A Theory of Electrons and Protons.

By P. A. M. Dirac, St. John's College, Cambridge.

(Communicated by R. H. Fowler, F.R.S.—Received December 6, 1929.)

§ 1. *Nature of the Negative Energy Difficulty.*

The relativity quantum theory of an electron moving in a given electromagnetic field, although successful in predicting the spin properties of the electron, yet involves one serious difficulty which shows that some fundamental alteration is necessary before we can regard it as an accurate description of nature. This difficulty is connected with the fact that the wave equation, which is of the form

$$\left[\frac{\mathrm{W}}{c}+\frac{e}{c}\mathrm{A}_0+\rho_1\left(\boldsymbol{\sigma},\ \mathbf{p}+\frac{e}{c}\mathbf{A}\right)+\rho_3 mc\right]\psi=0, \tag{1}$$

has, in addition to the wanted solutions for which the kinetic energy of the electron is positive, an equal number of unwanted solutions with negative kinetic energy for the electron, which appear to have no physical meaning. Thus if we take the case of a steady electromagnetic field, equation (1) will admit of periodic solutions of the form

$$\psi = u\, e^{-i\mathrm{E}t/h}, \tag{2}$$

where u is independent of t, representing stationary states, E being the total energy of the state, including the relativity term mc^2. There will then exist solutions (2) with negative values for E as well as those with positive values; in fact, if we take a matrix representation of the operators $\rho_1\sigma_1$, $\rho_1\sigma_2$, $\rho_1\sigma_3$, ρ_3 with the matrix elements all real, then the conjugate complex of any solution of (1) will be a solution of the wave equation obtained from (1) by reversal of the sign of the potentials A, and either the original wave function or its conjugate complex must refer to a negative E.

The difficulty is not a special one connected with the quantum theory of the electron, but is a general one appearing in all relativity theories, also in the classical theory. It arises on account of the fundamental fact that in the relativity Hamiltonian equation of the classical theory, namely,

$$\left(\frac{\mathrm{W}}{c}+\frac{e}{c}\mathrm{A}_0\right)^2-\left(\mathbf{p}+\frac{e}{c}\mathbf{A}\right)^2-m^2c^2=0, \tag{3}$$

there is an ambiguity in the sign of W, or rather $W + eA_0$. Although the operator on the wave function in (1) is linear in W, yet it is, roughly speaking, equivalent to the left-hand side of (3) and the ambiguity in sign persists. The difficulty is not important in the classical theory, since here dynamical variables must always vary continuously, so that there will be a sharp distinction between those solutions of the equations of motion for which $W + eA_0 \geqslant mc^2$ and those for which $W + eA_0 \leqslant -mc^2$, and we may simply ignore the latter.

We cannot, however, get over the difficulty so easily in the quantum theory. It is true that in the case of a steady electromagnetic field we can draw a distinction between those solutions of (1) of the form (2) with E positive and those with E negative and may assert that only the former have a physical meaning (as was actually done when the theory was applied to the determination of the energy levels of the hydrogen atom), but if a perturbation is applied to the system it may cause transitions from one kind of state to the other. In the general case of an arbitrarily varying electromagnetic field we can make no hard-and-fast separation of the solutions of the wave equation into those referring to positive and those to negative kinetic energy. Further, in the accurate quantum theory in which the electromagnetic field also is subjected to quantum laws, transitions can take place in which the energy of the electron changes from a positive to a negative value even in the absence of any external field, the surplus energy, at least $2mc^2$ in amount, being spontaneoulsy emitted in the form of radiation. (The laws of conservation of energy and momentum require at least two light-quanta to be formed simultaneously in such a process.) Thus we cannot ignore the negative-energy states without giving rise to ambiguity in the interpretation of the theory.

