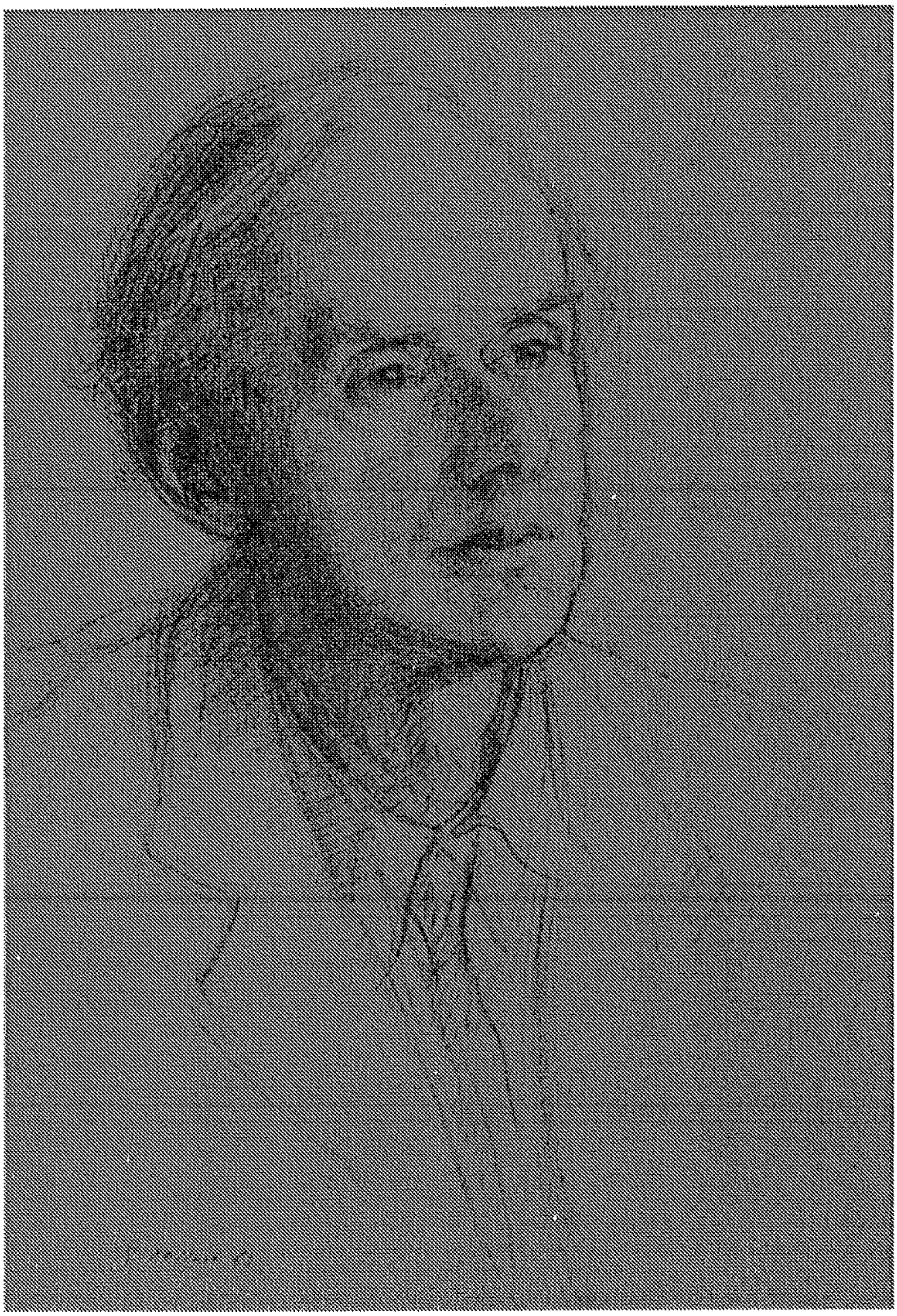
Paul Adrien Maurice Dirac was one of the founders of quantum theory and the author of many of its most important subsequent developments. He is numbered alongside Newton, Maxwell, Einstein and Rutherford as one of the greatest physicists of all time.

