

$$\frac{d^2u}{d\phi^2} + u = \frac{m}{h^2} + 3mu^2 \tag{63}$$

$$r^2\frac{d\phi}{dt} = h\left[c^2\gamma - \gamma^{-1}\left(\frac{dr}{dt}\right)^2 - r^2\left(\frac{d\phi}{dt}\right)^2\right]^{1/2} \tag{64}$$

where $u \equiv 1/r$, and where h is a constant of integration.[10]

Integrating (63), with approximations for convenience (though not of necessity)[11] we arrive, in the usual way at

[10] The form usually given for the equation is $r^2 d\phi/ds = h$. We have substituted for ds/dt from (61).

[11] Our wording is here a little careless, since there has been no mention of the sizes of quantities to warrant speaking of approximations. Our association of r, θ, ϕ, with polar coordinates is purely arbitrary, founded on a determination to find *some* solution of (54) and (55) which shall ultimately give an orbit approximating that which we find in terms of those coordinate measures which we actually make and which we call polar coordinate measures. Having made our association of coordinates, the ultimate association of the constants m and h with measured

$$u = \frac{m}{h^2}[1 + e \cos(\phi - \omega - \delta\omega)] \tag{65}$$

where e and ω are constants for the orbit, but

$$\delta\omega = \frac{3m^2}{h^2}\omega. \tag{66}$$

Apart from the small term $\delta\omega$, an examination of the orbit would serve to determine only m/h^2, and not m and h separately. The term $\delta\omega$ serves, however, to fix m^2/h^2, so that by its inclusion, both m and h are determined from experimental observations. The determination of h now enables c to be determined by the substitution of measured values of $d\phi/dt$ in (64).[12] The type of variation of ϕ with the time necessitated by (64) is of course, a requirement of the theory. In practise, since c turns out to be so large compared with dr/dt or $rd\theta/dt$, and since γ is very nearly equal to unity, (64) may be written with sufficient approximation as

$$r^2 d\phi/dt = hc$$

Without knowledge of h it would have been impossible to determine c. This quantity c thus finds its natural origin in the departure of the motion from the Newtonian law, and it plays no part in the principal part (the Newtonian part) of the motion.

But now there is something else. It can be shown that if, in the sense defined by

$$v = \left[\left(\frac{dr}{dt}\right)^2 + r^2\left(\frac{d\theta}{dt}\right)^2 + r^2 \sin^2\theta\left(\frac{d\phi}{dt}\right)^2\right]^{1/2} \tag{67}$$

a particle has a velocity v which is less than c, equations (63) and (64) will never allow it to attain a velocity greater than c. The existence of a maximum value c for the velocity attainable by a particle reminds us of the statement that no particles can move faster than the velocity of light, so we are tempted to inquire whether the velocity of light (understood as measured far from matter, of course,) will serve for the quantity c, which had its origin in the deviations from the Newtonian law of planetary motion. We find that the velocity of light will so serve.

Again, since the maximum velocity which a particle can attain is equal to the velocity of light, we are tempted to try whether a ray of light will follow the path of a particle which (at infinity of course) travels with the velocity of light. In this way we are led to the idea of the bending of light by the sun in a manner which may appear more suggestive to some than

---

quantities for this orbit insures the validity of neglecting $3mu^2$ compared with m/h^2 as a first approximation, and using the approximate solution thus obtained in the term $3mu^2$ itself. It is in the sense of anticipating this that our phrase concerning approximation has correct significance.

[12] dr/dt may be expressed, of course, in terms of $d\phi/dt$ through the equation of the orbit.

one which leads to it through what experts in the subject will recall as the Principle of Equivalence.

The Symmetry of the Line Element at Infinity. The fact that when $r = \infty$, (61) reverts to a form which when expressed in cartesian coordinates becomes

$$ds^2 = c^2dt^2 - dx^2 - dy^2 - dz^2 \tag{68}$$

may, at first sight, seem to carry with it the conclusion that the restricted theory must necessarily hold at infinity (at least as far as the laws under discussion are concerned). In accordance with the principles enunciated in the section on "The Relation between the Restricted Theory and the General Theory," this conclusion would indeed be expected if we were certain that the coordinates utilized in (68) represented the actual measures of observers. The fact that (68) has resulted from putting $r = \infty$ in (61) does not necessarily warrant this conclusion as to the coordinates. For suppose that, having written down (61), we had, before proceeding made a transformation such as[13]

$$x' = x - vt, \quad y' = y, \quad z' = z, \quad t' = t. \tag{69}$$

The new g's appropriate to the new expressions of the line element would be solutions of (55), for *no* transformation will rob them of this property. In proceeding to our solution in these coordinates, we should of course obtain exactly the solution usually obtained, i.e. the solution corresponding to (63) and (64), but transformed to the dashed coordinates specified in (69). We should obtain exactly the result which would be expected by a physicist of mid-Victorian times who had somehow or other been lead empirically to the equations of motion (63) and (64) as applicable to the sun at rest, and had then supposed the sun to move through space with a velocity v. When we proceed to infinity with our x', y', z', t', coordinates, however, we no longer obtain a line element of the form (68); and, all that the *general* theory predicts as regards a *restricted* theory is that if we make the transformation back from x', y', z', t', to x, y, z, t, we shall find ourselves in a set of coordinates such that the laws expressed in terms of them will be invariant under the Lorentzian transformation. They will not be so invariant to one who has expressed them in terms of x', y', z', t', however.

Electrodynamics

Statement of the Fundamental Equations. The laws of electrodynamics are embodied in the well-known scheme of equations[14]

$$\frac{1}{c}\left(\rho \boldsymbol{u} + \frac{\partial \boldsymbol{E}}{\partial t}\right) = \operatorname{curl} \boldsymbol{H}, \tag{70}$$

$$\rho = \operatorname{div} \boldsymbol{E}, \tag{71}$$

[13] We should really express this transformation in terms of polar coordinates, but for simplicity have left it in cartesians. This fact will cause the reader no difficulty.

[14] The units here used are the Heavisidian units, in which the unit of charge and unit magnetic pole are respectively $1/\sqrt{4\pi}$ times the electrostatic and magnetostatic units.

$$-\frac{1}{c}\frac{\partial \boldsymbol{H}}{\partial t} = \text{curl}\,\boldsymbol{E}, \tag{72}$$

$$0 = \text{div}\,\boldsymbol{H}, \tag{73}$$

together with a fifth relation whose aim is to describe the law of motion of an electron. This law states that the electron is a spherical shell of charge when viewed in a system of axes in which its center has zero velocity, and that in this system of axes, its acceleration and the other time derivatives of its motion are determined in such a manner that

$$\iiint \rho \boldsymbol{E} d\tau = 0. \tag{74}$$

In the above equations, ρ is the charge density, $\boldsymbol{E}$ and $\boldsymbol{H}$ the magnetic fields $\boldsymbol{u}$ the velocity, $d\tau$ an element of volume, and c a constant, which experiment shows to be equal to the velocity of light in vacuo.

It is customary to express (74) in another form. We first split the integral into two parts, one pertaining to the field $\boldsymbol{E}_i\boldsymbol{H}_i$, produced by the charge on the electron itself, disentangled from the total field in a manner to be explained later and the other, the field $\boldsymbol{E}_0\boldsymbol{H}_0$, due to all the remaining charges of the system. Equation (74) may be then written

$$\iiint \rho \boldsymbol{E}_0 d\tau = -\iiint \rho \boldsymbol{E}_i d\tau. \tag{75}$$

It is possible to calculate $\boldsymbol{E}_i$ in terms of the acceleration $\ddot{S}$, rate of change of acceleration, $\dddot{S}$, and higher time derivatives of the motion, so that expressing the right hand side of (75) in this way, the equation assumes the form

$$\iiint \rho \boldsymbol{E}_0 d\tau = \frac{e^2}{6\pi a c^2}\ddot{S} - \frac{e^2}{6\pi c^3}\dddot{S} + \cdots, \text{etc.} \tag{76}$$

where a is the radius of the spherical shell, and e the electronic charge.

The left hand side of (75) being called the external force on the electron, the quantity $e^2/6\pi a c^2$, which is the coefficient of the acceleration on the right hand side is appropriately known as the mass of the electron, it being understood that (76) is applicable only to the case when the velocity of the electron is zero in the system of axes to which $\boldsymbol{E}$ and $\boldsymbol{H}$ are referred. In many cases it is customary to neglect altogether the terms on the right hand side of (75) other than the first, and to assume moreover that the external field is constant over the electron. In this case, (76) assumes the form

$$e\boldsymbol{E}_0 = \frac{e^2\ddot{S}}{6\pi a c^2}\cdot \tag{77}$$

In extending the law of motion of the electron to cases where the velocity of the electron is not zero in the system of axes in which $\boldsymbol{E}$ is measured, it is customary to invoke the assumption of the restricted theory of relativity which, at any rate for the case where $\boldsymbol{E}_0$ and $\boldsymbol{H}_0$ may be assumed constant over the electron, leads via (75) to the result[15]

$$e\left(\boldsymbol{E}_0+\frac{[\boldsymbol{w}\boldsymbol{H}_0]}{c}\right)=\frac{d}{dt}\left\{m_0\boldsymbol{w}\Big/\sqrt{1-\frac{w^2}{c^2}}\right\} \tag{78}$$
$$+\text{ terms involving higher time derivatives}$$

where $m_0=e^2/6\pi ac^2$, and $\boldsymbol{w}$ is the velocity of the center of the electron. For the case of an acceleration parallel to the velocity, the first term on the right hand side of (78) assumes the form $m_0\dot{\boldsymbol{w}}/(1-w^2/c^2)^{3/2}$, and for the case of an acceleration perpendicular to the velocity it assumes the form $m_0\dot{\boldsymbol{w}}/(1-w^2/c^2)^{\frac{1}{2}}$. We consequently speak of

$$\text{The Longitudinal Mass }\frac{e^2}{6\pi ac^2}\left(1-\frac{w^2}{c^2}\right)^{-3/2}$$

and

$$\text{The Transverse Mass }\frac{e^2}{6\pi ac^2}\left(1-\frac{w^2}{c^2}\right)^{-1/2}.$$

On the foregoing view, the quantity which is really measured when the mass is measured for small velocities is the quantity $e^2/6\pi ac^2$, and the value of a is then calculated as the radius of a sphere which would give rise to that mass.

From the standpoint of classical electrodynamics and apart from quantum phenomena, these equations are regarded as applying universally outside, and even inside matter, so long as proper account is taken of the fields and charges arising from the atomic structure. The quantities usually denoted by $\boldsymbol{B}$ and $\boldsymbol{D}$, as distinct from $\boldsymbol{H}$ and $\boldsymbol{E}$ arise in the process of building up equations limited to application on a macroscopic scale. From the standpoint of the electron theory such equations owe their existence to an endeavor to amplify the simple equations to enable them to take account, at any rate approximately, of the presence of matter as far as its macroscopic effects are concerned, without invoking a detailed knowledge of the intramolecular structure. Since the present article is concerned with electrodynamics, mainly in its fundamental aspects, we shall not give any further consideration to the macroscopic equations.