Let us examine the wave functions representing states of negative energy a little more closely. If we superpose a number of these wave functions in such a way as to get a wave packet, the motion of this packet will be along a classical trajectory given by the Hamiltonian (3) with $W + eA_0$ negative. Such a trajectory, it is easily seen, is a possible trajectory for an ordinary electron (with positive energy) moving in the electromagnetic field with reversed sign, or for an electron of charge $+e$ (and positive energy) moving in the original electromagnetic field. Thus *an electron with negative energy moves in an external field as though it carries a positive charge.*

This result has led people to suspect a connection between the negative-energy electron and the proton or hydrogen nucleus.* One cannot, however,

* See for example, Weyl, 'Z. f. Phys.,' vol. 56, p. 332 (1929).

simply assert that a negative-energy electron *is* a proton, as that would lead to the following paradoxes :—

(i) A transition of an electron from a state of positive to one of negative energy would be interpreted as a transition of an electron into a proton, which would violate the law of conservation of electric charge.

(ii) Although a negative-energy electron moves in an external field as though it has a positive charge, yet, as one can easily see from a consideration of conservation of momentum, the field it produces must correspond to its having a negative charge, *e.g.*, the negative-energy electron will repel an ordinary positive-energy electron although it is itself attracted by the positive-energy electron.

(iii) A negative-energy electron will have less energy the faster it moves and will have to absorb energy in order to be brought to rest. No particles of this nature have ever been observed.

A closer consideration of the conditions that we should expect to hold in the actual world suggests that the connection between protons and negative-energy electrons should be on a somewhat different basis and this will be found to remove all the above-mentioned difficulties.

§ 2. *Solution of the Negative Energy Difficulty.*

The most stable states for an electron (*i.e.*, the states of lowest energy) are those with negative energy and very high velocity. All the electrons in the world will tend to fall into these states with emission of radiation. The Pauli exclusion principle, however, will come into play and prevent more than one electron going into any one state. Let us assume there are so many electrons in the world that all the most stable states are occupied, or, more accurately, that *all the states of negative energy are occupied except perhaps a few of small velocity*. Any electrons with positive energy will now have very little chance of jumping into negative-energy states and will therefore behave like electrons are observed to behave in the laboratory. We shall have an infinite number of electrons in negative-energy states, and indeed an infinite number per unit volume all over the world, but if their distribution is exactly uniform we should expect them to be completely unobservable. *Only the small departures from exact uniformity, brought about by some of the negative-energy states being unoccupied, can we hope to observe.*

Let us examine the properties of the vacant states or "holes." The

problem is analogous to that of the X-ray levels in an atom with many electrons. According to the usual theory of the X-ray levels, the hole that is formed when one of the inner electrons of the atom is removed is describable as an orbit and is pictured as the orbit of the missing electron before it was removed. This description can be justified by quantum mechanics, provided the orbit is regarded, not in Bohr's sense, but as something representable, apart from spin, by a three-dimensional wave function. Thus the hole or vacancy in a region that is otherwise saturated with electrons is much the same thing as a single electron in a region that is otherwise devoid of them.

In the X-ray case the holes should be counted as things of negative energy, since to make one of them disappear (*i.e.*, to fill it up), one must add to it an ordinary electron of positive energy. Just the contrary holds, however, for the holes in our distribution of negative-energy electrons. These holes will be things of positive energy and will therefore be in this respect like ordinary particles. Further, the motion of one of these holes in an external electromagnetic field will be the same as that of the negative-energy electron that would fill it, and will thus correspond to its possessing a charge $+e$. We are therefore led to the assumption that *the holes in the distribution of negative-energy electrons are the protons*. When an electron of positive energy drops into a hole and fills it up, we have an electron and proton disappearing together with emission of radiation.

A difficulty arises when we consider the field produced by the distribution of negative energy electrons. There is an infinite density of electricity which, according to Maxwell's equation

$$\text{div } \mathbf{E} = -4\pi\rho, \qquad (4)$$

should produce an electric field of infinite divergence. It seems natural, however, to interpret the ρ in Maxwell's equation (4) as the departure from the normal state of electrification of the world, which normal state of electrification, according to the present theory, is the one where every electronic state of negative energy and none of positive energy is occupied. This ρ will then consist of a charge $-e$ arising from each state of positive energy that is occupied, together with a charge $+e$ arising from each state of negative energy that is unoccupied. Thus the field produced by a proton will correspond to its having a charge $+e$.