This volume contains four lectures celebrating Dirac's life and work, and the text of an address by Stephen Hawking, which were given on 13 November 1995 on the occasion of the dedication of a plaque to him in Westminster Abbey. In the first lecture, Abraham Pais describes from personal knowledge Dirac's character and his approach to his work. In the second lecture, Maurice Jacob explains not only how and why Dirac was led to introduce the concept of antimatter, but also its central role in modern particle physics and cosmology. In the third lecture, David Olive gives an account of Dirac's work on magnetic monopoles and shows how it has had a profound influence in the development of fundamental physics down to the present day. In the fourth lecture, Sir Michael Atiyah explains the widespread significance of the Dirac equation in mathematics, its roots in algebra and its implications for geometry and topology. Together the four lectures in this volume give a unique insight into the relationship between Dirac's character and his scientific achievements.

PAUL DIRAC

THE MAN AND HIS WORK



PAUL DIRAC

THE MAN AND HIS WORK

ABRAHAM PAIS
Rockefeller University, New York

MAURICE JACOB
CERN, Geneva

DAVID I. OLIVE
University of Wales, Swansea

MICHAEL F. ATIYAH
Trinity College, Cambridge

Edited by Peter Goddard

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge CB2 1RP

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, United Kingdom
40 West 20th Street, New York, NY 10011-4211, USA
10 Stamford Road, Oakleigh, Melbourne 3166, Australia

First published 1998

Typeset in 10/13pt Monophoto Times [SE]

A catalogue record for this book is available from the British Library

Library of Congress cataloguing in publication data

Paul Dirac : the man and his work / Abraham Pais . . . [et al.].
p. cm.
ISBN 0 521 58382 9 (hardcover)
1. Dirac, P.A.M. (Paul Adrien Maurice), 1902– . 2. Physicists – Great Britain – Biography. I. Pais, Abraham, 1918– .
QC16.D57P38 1998
530′.092–dc21

ISBN 0 521 58382 9 hardback

Transferred to digital printing 2003

Contents

Preface

Paul Adrien Maurice Dirac was one of the founders of quantum theory and the author of many of its most important subsequent developments. He is numbered alongsidé Newton, Maxwell, Einstein and Rutherford as one of the greatest physicists of all time. He was born in Bristol on 8 August 1902 and died on 20 October 1984 in Tallahassee, Florida. On Monday 13 November 1995, after evensong, a plaque was dedicated in Westminster Abbey commemorating Paul Dirac. The simplicity and almost austere beauty of the plaque's design reflected in some ways the qualities of Dirac's unique intellect.

After graduating from Bristol University with a first class degree in engineering, Dirac stayed on to study mathematics there before obtaining a studentship in 1923 to enable him to undertake research at St John's College, Cambridge. In 1925, he became a Fellow of St John's College. In 1932, he was elected Lucasian Professor of Mathematics in the University. The Lucasian Professorship was once held by Sir Isaac Newton, and the present holder, Stephen Hawking, was present in the Abbey to give an address at the service of commemoration and the text of this address is included in this volume.

Dirac shared the 1933 Nobel Prize for Physics with Erwin Schrödinger. After retirement from the Lucasian chair in 1969, he accepted a research professorship at the Florida State University in Tallahassee. There he continued to work on fundamental physics, frequently returning to St John's College for summer visits, until shortly before his death.

This volume contains four lectures celebrating Dirac's life and work which were given at the Royal Society as a preface to the ceremonies in the Abbey, as well as the text of the address given by Stephen Hawking. The main force behind this commemoration was Richard Dalitz of Oxford University.

In the first lecture, Abraham Pais, of Rockefeller University, New York, and distinguished both for his contributions to fundamental physics and his works on its history, surveys the life and work of Paul Dirac. Although he was famous for his taciturnity and rather retiring nature, Dirac travelled frequently in order to maintain contact with leading physicists in many parts of the world. He visited the Institute for Advanced Study in Princeton several times over the years and there Pais came to know him quite well. His lecture conveys Dirac's singular personal qualities and their relationship to his approach to physics with its emphasis on beauty and simplicity.