Our statement above concerning the universal applicability of the equations for free space was qualified by the phrase "Apart from quantum phenomena." This reservation is of course very vital. For, in the last analysis, all phenomena are manifestions of atomic phenomena even when disguised in macroscopic form. The motions in a motor are manifestations of what are ultimately atomic phenomena. The best that can be said for the universal applicability of the equations is that while we do not regard them as applic-

[15] If $\boldsymbol{E}_0$, $d\tau$, and $\boldsymbol{H}_0$ are not constant over the electron, the left-hand side of (77) assumes the form $\iiint\rho(\boldsymbol{E}_0+[\boldsymbol{u}\boldsymbol{H}_0]/c)d\tau$, but with the various elements of the integral calculated at different times.

able to all cases of interactions between electrons, they do give, when averaged for such large scale phenomena as those concerned in the motor and other appliances of engineering practice, results which are correct and which are therefore the same as would be obtained by averaging from the fundamentally correct laws.

The status of the equations of electrodynamics in a scheme of modern quantum theory is hardly yet established in complete logical form. We shall, however, leave this matter for the present and discuss first the historical development and significance of the equations in relation to what we may call the classical period immediately prior to the advent of the quantum theory.

Historical Survey.—At the time when Maxwell commenced his epoch-making investigations the known laws of electrodynamics were summed up in conclusions which are the analytical equivalent of the following:

$$4\pi I = \oint H_s ds, \tag{79}$$

expressing the fact that the line integral of the magnetic field around a closed circuit encircling a current I is equal to $4\pi I$.

Then we have the relation[16]

$$-\frac{d}{dt}\iint H_n dS = \oint E_s ds, \tag{80}$$

expressing Faraday's law, that the rate of change of the surface integral of the normal component $\boldsymbol{H}_n$ of the magnetic field taken all over a cap bounded by a circuit is equal to the negative of the line integral of the complete electromotive intensity[17] taken all around the circuit. Finally we have a law having to do with the mechanical forces on a circuit carrying a current, and equivalent to the statement that in the case of a number of circuits expressed by generalized coordinates q_r, the generalized force F_r necessary to hold the circuit in position against the actions of the other circuits is[18]

$$F_r = -I_r \frac{\partial}{\partial q_r}\iint H_n dS. \tag{81}$$

[16] We are confining ourselves to the case of free aether, so that we use $\boldsymbol{H}$ instead of $\boldsymbol{B}$ in (80). The introduction of $\boldsymbol{B}$ and $\boldsymbol{D}$ into the discussion would carry with it more complexity of an irrelevant kind than is worth while for the purpose of the present historical survey. Our units are here, of course, the ordinary electromagnetic units, since the introduction of Heavisidean here would add needlessly to the apparent artificiality of the equations of the pre-Maxwell pe riod.

[17] Strictly speaking, the pre-Maxwellian era was acquainted only with the right hand side of (80) taken as a whole, i.e. with the electromotive force. The dissection of this electromotive force into the line integral of a quantity $\boldsymbol{E}_s$ anticipates the later stage of development in which the thought is concentrated on the force on an element of the electric current.

[18] The directions are such that viewing the circuit from that side from which the current appears to flow clockwise, the positive direction of the normal is drawn from the observer through the cap.

The negative of this quantity, $f_r(=-F_r)$ is the generalized force on the circuit due to the other circuits, so that:

$$f_r = I_r\frac{\partial}{\partial q_r}\iint H_n dS, \tag{82}$$

which is the form in which the expression is usually thought of.

Equation (82) is of course the analytical equivalent of the law obtained by calculating the forces between the current circuits by treating them in terms of their equivalent shells, and it is also the analytical equivalent of the statement that the forces on a current circuit as a whole, due to the other circuits is the same as that obtained by attaching to each element of length ds of the conductor a force δf_r, given by

$$\delta f_r = [\boldsymbol{I}_r\cdot\boldsymbol{H}]\delta s \tag{83}$$

Maxwell set himself the task of moulding equations (80) and (81) into a form which made them appear as the outcome of general dynamical laws.

In the hands of Maxwell, equation (79) amounted practically to a definition of $\boldsymbol{H}$ in terms of the currents. It is not difficult to show that (79) contains as an analytical consequence, that insofar as $\boldsymbol{H}$ is determined by the current alone, the flux of $\boldsymbol{H}$ through any circuit denoted by subscript r, is given by

$$\iint \boldsymbol{H}_n dS = L_{1r}I_1 + L_{2r}I_2 + \cdots L_{rr}I_r + \text{etc.} \tag{84}$$

where the L's are the ordinary coefficients of self and mutual induction, and are perfectly definite functions of the geometry of the system, given by

$$L_{mn} = \iint \frac{\cos\phi\, ds_m ds_n}{r} \tag{85}$$

ϕ being the angle between the elements ds_m and ds_n of the circuits concerned.

The Faraday law (80), for the circuit r becomes

$$-\frac{d}{dt}(L_{1r}I_1 + L_{2r}I_2 + \cdots L_{rr}I_r + \text{etc.}) = p_r \tag{86}$$

where p_r is the electromotive force $(=\int \boldsymbol{E}_s ds)$ produced in the circuit as a result of the varying currents in the circuit and the other circuits. Or, if $P_r = -p_r$ is the generalized external force necessary to result in the condition of changing currents in the circuit r.

$$\frac{d}{dt}(L_{1r}I_1 + L_{2r}I_2 + \cdots L_{rr}I_r + \text{etc.}) = P_r. \tag{87}$$

Again, the law (81) for the generalized force acting on the circuit r becomes[19]

[19] Of course, $\partial L_{rr}/\partial q_r$ is zero, and is merely included for symmetry.

$$F_r = -I_r\left(I_1\frac{\partial L_{1r}}{\partial q_r} + I_2\frac{\partial L_{2r}}{\partial q_r} + \cdots I_r\frac{\partial L_{rr}}{\partial q_r} + \text{etc.}\right). \tag{88}$$

Thus the problem is to represent (87) and (88) as a consequence of generalized dynamics. The procedure is now almost obvious. Lagrange's equations for generalized coordinates are of the form:

$$\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{q}_m}\right) - \frac{\partial T}{\partial q_m} = \Phi_m \tag{89}$$

where Φ_m is the generalized force associated with the coordinates q_r. Suppose, in addition to the mechanical coordinates, we think of electrical coordinates which we shall make appear only in our kinetic energy function through their velocities, which we shall associate with the various currents $I_1, I_2 \ldots\ldots$ Suppose in fact, we take for T the expression:

$$T = \tfrac{1}{2}L_{11}I_2^2 + L_{12}I_1I_2 + L_{13}I_1I_3 + \cdots \tfrac{1}{2}L_{22}I_2^2 + L_{23}I_2I_3 + L_{24}I_2I_4 + \cdots \text{etc.} + T_m, \tag{90}$$

where T_m is the expression we should have for the kinetic energy of the matter of the wires in absence of any currents. Then, applying Lagrange's equations to the electrical coordinate represented as regards its velocity by I_r, we obtain (since T does not involve the electrical coordinates)

$$\frac{d}{dt}(L_{1r}I_1 + L_{2r}I_2 + \cdots) = P_r \tag{91}$$

which is the same as (87) the equivalent of the Faraday law. And, by applying the equations to the mechanical coordinate q_r, we obtain

$$-I_r\left(I_1\frac{\partial L_{1r}}{\partial q_r} + I\frac{\partial L_{2r}}{\partial \dot{q}_r} + \cdots I_r\frac{\partial L_{rr}}{\partial q_r}\right) + \frac{d}{dt}\left(\frac{\partial T_m}{\partial \dot{q}_r}\right) - \frac{\partial T_m}{\partial q_r} = \Phi_r. \tag{92}$$

The part involving T_m is simply the part of Φ_r necessary to accelerate the brute matter of the conductors. In the case where the conductors are held fixed, Φ_r represents what we have called F_r and (92) becomes

$$-I_r\left(I_1\frac{\partial L_{1r}}{\partial q_r} + I_2\frac{\partial L_{2r}}{\partial q_r} + \cdots I_r\frac{\partial L_{rr}}{\partial q_r}\right) = F_r \tag{93}$$

which is the same as our law (88), which is the equivalent of (81).

We may regard the achievement of Maxwell as that of finding the necessary expression for T to permit of the moulding of the equations into the dynamical form in the sense we have described.

Maxwell next endeavored to extend the theory to open circuits, and to the motions of charges. Writing (79) in differential form it becomes

$$4\pi \boldsymbol{i} = \text{curl}\, \boldsymbol{H} \tag{94}$$

where $\boldsymbol{i}$ is now the current density, it being understood, however, that the stream lines of the flow continue themselves into complete circuits.

The thought is that $\boldsymbol{i}$ may be regarded as a flow of some density with a velocity $\boldsymbol{u}$, and that we may be able to write

$$\frac{4\pi\rho\boldsymbol{u}}{c} = \text{curl}\,\boldsymbol{H}. \tag{95}$$

The factor c is included because, if ρ is in electrostatic units, our thought of the similarity between what has been called current density and $\rho\boldsymbol{u}$ has implied only a proportionality between $\rho\boldsymbol{u}$ and $\boldsymbol{i}$, since the unit of $\boldsymbol{i}$ in equation (94) has been chosen so that $\boldsymbol{H}$ shall represent the force on unit pole in such cases where that test is applicable. c is in fact, the ratio of the electrostatic to the electromagnetic unit of charge.

With equation (95) however, we immediately encounter a difficulty; for, since div. curl $\boldsymbol{H}=0$, that equation would imply

$$\text{div}\,\rho\boldsymbol{u} = 0.$$

But, if we are to regard electricity in the light of an indestructible entity, its motion must be controlled by the equation of continuity

$$\text{div}\,\rho\boldsymbol{u} + \frac{\partial\rho}{\partial t} = 0 \tag{96}$$

so that (95) must be regarded as impossible for even such a simple case as that of a moving charge of finite spacial extent, since here the density at any point changes with the time.