In this way we can get over the three difficulties mentioned at the end of the preceding section. We require to postulate only one fundamental kind of particle, instead of the two, electron and proton, that were previously necessary.

The mere tendency of all the particles to go into their states of lowest energy results in all the *distinctive* things in nature having positive energy.

Can the present theory account for the great dissymmetry between electrons and protons, which manifests itself through their different masses and the power of protons to combine to form heavier atomic nuclei ? It is evident that the theory gives, to a large extent, symmetry between electrons and protons. We may interchange their rôles and assert that the protons are the real particles and the electrons are merely holes in the distribution of protons of negative energy. The symmetry is not, however, mathematically perfect when one takes interaction between the electrons into account. If one neglects the interaction, the Hamiltonian describing the whole system will be of the form ΣH_a, where H_a is the Hamiltonian or energy of an electron in state a and the summation is taken over all occupied states. This differs only by a constant (*i.e.*, by something independent of which states are occupied) from the sum $\Sigma(-H_a)$ taken over all unoccupied states. Thus we get formally the same dynamical system if we consider the unoccupied states or protons each to contribute a term $-H_a$ to the Hamiltonian. On the other hand, if we take interaction between the electrons into account we get an extra term of the form ΣV_{ab} in the Hamiltonian, the summation being taken over all pairs of occupied states (a, b), and this is not equivalent to any sum taken over pairs of unoccupied states. The interaction would therefore give an essentially different Hamiltonian if we regard the protons as the real particles that occupy states.

The consequences of this dissymmetry are not very easy to calculate on relativistic lines, but we may hope it will lead eventually to an explanation of the different masses of proton and electron. Possibly some more perfect theory of the interaction, based perhaps on Eddington's calculation* of the fine structure constant e^2/hc, is necessary before this result can be obtained.

§ 3. *Application to Scattering.*

As an elementary application of the foregoing ideas we may consider the problem of the scattering of radiation by an electron, free or bound. A scattering process ought, according to theory, to be considered as a double transition process, consisting of first an absorption of a photon with the electron simultaneously jumping to any state, and then an emission with the electron jumping into its final state, or else of first the emission and then the absorption.

* Eddington, 'Roy. Soc. Proc.,' A, vol. 122, p. 358 (1929).

We therefore have to consider altogether three states of the whole system, the *initial state* with an incident photon and the electron in its initial state, an *intermediate state* with either two or no photons in existence and the electron in any state, and the *final state* with the scattered photon and the electron in its final state. The initial and final states of the whole system must have the same total energy, but the intermediate state, which lasts only a very short time, may have a considerably different energy.

The question now arises as to how one is to interpret those scattering processes for which the intermediate state is one of negative energy for the electron. According to previous ideas these intermediate states had no real physical meaning, so it was doubtful whether scattering processes that arise through their agency should be included in the formula for the scattering coefficient. This gave rise to a serious difficulty, since in some important practical cases nearly all the scattering comes from intermediate states with negative energy for the electron.* In fact for a free electron and radiation of low frequency, where the classical formula holds, the whole of the scattering comes from such intermediate states.

According to the theory of the present paper it is absolutely forbidden, by the exclusion principle, for the electron to jump into a state of negative energy, so that the double transition processes with intermediate states of negative energy for the electron must be excluded. We now have, however, another kind of double transition process taking place, namely, that in which first one of the distribution of negative-energy electrons jumps up into the required final state for the electron with absorption (or emission) of a photon, and then the original positive-energy electron drops into the hole formed by the first transition with emission (or absorption) of a photon. Such processes result in a final state of the whole system indistinguishable from the final state with the more direct processes, in which the same electron makes two successive jumps. These new processes just make up for those of the more direct processes that are excluded on account of the intermediate state having negative energy for the electron, since the matrix elements that determine the transition probabilities are just the same in the two cases, though they come into play in the reverse order. In this way the old scattering formulas, in which no intermediate states are excluded, can be justified.

* I am indebted to I. Waller for calling my attention to this difficulty.