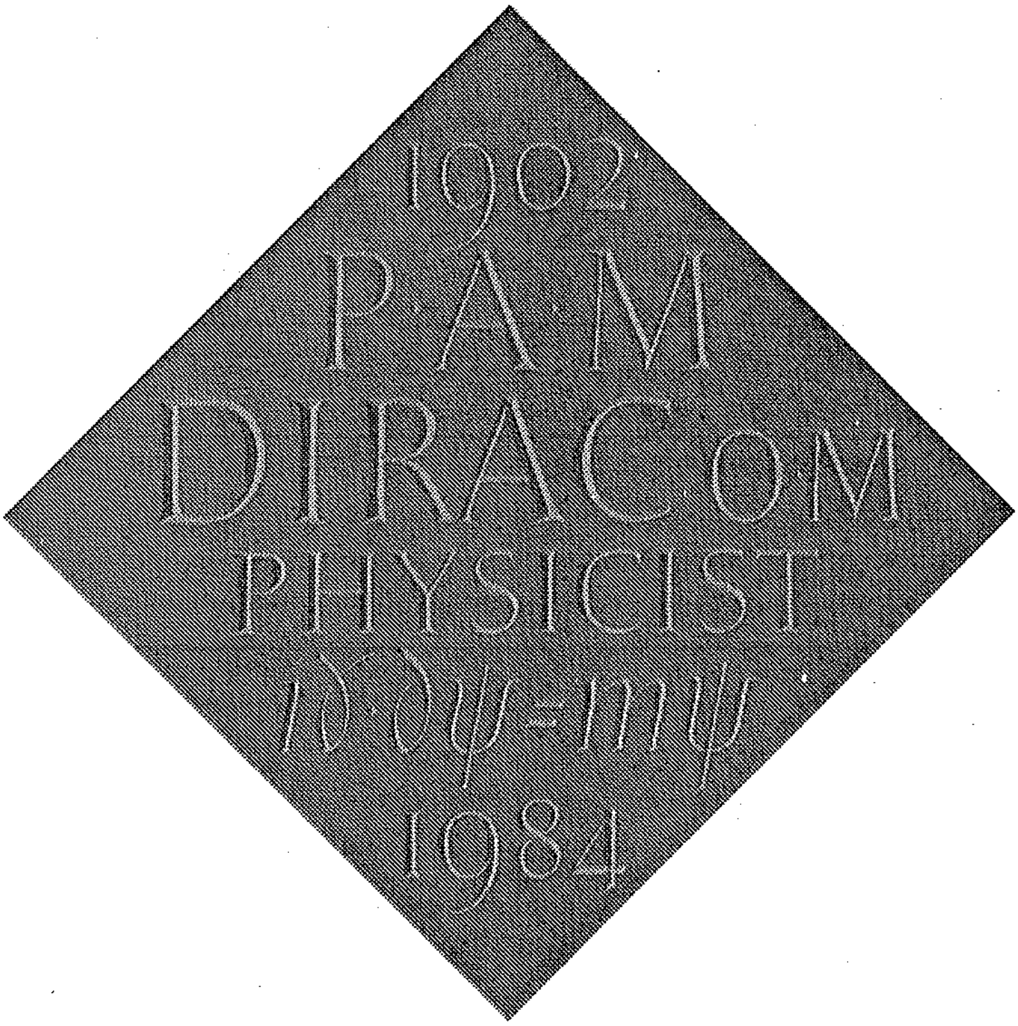
The work of Dirac which has most caught the popular imagination is his prediction of the existence of antimatter, which was described by Dirac's lifelong friend, Werner Heisenberg, as 'the most decisive discovery in connection with the properties or nature of elementary particles ... [It] changed our whole outlook on atomic physics completely'. Seeking to find a theory which reconciled quantum theory with relativity, Dirac found himself led inexorably to the equation which bears his name and which is now engraved on a plaque in the Abbey. As so often when two fundamental ideas are brought together, a third was born, and the existence of antimatter came to be seen as an inevitable consequence of the Dirac equation. Maurice Jacob, of CERN, Geneva, in the second lecture in this volume, explains not only how and why Dirac was led to introduce the concept of antimatter, but also its central role in modern particle physics and cosmology and its importance in practical applications.

Dirac cited mathematical beauty as the ultimate criterion for selecting the way forward in theoretical physics, and he would follow the paths he discerned with great consistency, clarity and courage, often far away from the more well-trodden routes. In the third lecture, David Olive, of the University of Wales Swansea, gives an account of Dirac's work on magnetic monopoles, initiated in 1931, showing that although they have remained undetected experimentally, the ideas Dirac initiated have had a profound influence on the development of fundamental physical theories down to the present day.

The influence of Dirac's ideas has been felt almost as much in mathematics as in physics. In the fourth lecture, Sir Michael Atiyah, speaking as President of the Royal Society and Master of Trinity, explains the significance of the Dirac equation in mathematics, its roots in algebra and its implications for geometry and topology, again taking us forward to very recent developments.