In order to escape this difficulty, Maxwell generalized (95), which is strictly true only when $\partial\rho/\partial t=0$, to the equation

$$\frac{4\pi}{c}(\rho\boldsymbol{u}) + \frac{1}{c}\,\frac{\partial\boldsymbol{E}}{\partial t} = \text{curl}\,\boldsymbol{H} \tag{97}$$

which, on taking the divergence of both sides gives

$$4\pi\,\text{div}\,\rho\boldsymbol{u} + \frac{\partial}{\partial t}\,\text{div}\,\boldsymbol{E} = 0 \tag{98}$$

so that, provided that

$$4\pi\rho = \text{div}\,\boldsymbol{E} \tag{99}$$

(98) becomes

$$\text{div}\,(\rho\boldsymbol{u}) + \frac{\partial\rho}{\partial t} = 0$$

so that the situation is now consistent with the equation of continuity.[20]

The quantity $\boldsymbol{E}$ occurring in (99) is of course, measured in electrostatic units, while the $\boldsymbol{E}$ in (80) is associated with a current I measured in the same

[20] The generalization on (95) implied in the addition of the term involving $\partial E/\partial t$, is of course not the only one possible. We might add in addition any vector whose divergence was zero without violating (96).

units as are employed in (79). Hence, if $\boldsymbol{E}$ is to be in electrostatic units, we must divide the left-hand side of (80) by c; and, expressed in differential form, and applied to a fixed circuit of infinitesimal dimensions, the equation would then become

$$-\frac{1}{c}\frac{\partial \boldsymbol{H}}{\partial t} = \operatorname{curl} \boldsymbol{E}. \tag{100}$$

It will here be noted that we have a partial time derivative in place of the total time derivative of (80) which latter equation is applicable to a moving circuit as well as to a fixed one.

Equation (82), with the current (there in electromagnetic units) regarded in the light of a generalization to moving charge density becomes the equivalent of stating that the force per unit volume on the electricity moving with velocity $\boldsymbol{u}$ is $\rho[\boldsymbol{u}\cdot\boldsymbol{H}]/c$, the c being introduced because ρ is in electrostatic units. If we introduce the further idea that, in the presence of an electric field $\boldsymbol{E}$ there would be a force per unit volume equal to $\boldsymbol{E}\rho$,[21] in addition to $\rho\ [\boldsymbol{u}\cdot\boldsymbol{H}]/c$ the total force per unit charge would be

$$\text{Force per unit charge} = \boldsymbol{E} + \frac{[\boldsymbol{uH}]}{c}\cdot \tag{101}$$

If we take a new unit of charge strength and a new unit of magnetic pole strength each respectively $1/(4\pi)^{\frac{1}{2}}$ of the electrostatic and magnetostatic units, and if the fields $\boldsymbol{E}$ and $\boldsymbol{H}$ be defined as the forces on these units, the 4π disappears, as is well known from (97) and (99), and our equations become

$$\frac{1}{c}\left(\rho\boldsymbol{u} + \frac{\partial \boldsymbol{E}}{\partial t}\right) = \operatorname{curl} \boldsymbol{H} \tag{102}$$

$$\rho = \operatorname{div} \boldsymbol{E} \tag{103}$$

$$-\frac{1}{c}\frac{\partial \boldsymbol{H}}{\partial t} = \operatorname{curl} \boldsymbol{E} \tag{104}$$

$$0 = \operatorname{div} \boldsymbol{H} \tag{105}$$

$$\text{Force per unit charge} = \boldsymbol{E} + \frac{[\boldsymbol{uH}]}{c}\cdot \tag{106}$$

The addition of the equations $0=\operatorname{div} \boldsymbol{H}$ implies the additional postulate that all magnetic fields are traceable to currents without the invocation of any magnetic poles.

The foregoing arguments can hardly be regarded as a proof of the relations (102)–(106). As we have already remarked, the generalization involved in the addition of the term involving $\partial\boldsymbol{E}/\partial t$ in (102) could have been made in an infinite number of other ways as far as serving the purpose of providing for the equation of continuity is concerned. Again, the deduction of (104) as

[21] This is not a foregone conclusion, because even in the sense of the definition of $\boldsymbol{E}$ in terms of the force on a charge, that charge is supposed at rest.

applied to the field at a point in space from the (80) as applied to a complete material circuit is not an entirely rigorous procedure. Similar remarks may be made regarding (106). Moreover, the sense in which the term "force" is considered as applying in equation (106) is to some extent vague when considered in relation to distributions of electricity, for the scheme of equations itself provides no idea of electric mass to combine with acceleration as a criterion for a realization of the actions of the so-called force.

The best that can be said for the differential forms (102)–(104) in the light in which they have thus been obtained is that they are forms which when applied to the large scale phenomena of wire circuits, reproduce the equations (79) and (80). And even this statement is not correct except for steady currents for equation (79) contains no $\partial \boldsymbol{E}/\partial t$ term. It is quite true that the scheme represented by (79), (80), and (81), together with its moulding into a dynamical form as carried out by Maxwell represents a logical scheme which might have corresponded to nature even for varying currents in closed circuits. There are no inconsistencies in the scheme itself. The extension to the general case of moving charges has, however, necessitated the introduction of the $\partial \boldsymbol{E}/\partial t$ term; and, this term having been introduced, it must not be discarded on return to the simpler case of closed circuits unless it, of its own accord, disappears for that case, and this it does not do.

Maxwell was not satisfied with a mere deduction of (104) by the process we have sketched; but, having modified (95) by the introduction of the $\partial \boldsymbol{E}/\partial t$ term, he proceeded to seek an expression for the kinetic and potential energies per unit volume in terms of the electric fields and the current densities, which, by the application of the dynamical equations for continuous media would result in equations (102)–(106), in a manner analogous to that which he had adopted for the case of continuous circuits. His analysis suffered from certain imperfections, and it was later modified by Lorentz, Larmor, and others, with the special additional postulate that the electricity existed in the form of definite entities or electrons. The procedure amounted to choosing an expression for the kinetic and potential energies per unit volume of the medium as a function of the electric field and motion of the electricity, and then by means of the Hamiltonian dynamical principle $\partial\int(T-V)dt=0$ applied to these energy expressions deducing two sets of equations, one of which was the equivalent of (104), and the other the equivalent of (106), but in a more unambiguous form. The form in question stated that the electron moves in such a manner that

$$\iiint\left(\boldsymbol{E}+\frac{[\boldsymbol{uH}]}{c}\right)\rho d\tau = 0 \tag{107}$$

the integral being taken over the electron, and the fields representing the total fields, including the parts produced by the electron itself.

Again, these derivations can hardly be regarded as having accomplished more than the finding of the expressions which it is appropriate to call kinetic and potential energy in order that the customary dynamical procedure shall lead to the equations desired.

As a matter of fact, the form of the force equation symbolized in (107) is one which would be inconsistent with the principle of relativity; for, in the manner which it is presented by the dynamical development, it is of course not restricted to the case where the electron is momentarily at rest in the system of axes to which $\boldsymbol{E}$ and $\boldsymbol{H}$ are referred.

We shall presently return to another form of the expression of the dynamics of the electron, and one which is consistent with the theory of relativity; but for the moment, it will be well to make a digression to discuss a few matters concerned with the significance of (70)–(73), quite apart from the question of their origin.

The Properties and Dynamics of the Electron not Contained in the Circuital Relation.—In the first place it is to be remarked that while equations (70)–(73) may be solved so as to give the fields $\boldsymbol{E}$ and $\boldsymbol{H}$ in terms of ρ and $\rho\boldsymbol{u}$ supposed assigned, they tell us nothing about the motions of the charges themselves. Since the equations are linear, we may add any two solutions and obtain as a result another solution. Thus, if we should find, as is in fact quite easy to do, the solution for the fields of an electron moving in a straight line with velocity v, and add to this the solution for another electron moving towards the first with velocity v, the result would also satisfy the equations. The equations themselves would be perfectly content for the electrons to go on in their motions quite oblivious of each other's presence. It is to the force equation entirely that we must look for an answer to how the motions of the electron influence each other. Had equations (70)–(73) been non-linear, this conclusion would not have held, solutions would not have been additive to give another solution; and, the equations themselves would have had something to say about the motions of the electrons relative to each other.

We frequently hear difficulties raised as to how the electron holds together, in view of the repulsion of its individual parts. Logically there is no difficulty. Equations (70)–(73) are perfectly content to have the various parts of an electron in mutual proximity without any tendency to fly apart. Even the force equation concerns only the motion of the electron as a whole; and it is no quibble to maintain that the ideas applicable to the whole entity are not applicable to its individual parts; for, as a matter of fact, even equation (107) while implying a motion of the electron which secures a zero value for $\iiint(\boldsymbol{E}+[\boldsymbol{uH}]/c)\rho d\tau$, by no means provides for the vector $\boldsymbol{E}+[\boldsymbol{uH}]/c$ being zero at each element of volume.

Solution of the Circuital Relations.—It is possible to solve equations (70)–(73) in such manner as to obtain the fields $\boldsymbol{E}$, $\boldsymbol{H}$, in terms of the charge density ρ, and the product $\rho\boldsymbol{u}$ supposed assigned. The results are:

$$\boldsymbol{E} = -\frac{1}{c}\frac{\partial \boldsymbol{U}}{\partial t} - \text{grad}\,\phi \tag{108}$$

$$\boldsymbol{H} = \text{curl}\,\boldsymbol{U} \tag{109}$$

where

$$4\pi c\boldsymbol{U} = \int\int\int \frac{(\rho \boldsymbol{u})}{r} d\tau \tag{110}$$

$$4\pi\phi = \int\int\int \frac{(\rho)}{r} d\tau \tag{111}$$

and where $\boldsymbol{U}$ and ϕ are still further subjected to the condition that

$$\operatorname{div} \boldsymbol{U} = -\frac{1}{c}\frac{\partial\phi}{\partial t}\cdot \tag{112}$$

The integrals are supposed taken throughout all space, and the parentheses are intended to denote that in finding the values of $\boldsymbol{U}$ and ϕ at time t, at any point (which is momentarily taken as the origin) we must insert in any particular element of volume the values of $\rho\boldsymbol{u}$, or ρ not at the time t, but at the earlier time $t-r/c$.[22]

Equations (70) and (71) of course contain the equation of continuity as a consequence. It can readily be shown that the relation (112) is the analytical guardian of the equation of continuity in preventing us from assigning in (110) and (111), values of $\rho\boldsymbol{u}$ and ρ which would of themselves violate that condition.

Thus, for example, if we were to assign $\boldsymbol{u}=0$, and $\rho=f(r)e^{-\alpha t}$ our assignment would of itself be inconsistent with

$$\operatorname{div} \rho\boldsymbol{u} + \frac{\partial\rho}{\partial t} = 0$$

Equation (112) would prevent such an assignment of ρ and $\rho\boldsymbol{u}$.

The fact that the solutions represent ϕ and $\boldsymbol{U}$ and so $\boldsymbol{E}$ and $\boldsymbol{H}$ as the sum of a number of parts, each determined independently by the separate elements of charge enables us to speak of the field contributed by a charge A as distinct from the field contributed by some other charge B. Had, for example, products of the values of $\rho\boldsymbol{u}$ in the different elements of volume been involved in the expression for the field at a point, it would have been impossible to speak of a part of the field being due to one charge and part to another.