Together the four lectures in this volume make clear how the purity of Dirac's nature and intellect guided his whole work and gave him a penetrating vision, revealing concepts of great depth and prevailing influence in mathematics and physics. They give a unique insight into the relationship between his character and his scientific achievements. Dirac wrote 'it is more important to have beauty in one's equations than to have them fit experiment ... It seems that, if one is working from the point of view of getting beauty in one's equations, and if one has really sound insight, one is on a sure line of progress. If there is not complete agreement between the results of one's work and experiment, one should not allow oneself to be too discouraged, because the discrepancy may be due to minor features ... that will get cleared up with further developments of the theory.' Dirac was writing about Schrödinger but it was his own work that showed just how powerful such an approach could be when adopted by someone with the deepest insight.

PETER GODDARD
MASTER OF ST JOHN'S COLLEGE
CAMBRIDGE



This green slate memorial at Westminster Abbey was designed and cut at the Cardozo Kindersley workshop in Cambridge, UK.

(Photo taken by Michael Manni)

Dirac Memorial Address

STEPHEN HAWKING

Paul Adrien Maurice Dirac (my speech synthesizer isn't very good with his name) was born in Bristol in 1902, to a Swiss father, and an English mother. He went on to become the Lucasian Professor at Cambridge, and to win a Nobel Prize, but was never well known to the public. His death in 1984 drew a short obituary in the *Times*, but otherwise it went almost unnoticed. It has taken 11 years for the nation to recognize that he was probably the greatest British theoretical physicist since Newton, and belatedly to erect a plaque to him in Westminster Abbey. It is my task to explain why. That is, why he was so great, not why it took so long.

In the early years of this century the way we picture the world, and our view of reality itself, were completely transformed by two discoveries: the Theory of Relativity and Quantum Mechanics. Dirac played a major role in quantum theory, and his efforts to make it compatible with relativity turned up new and unexpected phenomena.

Dirac was a research student at St John's College, Cambridge, when Werner Heisenberg visited his supervisor, R. H. Fowler, in the summer of 1925. Heisenberg told Fowler about his ideas on what he called 'matrix mechanics' and sent him a proof copy of his paper on the subject. Fowler passed this to Dirac, who recognized a striking similarity with objects called Poisson brackets in classical mechanics. This led him to write a remarkable paper in which he formulated general rules for the quantum mechanics of any system. These rules incorporated the ideas of both Heisenberg and Schrödinger and showed they were equivalent. Of the three founders of modern quantum mechanics, Heisenberg and Schrödinger can claim to have caught the first glimpses of the theory. But it was Dirac who put them together and revealed the whole picture.

For that alone he would be worthy of a memorial in Westminster Abbey. But he went on, working on how to combine the Special Theory of Relativity with Quantum Theory. In 1928 he discovered what he called the relativistic equation for the electron, but everyone else calls the Dirac equation. As Dirac himself said, this equation governs most of physics and the whole of chemistry. If Dirac had patented his equation, like some people are now patenting human genes, he would have become one of the richest men in the world. Every television set or computer would have paid him royalties.

Dirac made a number of other important contributions to physics, but I won't go into them now. Dirac saw things in very simple and clear terms, and wasn't always able to understand why other people didn't see things similarly. This led to a whole host of Dirac stories. I won't repeat other people's stories, but would like to tell one of my own.

I was a member of the same department as Dirac from 1962 to 1969, but I never saw him. That was because Dirac belonged to the old school who didn't believe in these new-fangled departments of pure and applied mathematics, but worked in their college rooms. And I was working on classical general relativity and not quantum theory at that time, so I didn't go to his lectures. It was not until

1975 that I met him in Rome. I had just been awarded a gold medal by the Pope. Dirac told me that he had nominated someone else for the medal, but had then decided I was better and told the Pontifical Academy so. That was why they had given it to me.

After that I saw Dirac almost every year until his death, when he came back from the University of Florida where he had retired and visited Cambridge in the summer. He never said much, in contrast to his wife, who was Hungarian, and a great character. It was said that this silence was a result of his childhood, when his father would allow him only to speak perfect French at meal times. That may be true, but I suspect he would have been silent even without that. But when he did speak, it was all the more worth hearing.

Dirac has done more than anyone this century, with the exception of Einstein, to advance physics and change our picture of the universe. He is surely worthy of the memorial in Westminster Abbey. It is just a scandal that it has taken so long.

1 Paul Dirac: aspects of his life and work[1]

ABRAHAM PAIS
Rockefeller University, New York

'Of all physicists, Dirac has the purest soul.'
Niels Bohr

In the year 1902, the literary world witnessed the death of Zola, the birth of John Steinbeck, and the first publications of *The Hound of the Baskervilles*, *The Immoralist*, *Three Sisters*, and *The Varieties of Religious Experience*. Monet painted *Waterloo Bridge*, and Elgar composed *Pomp and Circumstance*, Caruso made his first phonograph recording and the Irish Channel was crossed for the first time by balloon. In the world of science, Heaviside postulated the Heaviside layer, Rutherford and Soddy published their transformation theory of radioactive elements, Einstein started working as a clerk in the patent office in Berne, and, on August 8, Paul Adrien Maurice Dirac was born in Bristol, one of the children of Charles

Dirac (1866–1936), a native of Monthey in the Swiss canton of Valais, and Florence Holten (1878–1941), daughter of a British sea captain. There was also a brother two years older, Reginald, whose life ended in suicide, in 1924, and Beatrice, a sister four years younger. About his father Dirac has recalled:

> My father made the rule that I should only talk to him in French. He thought it would be good for me to learn French in that way. Since I found that I couldn't express myself in French, it was better for me to stay silent than to talk in English. So I became very silent at that time – that started very early.[2]

The first edition of Dirac's book, *The Principles of Quantum Mechanics*, has stood on my shelves since my graduate days in Holland. Learning from it the beauty and power of that compact little Dirac equation was a thrill I shall never forget. Years later, in January 1946, I first met Dirac and his wife on a brief visit to their home at 7 Cavendish Avenue in Cambridge. I saw much more of him in the autumn of that year when we met at the Institute for Advanced Study in Princeton. He had spent the academic year 1934–5 there, and also, during my own time at the Institute, the fall term of 1946, and academic years 1947–8, 1958–9, and 1962–3. In the course of all these visits to Princeton I came to know Dirac quite well. A friendship developed. In the course of joint talks and walks and wood chopping expeditions, I developed a good grasp of his views on physics. I also met him elsewhere, especially in Tallahassee where, in 1972, at age 70, he had started a new career: Professor of Physics at the Florida State University.

I shall presently tell of those encounters with Dirac, and of my impressions of his personality. First, however, I should like to speak of his career prior to the time of my personal contacts with him.

Young Paul first attended the Bishop Road primary school, then, at age 12, the secondary school at the Merchant Venturer's

Technical College, both in Bristol, where his father taught French. Much later he has recalled that

> [This] was an excellent school for science and modern languages. There was no Latin or Greek, something of which I was rather glad, because I did not appreciate the value of old cultures . . . I played soccer and cricket . . . and never had much success. But all through my schooldays, my interest in science was encouraged and stimulated.[3]

At the suggestion of his father, Dirac started in 1918 to study at the electrical engineering department of the University of Bristol, from which he graduated with first-class honours in 1921. Forty years later he wrote:

> I would like to try to explain the effect of this engineering training on me. I did not make any further use of the detailed applications of this work, but it did change my whole outlook to a very large extent. Previously, I was interested only in exact equations. Well, the engineering training which I received did teach me to tolerate approximations, and I was able to see that even theories based on approximations could sometimes have a considerable amount of beauty in them . . . I think that if I had not had this engineering training, I should not have had any success with the kind of work that I did later on . . . I continued in my later work to use mostly the nonrigorous mathematics of the engineers, and I think that you will find that most of my later writings do involve nonrigorous mathematics . . . The pure mathematician who wants to set up all of his work with absolute accuracy is not likely to get very far in physics.[4, 5]

During those years as an engineering student,

> A wonderful thing happened. Relativity burst on the world . . . It is easy to see the reason for this tremendous impact. We had just been living through a terrible and very serious war . . . Everyone

> wanted to forget it. And then relativity came along . . . It was an escape from the war.
>
> Previously, as a schoolboy I had been much interested in the relations of space and time. I had thought about them a great deal, and it had become apparent to me that time was very much like another dimension, and the possibility had occurred to me that perhaps there was some connection between space and time, and that we ought to consider them from a general four-dimensional point of view. However, at that time the only geometry that I knew was Euclidean geometry.[4]

In 1921, Dirac looked without success for an engineering job. Then, to his luck, he was offered free tuition for two years to study mathematics at the University of Bristol.