Definitions.—Before proceeding further, it is well that we pause for a moment to consider the significance which is to be attached to quantities like $\boldsymbol{E}$ and $\boldsymbol{H}$ in a scale of analysis which is contemplated in the equations as they are used in atomic physics. The definition of the magnetic field as the force on a unit magnetic pole, and of the electric field as the force on unit charge, are permissible so long as we are dealing with large scale phenomena; but, a unit pole has only existence in a macroscopic sense where a point is identified in a crude way with a region sufficiently large to contain many molecules. Even though, in the case of the charge, we can go to as

[22] We shall not pause to consider the possibilities inherent in the use of the advanced potentials.

small a unit as the electron, at any rate in imagination, for a charge entity in terms of which to define, by the magnitude of its acceleration, the electric field, our procedure becomes illogical when we contemplate a situation in which the field $\boldsymbol{E}$ varies enormously in magnitude and direction over the space comprised by a single electron. The question here involved is nor one of experimental difficulty, but of logical meaninglessness. It is not one purely ideal in its character, moreover; for in one of the most important applications of the equations, that involved in finding the quantity represented by the right-hand side of (75) we encounter a situation in which the quantity $\boldsymbol{E}$ with which we deal, varies to the extent of the whole of its value over the electron. Why is it that we are not led into hopeless error in cases of the above kind? It is because we do not *use* $\boldsymbol{E}$ and $\boldsymbol{H}$ in the sense of the definition in terms of unit charges and unit magnetic poles.

If we examine what we actually do in all problems of atomic dimensions we shall find that the starting point is the *assignment* of the charges and of their motions—the assignment of ρ and $\rho\boldsymbol{u}$. This is followed by the *calculation* of $\boldsymbol{E}$ and $\boldsymbol{H}$ by aid of equations (108)–(109), and the subsequent ultilization of the same $\boldsymbol{E}$ and $\boldsymbol{H}$ for some specific purpose, as, for example, the determination of the integral on the right-hand side of (75) (where only $\boldsymbol{E}$ is required, however).

From the standpoint which is the standpoint of their actual use in nearly all problems of atomic importance, equations (70)–(73) really constitute the definitions of $\boldsymbol{E}$ and $\boldsymbol{H}$ in terms of ρ and $\rho\boldsymbol{u}$ supposed assigned.

But you may well ask what we gain by mere definition. We cannot discover laws of physics by definitions. That is true; but, we can, by well chosen definitions insure useful properties for the quantities defined. Thus, the quantities $\boldsymbol{E}$ and $\boldsymbol{H}$ defined by equations (70)–(73) vary with space and time in free space in accordance with the wave equation. Thus, for example:

$$\nabla^2\boldsymbol{E} - \frac{1}{c^2}\frac{\partial^2\boldsymbol{E}}{\partial t^2} = 0$$

and the wave equation insures for them certain other important properties—such, for example, as the property that in a plane wave the vectors $\boldsymbol{E}$ and $\boldsymbol{H}$ are perpendicular to each other, and lie in the plane of the wave-front. It would be useless to define things like $\boldsymbol{E}$ and $\boldsymbol{H}$ if we could not express in terms of them the behavior of something we can observe. The justification for their definition in electrodynamics lies in their being appropriate quantities in terms of which to express the motions of electrons through such equations as (75), or (78), for example; and, the properties provided for in their definitions provide in turn for properties which we can observe in the motions of the electrons referred to. Now it is quite true that, in the majority of problems of atomic dynamics, the things which happen to electrons do not go on in such manner as to indicate that they have been controlled by the classical force equation. Nevertheless, in formulating the new laws of such a phenomenon as the photoelectric effect, for example, the properties of $\boldsymbol{E}$ and $\boldsymbol{H}$ inherent in their definitions through the electromagnetic equations make them useful props in a scaffolding appropriate to the new theory.

All that we need as a basis in terms of which to define a set of electromagnetic vectors satisfying equations (70)–(73) is a quantity ρ and a quantity $\rho \boldsymbol{u}$ satisfying the equation of continuity

$$\operatorname{div} \rho \boldsymbol{u} + \frac{\partial \rho}{\partial t} = 0$$

As a matter of fact, the density and density times velocity occur in the solutions for ϕ and $\boldsymbol{U}$, and so in those for $\boldsymbol{E}$ and $\boldsymbol{H}$ in the forms they do, not because of any property inherent in themselves, but simply on account of their relative positions in (70)–(73). If we should replace ρu_x, ρu_y, ρu_z, ρ by any four quantities S_x, S_y, S_z, R, satisfying

$$\operatorname{div} \boldsymbol{S} + \frac{\partial R}{\partial t} = 0$$

the solutions would be of exactly the same form as before, with $\boldsymbol{S}$ and R replacing $\rho \boldsymbol{u}$ and ρ. We shall return later to the significance of this aspect in relation to modern atomic theory.

Definitions in Terms of Assigned Entities.—For one to whom the aspect of (70)–(73) as definitions of $\boldsymbol{E}$ and $\boldsymbol{H}$ is too artificial for his mental satisfaction, and who insists upon definitions of $\boldsymbol{E}$ and $\boldsymbol{H}$ in terms of the supposed action of these vectors on something or other, a way is still open, at least to a scale of precision which regards the electron as no more than a point entity. It is only at the expense of considerable complication, however, that this way can be made precise.

The most drastic change from the classical procedure must be made in the matter of the magnetic pole. And, as is often the case, if we follow the plan of what actually happens in nature rather than invent artificial concepts, we gain in the long run. The action of a magnetic field on a pole is an idealization of its action on a magnet, and that is an idealization of the really more simple concept of its action on a moving electron, since the motions of electrons endow the magnet with its magnetic properties. It is then in the spirit of its action on a moving electron that we shall seek to define our magnetic field.

Let us suppose for a moment, that $\boldsymbol{E}$ and $\boldsymbol{H}$ had been defined in a manner which, for the moment, we need not specify, and let us suppose that the motion of an electron in terms of the fields so defined were given by (78), where we shall suppose the terms on the right-hand side other than the first to be absent. Then by putting $\boldsymbol{w}=0$, it would be a fact that

$$E_0 = \frac{m_0}{e} \operatorname*{Lt}_{w=0} \frac{d}{dt} \frac{\boldsymbol{w}}{(1 - w^2/c^2)^{1/2}} \tag{113}$$

and that when $\boldsymbol{w}$ was not zero

$$\frac{[wH_0]}{c} = \frac{m_0}{e}\left\{\left(\frac{d}{dt}\,\frac{w}{(1-w^2/c^2)^{1/2}}\right)_{w\neq 0} - \underset{w=0}{\mathrm{Lt}}\,\frac{d}{dt}\,\frac{w}{(1-w^2/c^2)^{1/2}}\right\} \tag{114}$$

If the direction of w be adjusted so that the quantity on the right-hand side is a maximum, we shall have

$$\frac{H_0}{c} = \frac{m_0}{|ew|}\left\{\left(\frac{d}{dt}\,\frac{w}{(1-w^2/c^2)^{1/2}}\right)_{w\neq 0} - \underset{w=0}{\mathrm{Lt}}\,\frac{d}{dt}\,\frac{w}{(1-w^2/c^2)^{1/2}}\right\} \tag{115}$$

where the w inserted is chosen in the direction to make H_0 a maximum. The direction of H_0 is that consistent with the vector product of H_0 and the velocity w being in the direction of the vector specified as the right-hand side of (114).

Now the facts embodied in the above equations provide us with a means of defining E_0 and H_0. For, suppose, discarding the fact that E_0 and H_0 were supposed defined originally, we take (113) as a definition of E_0; and, after proceeding to the limit when $w=0$ in order to secure uniqueness, suppose we take (115) as the basis of our definition of H_0. That is:

$$\frac{H_0}{c} = \frac{m_0}{e}\underset{=0}{\mathrm{Lt}}\,\frac{\left\{\left(\frac{d}{dt}\,\frac{w}{(1-w^2/c^2)^{1/2}}\right)_{w\neq 0} - \underset{w=0}{\mathrm{Lt}}\,\frac{d}{dt}\,\frac{w}{(1-w^2/c^2)^{1/2}}\right\}}{|w|} \tag{116}$$

Then, the definitions so found will be consistent with the equation

$$e\left(E_0 + \frac{[wH_0]}{c}\right) = \frac{d}{dt}\,\frac{m_0 w}{(1-w^2/c^2)^{1/2}} \tag{117}$$

for all values of w. Of course, they would not *necessitate* that (117) should hold for all values of w.

If it did not turn out that (117) were true for all values of w when E_0 and H_0 were defined in terms of the limiting value of w, our definitions would be definite, at least, apart from further criticisms which I shall presently raise, but they would be useless. It would then be necessary to find in place of

$$\frac{d}{dt}\,\frac{m_0 w}{(1-w^2/c^2)}$$

a quantity W which, when we again made our definitions in a manner analogous to the above, gave us quantities E_0 and H_0, which were consistent with

$$e\left(E_0 + \frac{[wH_0]}{c}\right) = W \tag{118}$$

for all values of w.

Of course, it will be understood that our experimental realization of the above definitions would involve the measurement of such quantities as

$$\underset{w=0}{Lt}\,\frac{d}{dt}\,\frac{w}{(1-w^2/c^2)^{1/2}}$$

and would necessitate a more or less ideal experiment. In the case cited, we should have nothing but an acceleration to measure; but in general, this would not be the case. Thus, to take an example, suppose there existed certain quantities $\boldsymbol{E}_0$ and $\boldsymbol{H}_0$ in terms of which the equation of motion of the electron was

$$e\left(\boldsymbol{E}_0 + \frac{[\boldsymbol{w}\boldsymbol{H}_0]}{c}\right) = a_0\ddot{S} + a_1\dddot{S} + a_2\ddddot{S} + \cdots, \tag{119}$$

where $\ddot{S}$ represents the acceleration, and the a's are functions of the velocity. On putting $\boldsymbol{w}=0$, we obtain

$$e\boldsymbol{E}_0 = \bar{a}_0\ddot{S} + \bar{a}_1\dddot{S} + \bar{a}_2\ddddot{S} + \cdots, \tag{120}$$

where the $\bar{a}$'s are the limiting values of the a's for zero velocity. We see that, in order to complete our definition it would be necessary to specify $\ddot{S}$, $\bar{S}$, . . . , etc., when $\boldsymbol{w}=0$. Moreover, we cannot specify these arbitrarily. For example, we cannot say "consider an electron for which the velocity is zero and all the time derivatives of the motion are zero except $\ddot{S}$," because, although $\ddot{S}$ could adjust itself so that $e\boldsymbol{E}_0=\bar{a}_0\ddot{S}$ at the instant chosen, it would have to be in the process of change in order to make (119) fit the possibly new value of $\boldsymbol{E}_0$ and of $\boldsymbol{H}_0$ and $\boldsymbol{w}$ at the next and subsequent instants. Our only hope, in the case of such a situation as is represented in (120) would be, having made a satisfactory guess at the final form (119), to proceed to its limiting case (120), and then define $\boldsymbol{E}_0$ in terms of the actual values of $\ddot{S}$, $\bar{S}$, etc., as measured for that case. It is, of course, to be remarked that these actual values $\ddot{S}$, $\bar{S}$, $\overline{S}$, . . . , etc., would not be entirely unrelated. For, the specification of all the time derivatives of the motion at an instant determines that motion for all time if the function is analytic for all time; and, in general it determines the motion for some finite time if the function is analytic for a finite time. Now it is manifest that the motion so determined must be such that it agrees continually with (118) or with whatever equation happens to be the correct one; and, the left hand side of this equation involves the way in which the field $\boldsymbol{E}_0$ and $\boldsymbol{H}_0$ change with the time. We shall return later to consideration of some of these questions having to do with the motion of the electron.

Electrodynamics and Restricted Relativity.—In the section dealing with restricted relativity, we have already called attention to the significance of the invariance of the circuital relation under the Lorentzian transformation and have remarked that before it is possible to consider the quantities $\boldsymbol{E}'$, $\boldsymbol{H}'$, defined by equations (20) and (21) as the true fields in the dashed systems, it is necessary to appeal to the definitions of these quantities and ascertain whether, in the sense of these definitions, $\boldsymbol{E}'$ and $\boldsymbol{H}'$ have, in the dashed system, a significance corresponding to that of $\boldsymbol{E}$ and $\boldsymbol{H}$ in the undashed system. Two cases are open for consideration. The first is that applicable to the case where $\boldsymbol{E}$ and $\boldsymbol{H}$ are defined in terms of the motion of an assigned entity and the second is that applicable to the case where equations (70)–(73) constitute really the definitions of $\boldsymbol{E}$ and $\boldsymbol{H}$.

Significance of Relativity when $\boldsymbol{E}$ and $\boldsymbol{H}$ are defined in terms of the motion of an Assigned Entity. It is a matter of simple algebra to show that if $\boldsymbol{E}'$ and $\boldsymbol{H}'$ are related to $\boldsymbol{E}$ and $\boldsymbol{H}$ by equations (20) and (21), then if $k'=(1-w^2/c^2)^{-\frac{1}{2}}$ the four quantities $k'(\boldsymbol{E}'+\boldsymbol{w}'\boldsymbol{H}'/c)$, $k'i(\boldsymbol{E}'\boldsymbol{w}'/c)$ are related to the corresponding four quantities without dashes by the transformation of a 4-vector as specified in equations (53). So far there is no implication as to the meaning of $\boldsymbol{E}'$ and $\boldsymbol{H}'$, other than that defined by (20) and (21). Thus we know that if there be any such equation as (118), then provided that k times the three components of the right-hand side form three components of a 4-vector whose fourth component is $ki(\boldsymbol{W}\boldsymbol{w})/c$, the Lorentzian transformation when applied to this equation will lead to an exactly similar equation in dashed letters, with $\boldsymbol{E}'$ and $\boldsymbol{H}'$ having the significance implied in (20) and (21).[23]

Now since $\boldsymbol{E}_0$ and $\boldsymbol{H}_0$ are defined in terms of limiting cases of (118), the same definitions in terms of the corresponding limiting cases in the dashed system will insure that the quantities which we have called $\boldsymbol{E}_0'$ and $\boldsymbol{H}_0'$ will be the fields in the dashed system in terms of their proper definitions. Thus the correspondence between $\boldsymbol{E}_0$, $\boldsymbol{H}_0$, and $\boldsymbol{E}_0'$, $\boldsymbol{H}_0'$ is provided for, if there is an equation of motion of the type (118) where $\boldsymbol{W}$ has the properties specified above. If there were no such equation which was true for all velocities in the undashed system we might still use such an equation for the limiting cases to define $\boldsymbol{E}_0$ and $\boldsymbol{H}_0$, as pointed out in the text between equations (117) and (118), but not only would the quantities thus defined be useless, but the corresponding equation when used for the purpose of providing corresponding definitions in the dashed system could not lead to quantities the equivalent of $\boldsymbol{E}_0'$ and $\boldsymbol{H}_0'$. If it did, the truth of the equation would be implied for finite velocities since the limiting cases in the dashed system would correspond to finite velocities in the undashed system.

Significance of Relativity when Equations (70)–(73) constitute the Definitions of E and H. Let us speak of the undashed and the dashed systems as S and S' respectively, and let us think of the elements of volume of

[23] Here we note in the first place that a factor k is necessary on each side of (118) to convert the quantities to the first three components of a 4-vector. This factor k may immediately be cancelled out however. In the second place, it will be noted that the fourth component of the right-hand side has to be $ki\ (\boldsymbol{W}.\boldsymbol{w})/c$. The point involved here is simply that we are not permitted to equate any two 4-vectors simply because they are 4-vectors, just as we are not permitted to equate any two ordinary vectors. The vectors must be in the same direction. The fact that the 4th component of the 4-vector must be $ki(\boldsymbol{W}.\boldsymbol{w})/c$ is bound up with a requirement analogous to that of equality of direction in three dimensions. There is no matter of fundamental interest involved here The question is one easily understood directly the mathematical technique is studied, and we shall not pursue it further. Finally, it may be remarked, we have talked of the quantity $\boldsymbol{W}$, rather than the more specific quantity $m_0/e\cdot d/dt\cdot \boldsymbol{w}/(1-w^2/c^2)^{\frac{1}{2}}$ simply for generality. It may be well for the reader to think in the first instance of the more specific quantity referred to.

space in S as marked out with identifiable points at their boundaries, so that those boundaries may be recognized in S'. The elements of volume may be allowed to move about. Let A and A' be observers in S and S'.

Let us suppose that A and A' assign the same number ΔQ to corresponding elements of volume at times t and t' related by

$$t' = \beta\left(t - \frac{v}{c^2}x\right).$$

It is easy to show that corresponding elements of volume $\Delta\tau$ and $\Delta\tau'$ are related by the equation

$$\Delta\tau = \beta\left(1 - \frac{u_x v}{c^2}\right)\Delta\tau'$$

where u_x is the x component of the velocity of the element of volume in S.

If now we *define* ρ and ρ'' by[24]

$$\rho \equiv \frac{\Delta Q}{\Delta\tau}; \quad \rho'' \equiv \frac{\Delta Q}{\Delta\tau'} \tag{121}$$

it follows that

$$\rho'' = \beta\rho\left(1 - \frac{u_x v}{c^2}\right).$$

Thus the quantity ρ'' defined by (121) is the very same quantity as that defined by equation (22).

Suppose now A in S uses equations (70)–(73) or their solutions (108)–(112) to define $\boldsymbol{E}$ and $\boldsymbol{H}$ in accordance with the ideas outlined in the section on definitions. It will of course be true that the quantities $\boldsymbol{E}'$ and $\boldsymbol{H}'$ defined in terms of $\boldsymbol{E}$ and $\boldsymbol{H}$ by (20) and (21) will satisfy the equations obtained from (70)–(73) by dashing all the letters, and with the understanding that $\rho' = \beta\rho(1 - u_x v/c^2)$. If, however, the observer A' in S' defines what to him are the electric and magnetic fields he will arrive at those very fields $\boldsymbol{E}'$ and $\boldsymbol{H}'$ given by (20) and (21); for his charge density ρ'' has been proved equal to ρ', and the solutions of the electromagnetic equations in terms of charge density and velocity are unique insofar as we concern ourselves with the parts related to the charges.

Thus, the reality of the invariance, and the significance of $\boldsymbol{E}'$, $\boldsymbol{H}'$ and ρ' as true electric and magnetic fields and charge density depend ultimately upon the assignment of the same numbers by A and A' to the charges in corresponding elements of volume in the two systems.[25] But we must now

[24] We have used ρ'' instead of ρ' because ρ' has already been used in equation (22) as *defined* by $\rho' = \beta\rho(1 - u_x v/c^2)$, ,and we do not wish to assume that $\Delta Q/\Delta\tau'$ has this value until we have proved it.

[25] It is to be noted that although *invariance* of the charge is demanded, that is not the same thing as demanding *constancy* of the charge. As far as *invariance* was concerned the charge on an electron, for example, might vary with the time and all would be well, so long as A and A' measured equal charges in corresponding volume elements at corresponding times. It is true that the equations themselves provide for *constancy* of the charge as well, however.

inquire by what process A and A' arrive at conclusions as to the magnitudes of charges in corresponding volume elements. Of course they can only arrive at such conclusions by observing something which is observable, and which is calculable in terms of the fields. Thus, suppose having defined our fields as above, we invoke a force equation, for example, of the form (78). The magnitudes of the fields which we shall insert for $\boldsymbol{E}_0$ and $\boldsymbol{H}_0$ will depend upon the magnitudes which we assign to the charges in the elements of volume of space. It must be our business to assign the charges in such a manner that our equation of motion is true.[26] If we cannot make this equation of motion true, our business must be to find some other equation which is invariant under the transformation in the sense that (78) is invariant, and such that it will be possible to assign charges to the volume elements of space which will lead to the calculation of fields which make the equation true.[27] If we succeed in doing this in S, there is no further assumption involved in stating that in S' the corresponding charges will be the same. That they will be the same is a mathematical consequence of the forms of the equations. The dashed and the undashed system of equations are the expression of the same facts provided that the *corresponding charges in the two systems are chosen equal*; and, the process of fitting these equations to the same phenomena will insure that A and A' shall assign the corresponding charges as equal.

In order to crystallize the whole situation, even at the expense of some repetition, it may then be worth while to state the significance of invariance in relation to the electromagnetic equations. Our statement is as follows:

First we must imagine equations (70)–(73) supplemented by some such equation as (78) concerning which all we need say is that it describes the motion of some observable entity, and describes it in an equation which under the Lorentzian transformation assumes the same form in dashed letters, and in particular with $\boldsymbol{E}_0$ and $\boldsymbol{H}_0$ replaced by $\boldsymbol{E}_0'$ and $\boldsymbol{H}_0'$ as defined by, (20)–(21).

Then suppose that observer A uses equations (70)–(73) together with this extra equation to describe the motion of his observable entities. In this process, in order to calculate the fields $\boldsymbol{E}_0$, $\boldsymbol{H}_0$, he will have to assign certain charges to the volume elements of space (although of course, these charges will move about), and his charge densities must be supposed obtained by dividing these charges by the elements of volume which contain them. In terms of the charge densities and velocities assigned to the charges[28] he

[26] The constants m_0 and e in (78) combine to one constant m_0/e. All matters concerned with e being the same as the charges on the other electrons, or of m_0 being expressed in terms of e and of a radius of the electron arc matters perfectly consistent with, but not required by the considerations of invariance.

[27] Of course, there is no reason to expect that the charges so calculated will be grouped into equal units corresponding to electrons.

[28] It must be remembered that the velocity of an element of charge is not a quantity directly observable, and strictly speaking, these velocities will be determined as the numbers necessary to fit the facts, and, incidentlally, to preserve the equation of continuity $\partial\rho/\partial t+\text{div.}\,\rho\boldsymbol{u}=0$, which equation is contained as a consequence of (70)–(73).

will calculate from (70)–(73) the fields $\boldsymbol{E}_0$, $\boldsymbol{H}_0$. He will then substitute the results so obtained in the additional equation to calculate the motions of the observable entity. The agreement of his calculations with the experimental facts is the criterion that he has assigned the correct charges and velocities; and, the fact that it is *possible* to find suitable charges and velocities to do this is the criterion that the theory is correct. Now the position of relativity in this matter is as follows: No extra assumption is involved. It is a mathematical consequence of the forms of the equations that if A' goes through exactly the same procedure to discuss the identical phenomenon which A is examining, he may use exactly the same equations. He need not even dash the letters, but it will of certainty result that if A was successful in calculating the observed motion, A' will be successful, and that moreover, the charges which A' will have to assign will be equal to those assigned by A, to corresponding elements at corresponding times. Of course, the fields which A' calculates will not be the same as those calculated by A. They will be related to them by equations (20) and (21). Moreover, the velocities will be related by (19) and the densities by (22). Of course, moreover, the form of the motion of the observable entity as calculated by A', in his coordinates, will differ from that calculated by A in his coordinates. It will be a motion obtainable from that calculated by A by the Lorentzian transformation. The point is that if the motion calculated by A agrees with A's measurements, the motion calculated by A' will agree with A''s measurements.

Perhaps one of the vital things to observe in all this is that the things which are assigned, the charges and the velocities, are either the same for the two observers as in the case of the charges, or are related by a law immediately predictable from the transformation itself, as in the case of the velocities. Had it been impossible to pass back from the relations between corresponding densities to the invariant charges, had it been necessary in fact, to say that A and A' must *assign* densities related by (22), much of the purity of meaning of the theory of relativity as applied to electrodynamics would have been lost.

The Relativity Dynamics of the Electron. Although the recent developments of the quantum theory have rendered it necessary for us to change our ideas as to the motion of an electron, and to discard the older force equations of electrodynamics as representative of the laws of motion in most atomic problems, nevertheless, the fact that classical laws do hold in limiting cases, of slowly varying forces, causes them still to form an important part in the dynamic story of atomic processes. In particular, the approximate equation

$$\left(\boldsymbol{E}_0 + \frac{[\boldsymbol{w}\boldsymbol{H}_0]}{c}\right)e = m\frac{d}{dt}\,\frac{\boldsymbol{w}}{(1-\epsilon^2)^{1/2}}$$

where $\epsilon = (w_x^2 + w_y^2 + w_z^2)^{\frac{1}{2}}/c$, still plays an interesting role, and it is customary to seek for it a parentage not so much in laws involving concepts of the type represented by (107) as in laws as nearly as possible like those of classical dynamics.

According to the principles of general classical dynamics, the equation of motion of a system may be written in the form of Lagrange's equation

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_r}\right) - \frac{\partial L}{\partial q_r} = 0\,. \tag{122}$$

Here no restriction is placed on L, other than the restriction that it is a function of the coordinates and the generalized velocities, and possibly also of the time. (122) is the analytical equivalent of

$$\delta \int L dt = 0. \tag{123}$$

where the values of δq_r at the time limits are zero. (122) contains as a consequence

$$\frac{d}{dt}\left\{ \sum \dot{q}_r \frac{\partial L}{\partial \dot{q}_r} - L \right\} + \frac{\partial L}{\partial t} = 0\,. \tag{124}$$

If $L = T - V$ and T is a homogeneous quadratic function of the time rates of change of the coordinates, while V is a function of the coordinates and time, it is easy to show that

$$\sum \dot{q}_r \frac{\partial L}{\partial q_r} = 2T$$

and equation (124) becomes

$$\frac{d}{dt}(T + V) + \frac{\partial L}{\partial t} = 0$$

so that if V does not involve the time explicitly

$$\frac{d}{dt}(T + V) = 0$$

and

$$T + V = \text{constant}$$

$T + V$ is of course the energy. Even where L is not of the above form, however, it is still true that if L does not involve t explicitly, $\Sigma \dot{q}_r(\partial L/\partial \dot{q}_r) - L$ is constant with time, so it is convenient to give the name "energy" to this quantity for the general case

$$\text{Energy} = \sum \dot{q}_r \frac{\partial L}{\partial q_r} - L$$

and understand by this the energy, even for cases where L does involve t explicitly. An analytical equivalent of (122) or (123) is

$$\dot{q}_r = \frac{\partial H}{\partial p_r};\ \dot{p}_r = -\frac{\partial H}{\partial p_r} \tag{125}$$

where p_r is defined as

$$p_r \equiv \frac{\partial L}{\partial \dot{q}_r} \tag{126}$$

and H is defined as

$$H \equiv \sum p_r \dot{q}_r - L \tag{127}$$

where, however, the $\dot{q}$'s in H are expressed in terms of the q's and p's through equation (126), so that H is a function of the p's and q's and of t.

It is an analytical consequence of (125) that

$$\frac{dH}{dt} = -\left(\frac{\partial L}{dt}\right)_{p\text{'s and }q\text{'s constant}} \tag{128}$$

Again, if $L = T - V$ and T is a homogeneous quadratic function of the velocities and V a function of coordinates and time, $H = T + V$, and is constant with time if L does not involve t explicitly.

In general, if

$$L = T' - V,$$

where T' is merely a function of velocities, coordinates and time, but not a homogeneous quadratic function, the energy $\Sigma p_r \dot{q}_r - L$ will be something other than $T' + V$, and it will be this something which will be constant with time or equal to $-(\partial L/\partial t)$ according as L does or does not involve t explicitly.

Let us consider the consequences of taking for L,

$$L = \sum_i - m_{0i}c^2(1 - \epsilon_i^2)^{1/2} - e\phi_i + \frac{e}{c}(\boldsymbol{U}_{xi}\boldsymbol{w}_{xi} + \boldsymbol{U}_{yi}\boldsymbol{w}_{yi} + \boldsymbol{U}_{zi}\boldsymbol{w}_{zi}) \tag{129}$$

where

$$c^2\epsilon_i^2 = \boldsymbol{w}_{xi}^2 + \boldsymbol{w}_{yi}^2 + \boldsymbol{w}_{zi}^2$$

where $\boldsymbol{U}_i$ and ϕ_i are the vector and scalar potentials contributed by the charges other than that of the ith electron. The quantity L satisfies the requirements we have imposed on L, the requirements of being a function of the q's, the $\dot{q}$'s and Equation (126) then leads to

$$p_i = \frac{m_{0i}\boldsymbol{w}_i}{(1 - \epsilon_i^2)^{1/2}} + \frac{e}{c}\boldsymbol{U}_i. \tag{130}$$

On making use of this in (127) we find

$$H \equiv \sum_r p_r \dot{q}_r - L = \sum_i \left\{\frac{m_{0i}c^2}{(1 - \epsilon_i^2)^{1/2}} + e\phi_i\right\} \tag{131}$$

so that if ϕ_i and U_i are independent of t explicitly, so that L is independent of t explicitly

$$\sum_i \left\{ \frac{m_{0i}c^2}{(1 - \epsilon_i^2)^{1/2}} + e\phi_i \right\} = \text{constant}.$$

Whether this quantity is constant with time or not, it is, as we have remarked what we call the energy, since it is the quantity which is constant when L is independent of t explicitly.

In the case of a single electron[29] and an external field specified by ϕ_i and U_i, as functions of position and time, ϕ_i involves only the coordinates and time, so that it is appropriate to call $e\phi_i$ the potential energy. It is then appropriate to call $m_{0i}c^2/(1-\epsilon_i^2)^{1/2}$ the kinetic energy T. Or since it is customary to regard T as zero when ϵ is zero, we take

$$T \equiv m_{0i}c^2\left(\frac{1}{(1 - \epsilon^2{}_i)^{1/2}} - 1\right).$$

The addition of the constant term does not, of course, affect the condition

$$\frac{d}{dt}(T + V) = -\frac{\partial L}{dt}.$$

The quantity we have called T' is, in the present case,

$$T' = -m_{0i}c^2(1 - \epsilon_i^2)^{1/2} + \frac{e}{c}\left(U_{xi}w_{xi} + U_{y.}w_{yi} + U_{zi}w_{zi} \right).$$

Whether ϕ_i and U_i are, or are not, independent of t explicitly, it results on applying (122) or (123) to L, with due regard to the proper restricting conditions under which the variation is performed, that

$$\left\{ -\frac{1}{c}\frac{\partial U_i}{\partial t} - \operatorname{grad}\phi_i + \frac{1}{c}[w_i \operatorname{curl} U_i] \right\} e = \frac{d}{dt}\left(\frac{m_{0i}w_i}{(1 - \epsilon_i^2)^{1/2}}\right) \tag{132}$$

so that from (108) and (109)

$$\left(E_{0i} + \frac{[w_i H_{0i}]}{c} \right) e = m_{0i}\frac{d}{dt}\left\{ \frac{w_i}{(1 - \epsilon_i^2)^{1/2}} \right\}. \tag{133}$$

Thus we may regard (129) as the appropriate expression for L in order to give rise to the equations of motion (133) by the usual dynamical procedure. In (132) we should regard U_i and ϕ_i as defined in terms of the charges of

[29] Some little care is necessary in interpreting the exact significance of this matter for the case of more than one electron, since the retarded nature of U and ϕ cause them to involve higher time derivatives of the motion other than the velocity. It is, however, possible to regard U and ϕ as functions of the time explicitly, and then take account of their relation to the electronic motions through the introduction of undetermined multiples in the usual manner when L is subjected to the variational operation. It is not my intention here to do more than call attention to the point as a complete discussion of it would involve more space than is consistent with its importance.

the *external* system by (110) and (111). The quantities $\boldsymbol{E}_{0i}$ and $\boldsymbol{H}_{0i}$ need never be used at all except in the light of a shorthand notation. If they are introduced as a matter of convenience, they must be regarded in the light of being defined by (108) and (109); and, in the light of this definition, their obedience to (70)–(73) is guaranteed. We have to choose our Lagrangian function L in any case; and, having chosen it in terms of ϕ and $\boldsymbol{U}$, the complete story is told in a way which is definite. It may be wrong, but it is definite, and $\boldsymbol{E}_{0i}$ and $\boldsymbol{H}_{0i}$ need never make their appearance at all except in the character of a scaffolding. It is of interest to note that the quantity which we usually call the momentum of the ith electron, viz.: $m_{0i}\boldsymbol{w}_i/(1-\epsilon_i^2)^{1/2}$ is not the same thing as the generalized momentum of the system *associated* with the electron i, the latter quantity is given by (130).

It is perhaps worthwhile to emphasize that the application of (122) or (123) to (129) to obtain (132) or (133) is not to be regarded as a *proof of* (132) or (133), since the choice of (129) rests on its being the suitable function for this purpose. All that is shown is that (132) or (133) is the analytical equivalent of applying (122) or (123) to (129). Indeed, it is probable that most so-called proofs in physics should be regarded more in the light of showing the analytical equivalence of different statements of a law.

We know that, with suitable definitions of the quantities concerned as already discussed, (133) for a single electron is invariant under the restricted relativity transformation. We should expect that this would be the case if Ldt for a single electron were invariant in form under the Lorentzian transformation. It is easy to see that this invariance holds. For, writing $k_i \equiv (1-\epsilon_i^2)^{-1/2}$, dt/k_i is known to be an invariant. Also, $k_{xi}w_{xi}$, $k_{yi}w_{yi}$, $k_{zi}w_{zi}$, $kc\sqrt{-1}$ is a 4-vector. It is the velocity 4-vector. Also $\boldsymbol{U}_{xi}$, $\boldsymbol{U}_{yi}$, U_{zi}, $\phi\sqrt{-1}$ is a 4-vector. Thus $m_{oi}c^2(1-\epsilon_i^2)^{1/2}dt = m_oc^2dt/k_i$ and is invariant. Also $\boldsymbol{U}_{xi}\boldsymbol{w}_{xi}+\boldsymbol{U}_{yi}\boldsymbol{w}_{yi}+\boldsymbol{U}_{zi}\boldsymbol{w}_{zi}-c\phi_i=(\boldsymbol{U}_{xi}k_i\boldsymbol{w}_{xi}+\boldsymbol{U}_{yi}k_i\boldsymbol{w}_{yi}+\boldsymbol{U}_{zi}k_i\,\boldsymbol{w}_{zi}-ck_i\phi_i)dt/k_i$ and is therefore invariant. Hence, Ldt is invariant.

It is, of course, possible to deduce (133) from the Hamiltonian equations (125). For this purpose we need the Hamiltonian function for an electron. This is to be obtained by substituting for the $\dot{q}$'s in (127), their values in terms of the q's and p's from (130). In this way we readily obtain for a single electron

$$H = c\left[\left(p_x - \frac{e}{c}U_x\right)^2 + \left(p_y - \frac{e}{c}U_y\right)^2 + \left(p_z - \frac{e}{c}U_z\right)^2 + m_0^2c^2\right]^{1/2} + e\phi. \quad (134)$$

Intuitive Concepts on Electrodynamics: In the historical developments of electrodynamics, attempts to visualize the processes going on in the field have lead to certain forms of statement of the content of the equations which are applicable only in certain cases, but which have in certain minds become so fundamentally engrained as to assume a priority in respect to the equations themselves. Prominent among these concepts is that of moving lines of magnetic force.[30] It is customary to say that a moving magnetic field produces an electric field perpendicular to the lines of force and to the direction of motion of the lines of force, the magnitude of the field being proportional to the product of the velocity and the magnetic induction. It is easy to invent cases where this statement leads obviously to error. Thus imagine two cylindrical bar magnets with axes in line, separated by a gap, so that the north end of one magnet faces the south end of the other. It would be natural to think of the rotation of this system about its axes as giving rise to a squirrel cage-like motion of the lines of magnetic force; and one who supposed that there was an electric field perpendicular to the lines of force and to the velocity at each point would have to conclude that the electric field was radial, and he would be lead to the unwelcome conclusion that the lines of electric force must converge to a linear charge on the axis of the system between the poles. A direct appeal to the electromagnetic

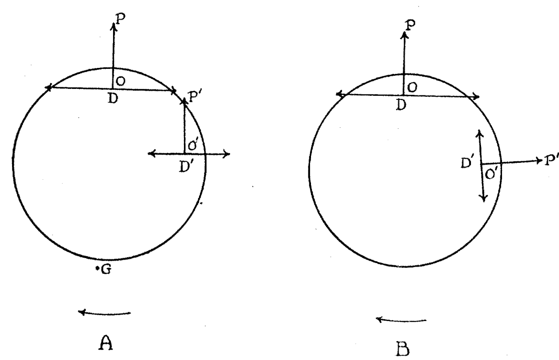

Fig. 3.

equations would, of course, deny the existence of such an electric field. But, while it is easy to construct cases of this kind, and show that the fallacious conclusions would not be predicted by a proper application of the electromagnetic equations, one must face the conclusion that in certain important and simple cases, such as the bodily translatory movement of a magnet in the vicinity of a wire, the law of the rate of cutting of lines of force does lead to a correct calculation of the electromotive force in the wire; and, an exponent of the moving line idea may well claim that there must be something in his statements.

Let us imagine ourselves facing the end of a cylindrical magnet represented by the large circles in Fig. 3, *A* and *B*. Let us further suppose that the magnet is in rotation in the sense indicated by the curved arrow. Let us moreover, think of the individual molecular magnets of which the large magnet is composed. In the first place, it is easy to see that there is an am-

[30] A fairly full discussion of the whole subject concerned with moving lines of force will be found in a paper "Unipolar Induction," by W. F. G. Swann, Phys. Rev., Vol. XV, No. 5, pp. 365–398, May, 1920; also in "Unipolar Induction" by John T. Tate, a section from "Electrodynamics of Moving Media," Bulletin No. 24, of the National Research Council, Part II, pp. 75–95.

biguity in meaning as regards the motion of the lines of force of the resultant field and the motions of the lines of force of the individual molecular magnets. Thus, the resultant lines of force of the magnet at such a point as G are moving from right to left, whereas the contribution to the lines of force at G from the molecular magnet D is such as to correspond to a motion from left to right. Thinking for a moment of the molecular magnets and representing by the radial arrows drawn at D and D', the lines of force coming out of a pole of the molecular magnet, there are two extreme possibilities as regards the meanings to be attached to the motion of the line of force OP for example. The motion may be thought of as corresponding to Fig. 3, A, in which in the motion of the molecular magnet from D to D', the line of force remains parallel to itself. Or, it may be thought of as corresponding to Fig. 3, B, in which, in the motion from D to D', the line of force OP assumes the position there shown. It can be shown that it is by thinking of the motion of lines of force in the former sense that we reproduce the equivalent of the electromagnetic equations. But, in this picture, it is impossible to combine the lines from the different molecular magnets into a resultant, and use the motion of that resultant as a basis for the discussion. On the other hand, in the latter alternative, as represented at Fig. 3, B, it is easy to see that the story predicted by the combined effect of all the molecular magnets would be the same as that predicted by combining the fields into a resultant which rotated rigidly with the magnet as a whole. The attachment of this significance to the motion of lines of force leads, however, to results totally inconsistent with those of the electromagnetic equations.

In the case of the pure translatory motions of magnets without rotation, the former case is the only one which can arise so that in such cases the moving line of force idea fits in with the facts.

The Position of Electrodynamics in Modern Atomic Theory, in Wave-Mechanics, etc. This is not the place to discuss the origin and content of Wave mechanics, and we must be content to dissect out for discussion simply those parts in which electrodynamics figure; treating the remainder as mere statements of methods of procedure. Even here, it will be necessary to confine attention in the present report to a very limited phase of the subject, and omit entirely the electrodynamics theory of Dirac, since to include it would lead us too far into subjects with which it is not the primary purpose of this article to deal.

It will then be recalled that in discussing a problem in wave mechanics, the procedure is first to write down the Hamiltonian function H which would have figured had the intention been to discuss the problem on the lines of classical dynamics. H is a function of, say, n coordinates q_r and n momenta p_r, and of the time t

$$H = H(q_1, q_2 \cdots q_n, p_1, p_2, \cdots t) \tag{135}$$

The next thing is to construct for a quantity ψ a differential equation

$$-\epsilon\frac{\partial\psi}{\partial t} = H\left(q, q \cdots \epsilon\frac{\partial}{\partial q}, \epsilon\frac{\partial}{\partial q_2} \cdots t\right)\psi \tag{136}$$

where $\epsilon = h/2\pi i$, and p_r in (135) is replaced by $\epsilon(\partial/\partial q_r)$ in the operator $H(q_1, q_2 \cdots \epsilon\partial/\partial q_1, \cdots, t)$ which operates on ψ in (136). The next procedure must be to solve this differential equation subject to the boundary conditions appropriate to the problem, and when the solution is found to construct $\bar{\psi}$ the conjugate to ψ. The content of the theory lies in the supposition that if we consider any entity of the system specified by coordinates $q_s, q_{s+1}, \ldots . q_{s+p}$, the probability that these coordinates lie between $q_s, q_{s+1} \cdots q_{s+p}$, and $q_s+dq_s, q_{s+1}+dq_{s+1}, \cdots, q_{s+p}+dq_{s+p}$ is

$$\delta P = dq_s dq_{s+1} \cdots dq_{s+p} \cdots \int \cdots \int \psi\bar{\psi}(dq_1 dq_2 \cdots dq_{s-1})(dq_1 \cdots dq_{s1}) \\ (dq_{s+p+1} \cdots dq_n) \tag{137}$$

Incidentally, it may be remarked that the functions $\psi\bar{\psi}$ are supposed normalized so that $\int \cdots \int\psi\bar{\psi}\, dq_1 dq_2 \cdots dq_{n=1}$.

Everything depends, of course, upon the choice of the Hamiltonian function; and, in the developments of the theories of various phenomena, it cannot be said that wave mechanics has been made to occupy a very consistent position. It will be sufficient, in the first instance to confine ourselves to a case of three dimensions i.e. the case of an electron in a prescribed field of force.

The first problem to engage our attention is naturally that of line spectra. Here the procedure has been to write down the simple Hamiltonian function for the atom in the Newtonian form, with no terms depending upon electrodynamics. The determination of ψ then follows from the solution of (136) subject to the usual conditions of vanishing at infinity,[31] and it corresponds to the sum of a number of terms simple harmonic in t. An expression for $\psi\psi$ is then obtained, and this expression changes with time at any point in a manner representative of the sum of a number of simple harmonic periods corresponding to those of the line spectra of the atom. For the usual case, that where (136) involves the p's as the sum of squares, it turns out as a direct consequence of that equation that the integral of $\psi\bar{\psi}$ throughout all space is independent of the time.

For the usual case, viz. that where (136) involves the p's as the sum of squares, of the form $(p_x^2+p_y^2+p_z^2)/2m$ it turns out as a direct consequence of that equation that[32]

$$\frac{\partial\psi\bar{\psi}}{\partial t} = \frac{h}{4\pi i m}\,\mathrm{div}\,[\bar{\psi}\,\mathrm{grad}\,\psi - \psi\,\mathrm{grad}\,\bar{\psi}] \tag{138}$$

If we define a quantity w by the equation

[31] A condition which places restrictions upon the values of the possible constants of integration of (136), a restriction limiting them to the well-known Eigenwerte.

[32] Strictly speaking what results is that the space integrals of the left and right handsides of this expression are equal. Equation (138) however provides at any rate a sufficient condition for the conservation of the space integrals of $\psi\bar{\psi}$.

$$\psi\bar{\psi}\cdot w = \frac{h}{4\pi i m}\,\text{div}\,[\bar{\psi}\,\text{grad}\,\psi - \psi\,\text{grad}\,\bar{\psi}] \tag{139}$$

equation (138) may be regarded as an equation of continuity with $\psi\bar{\psi}$ functioning in the position of a density and w in the position of a velocity.

The equation of continuity having been provided for, the quantities $\psi\bar{\psi}$ and $\psi\bar{\psi}\cdot w$ provide suitable quantities in terms of which to define a set of vectors $\boldsymbol{E}$ and $\boldsymbol{H}$, satisfying equations of the type (70)–(73), in terms of which the action of radiation on some other atom may be discussed.

By proceeding to the limiting case for a doublet of moment $\boldsymbol{M}$ in electromagnetic theory, we find that ϕ and $\boldsymbol{U}$, and so $\boldsymbol{E}$ and $\boldsymbol{H}$ are completely determined by the assignment of $\boldsymbol{M}$ and $\dot{\boldsymbol{M}}$, so that any quantity $\boldsymbol{M}$ varying with time and assigned to a particular point of space may be used as a basis for defining a set of electromagnetic vectors conforming to the frequencies contained in the time variations in $\boldsymbol{M}$. The procedure for obtaining, for example, the x component of the moment, is to evaluate $\iiint\psi\bar{\psi}x\,dt$ integrated throughout all space, and take this quantity for $\boldsymbol{M}$.

The vectors $\boldsymbol{E}$ and $\boldsymbol{H}$ having been provided for in some such manner as the foregoing, we may then proceed to calculate their values at other points in space where we wish to discuss some phenomenon such as the Compton effect, for example.

The electromagnetic fields defined in this way are, of course, not ends in themselves; and, in discussing such a phenomenon as the photoelectric effect, for example, produced by these fields, the procedure is to write down for an electron in the atom a Hamiltonian function of the form (134) with the quantities ϕ and $\boldsymbol{U}$ as the scalar and vector potentials corresponding to the electromagnetic radiation as above defined. The subsequent procedure is to solve the equation of the type evolved from this Hamiltonian function and the value of $\psi\bar{\psi}$ which results from this solution provides for the history of the ejected photoelectrons by giving the probability that an electron may subsequently be found in any assigned region of space.

A slightly different type of problem is represented by the scattering of x-rays by an electron. Here the customary starting point is the assignment of a Hamiltonian function of the type (134) for the electron, $\boldsymbol{U}$ being now the vector potential of the primary x-rays and ϕ the total potential due to x-rays and atom. The solution of the corresponding equation of the type (136) then follows, subject to appropriate boundary conditions, and the value of $\psi\bar{\psi}$ is calculated at each point of space. The secondary, or scattered radiation, is then calculated as the electromagnetic radiation emitted with $\psi\bar{\psi}$ regarded as the charge density.[33] Strictly speaking, the detection of this radiation would involve the observation of such a phenomenon as the subsequent emission of a photoelectron by the light, and the story in this connection would have to be told in manner analogous to that indicated above for the detection of the light emitted in line spectra.

[33] In the treatment by W. Gordon, (Zeits. f. Physik **40**, 117, 1926), the charge density is taken as a quantity somewhat different from $\psi\bar{\psi}$, but this consideration is not pertinent to the present discussion.

The most unsatisfactory part of the situation concerned in such problems as the foregoing lies in the association of $\psi\bar{\psi}$ with the charge density. Sometimes an effort is made towards a reconciliation of the situation with the logical demands by saying that since $\psi\bar{\psi}$ at a point represents the probability of the presence of an electron in unit volume at that point, the time changes of $\psi\bar{\psi}$ in a volume element represent on the average the time changes of concentration of actual electrons in the volume element. Such a view ignores the fact that, in calculating the fields from the $\psi\bar{\psi}$ distribution, the phase relations of the time oscillations of $\psi\bar{\psi}$ at different places play an important part; and, such phase relations would not necessarily be preserved in the time rates of change of concentration of real electrons if $\psi\bar{\psi}$ represents only the *probability* of the presence of an electron per unit volume. The situation is analogous to that in which a person has two flames of equal intensity, and hopes to get interference between them, forgetful of the fact that the absence of phase relations between the flames would destroy the possibility of occurrence of such phenomena. It is, however, possible to discard entirely the idea that the real electrons have anything to do with the production of the electromagnetic fields associated with the potentials ϕ and $\boldsymbol{U}$, and to regard these quantities as determined entirely by the quantity $\psi\bar{\psi}$. The procedure amounts to actually understanding by the vector and scalar potentials occurring in the original equation for ψ not scalar and vector potentials calculated in terms of the actual electrons, but quantities defined by

$$\left.\begin{aligned} 4\pi\phi &= \int\int\int\frac{(\psi\bar{\psi})d\tau}{r} \\ 4\pi c\boldsymbol{U} &= \int\int\int\frac{(\psi\bar{\psi}\cdot\boldsymbol{w})}{r}d\tau \end{aligned}\right\} \tag{140}$$

with $\boldsymbol{w}$ defined by (139) and with the potentials taken in the retarded sense.

A word may be added concerning problems involving many dimensions. Returning to expression (137) for the probability, and considering the entities concerned as n electrons, the probability ΔP_1 that electron No. 1 lies in the range dq_1 regardless of the positions of the others is

$$\Delta P_1 = dq_1\int\int\int\cdots\int\psi\bar{\psi}dq'$$

where here, by dq_1, we shall understand the product $(dq_1)_x(dq_1)_y(dq_1)_z$ associated with the three coordinates of electron No. 1, and where dq' refers to the product of all the differentials other than $(dq_1)_x$, $(dq_1)_y$, and $(dq_1)_z$. Similarly the probability ΔP_2 that electron No. 2 lies in the range dq_2 regardless of the positions of the others is

$$\Delta P_2 = dq_2\int\int\cdots\int\psi\bar{\psi}dq''.$$

A similar result holds for all of the electrons. Thus, the probability ΔP that some unspecified electron lies in the element $dx\, dy\, dz$ of three dimensional space, regardless of the positions of the other electrons is

$$\Delta P = \sum_{\mu} dq_{\mu} \int\int \int \cdots \int \psi\bar{\psi} dq_{\mu}^{*}$$

where dq_{μ}^{*} refers to the product of all of the differentials other than dq_{μ}, where the sum is taken for all the coordinate elements dq_{μ} arising from the electrons, and where ultimately, all the dq_{μ}'s outside the integrals are replaced by $dx\, dy\, dz$, so that

$$\Delta P = dxdydz \sum f_{\mu}(x \cdot y \cdot z) \tag{141}$$

Here $f_{\mu}(x \cdot y \cdot z)$ is a function of $x \cdot y \cdot z$ in ordinary three dimensional spaces.

Thus, in any many dimensional problem it is this quantity which may function as a charge density for building up vector and scalar potentials to introduce into the Hamiltonian function. With it must be associated a velocity so chosen as to satisfy the equation of continuity for the apparent charge density. As a matter of fact in order to provide consistency with a case in which the electrons associated with the problem may be supposed to have different charges, it is desirable to associate with each component of the sum in (141) a separate velocity w_{μ}, defined by (139) with the $\psi\bar{\psi}$ of that expression replaced by $f_{\mu}(x \cdot y \cdot z)$.

In line with equation (134), the natural Hamiltonian function to assume for a group of electrons would be

$$H = \sum c\left[\left(p_{xi} - \frac{e_i}{c}U_{xi}\right)^2 + \left(p_{yi} - \frac{e_i}{c}U_{yi}\right)^2 + \left(p_{zi} - \frac{e_i}{c}U_{zi}\right)^2 + m_0^2c^2\right]^{1/2} + \sum e_i\phi_i \tag{142}$$

and, writing $\epsilon = h/2\pi i$, the corresponding equation for ψ would be

$$-\epsilon\frac{\partial\psi}{\partial t} = \left\{ \sum_i c\left[\left(\epsilon\frac{\partial}{\partial x_i} - \frac{e_i}{c}U_{xi}\right)^2 + \left(\epsilon\frac{\partial}{\partial y_i} - \frac{e_i}{c}U_{yi}\right)^2 + \left(\epsilon\frac{\partial}{\partial z_i} - \frac{e_i}{c}U_{zi}\right)^2 + m_0^2c^2\right]^{1/2} + \sum e_i\phi_i \right\}\psi \tag{143}$$

where

$$\left.\begin{aligned} 4\pi\phi_i &= \sum \int\int\int \frac{\{f_{\mu}(x \cdot y \cdot z)\}}{r} d\tau \\ 4\pi c\boldsymbol{U}_i &= \sum \int\int\int \frac{\{\boldsymbol{w}_{\mu} f_{\mu}(x \cdot y \cdot z)\}}{r} d\tau \end{aligned}\right\} \tag{144}$$

where the integrals are retarded in the usual manner, and where the summation in the case of ϕ_i and $\boldsymbol{U}_i$ is taken for all the f_{μ} other than that originating on account of the i^{th} electron.

A ψ equation based upon the Hamiltonian function above given has been used by W. Gordon,[34] but without the expression for ϕ_i and U_i in terms of the f_μ as above. The use of a ψ equation based upon this Hamiltonian function has the disadvantage of requiring, for the strict applicability of the equation of continuity, a definition of equivalent charge density and charge velocity other than those which we have given, and as a matter of fact this inconsistency exists even in equations (142)–(144). It may be avoided in a formal manner by holding to our definition of equivalent charge density constituted by each electron and associating with it a velocity w_μ defined by[35]

$$w_\mu \cdot f_\mu(x \cdot y \cdot z) = w\mu \frac{1}{4\pi} \operatorname{grad} \int\int\int \frac{\frac{\partial}{\partial t}\{f_\mu(x \cdot y \cdot z)\} d\tau}{r}$$

which, since the divergence of the right-hand side is $-\partial/\partial t\{f_\mu(x \cdot y \cdot z)\}$, leads to an equation of continuity

$$\operatorname{div}\{f_\mu(x \cdot y \cdot z)\} + \frac{\partial}{\partial t}\{f_\mu(x \cdot y \cdot z)\}. = 0$$

It will be recalled that in our former discussion, the invariance of the electromagnetic equations implied the assignment of equal charges by the observers A and A' to corresponding elements at corresponding times in the two systems. In the present formulation, this becomes translated into the assignment of equal values of $f_\mu(x, y \cdot z)d\tau$; in other words, of equal probability of an electron being found in corresponding volume elements at corresponding times. This condition is a very natural one, and is indeed demanded by the concept of such an entity as an electron, which is a thing primarily supposed as recognizable by one observer if it is recognizable by the other. It is a comparatively easy matter to show that, for the three dimensional case with, of course ϕ_i and U_i given as explicit functions of the time, (143) is invariant under the Lorentzian transformation. For the case of many dimensions, however, this result does not hold.

In conclusion, the reader must again be warned against regarding the foregoing as anything like an adequate treatment of electrodynamics in relation to modern wave mechanics. The excuse for the relatively fragmentary treatment given is that it perhaps clarifies certain of the major points which may have caused difficulty in the minds of those whose interest in the subject has been more or less that of the general reader.

[34] loc. cit.

[35] The integral in this expression is, of course, not retarded