Those years conclude what one may call the prelude to Dirac's scientific career.

In the autumn of 1923, Dirac enrolled at Cambridge with a maintenance grant from the Department of Scientific and Industrial Research. Nine years later he would succeed Joseph Larmor (1857–1942) to the Lucasian Chair of Mathematics, once held by Newton.[6] It was Ralph Fowler (1889–1944) who, in Cambridge, introduced Dirac to the old quantum theory, and it was from him that he first learned of the atom of Rutherford, Bohr, and Sommerfeld.

Dirac first met Bohr in May 1925 when the latter gave a talk in Cambridge on the fundamental problems and difficulties of the quantum theory. Of that occasion Dirac said later:

> People were pretty well spellbound by what Bohr said . . . While I was very much impressed by [him], his arguments were mainly of a qualitative nature, and I was not able to really pinpoint the facts behind them. What I wanted was statements which could be expressed in terms of equations, and Bohr's work very seldom

> provided such statements. I am really not sure how much my later work was influenced by these lectures of Bohr's . . . He certainly did not have a direct influence, because he did not stimulate one to think of new equations.[4]

In July 1925 Dirac first met Heisenberg, also in Cambridge. In that month, Heisenberg's first paper on quantum mechanics had come out.

> I learned about this theory of Heisenberg in September, and it was very difficult for me to appreciate it at first. It took two weeks; then I suddenly realized that the noncommutation was actually the most important idea that was introduced by Heisenberg.[7]

The result was Dirac's first paper on quantum mechanics.[8] Prior to that time he had already published seven respectable papers which had not caused any particular response. Number eight caused a stir, however. It contained the relation $pq - qp = h/2\pi i$, independently derived shortly before by Born and Jordan. The respective authors were unaware of one another's results. Born has described his reaction upon receiving Dirac's paper:

> This was – I remember well – one of the greatest surprises of my scientific life. For the name Dirac was completely unknown to me, the author appeared to be a youngster, yet everything was perfect in its way and admirable.[9]

In those days, Dirac invented several notations which are now part of our language: q-numbers, where 'q stands for quantum or maybe queer'; c-numbers, where 'c stands for classical or maybe commuting.'[4] He has described his work habits in those years: 'Intense thinking about those problems during the week and relaxing on Sunday, going for a walk in the country alone.'[4] Dirac was forever much attracted by the beauty of nature, particularly of mountain areas. He liked to climb mountains, for which he practiced by climbing trees on the Gog–Magog hills outside

Cambridge, even then wearing his perennial dark suit. He avoided technical climbs but nevertheless ascended impressive peaks, in the Rockies, the Alps, and the Caucasus. In 1936, accompanied by Igor Tamm (1895–1971), he managed to reach the 5640 m high top of the Elbruz, Europe's highest mountain, but collapsed at a high altitude, where he had to rest for 24 hours before completing the descent.[10]

In May 1926, Dirac received his Ph.D. on a thesis entitled 'Quantum Mechanics.'[11] Meanwhile Schrödinger's papers on wave mechanics had appeared, to which Dirac reacted with initial hostility, then with enthusiasm. He quickly applied the theory to systems of identical particles.[12] At almost the same time, that problem also attracted Heisenberg,[13] whose main focus, on a few particle systems, resulted in his theory of the helium atom.[14] Dirac's paper[12] (August 1926), on the other hand, will be remembered as the first in which quantum mechanics is brought to bear on statistical mechanics. Recall that the earliest work on quantum statistics, by Bose and by Einstein, predates quantum mechanics. Also, Fermi's introduction of the exclusion principle in statistical problems, though published[15] after the arrival of quantum mechanics, is still executed in the context of the 'old' quantum theory.[16] All these contributions were given their quantum mechanical underpinnings by Dirac, who was, in fact, the first to give the correct justification of Planck's law, which started it all: 'Symmetrical eigenfunctions . . . give just the Einstein–Bose statistical mechanics . . . (which) leads to Planck's law of black-body radiation.'[12]

It is edifying to remember that it took some time before it was sorted out when Bose–Einstein and Fermi–Dirac statistics respectively apply. Dirac in August 1926:

> The solution with anti-symmetric eigenfunctions (F.D. statistics) . . . is probably the correct one for gas molecules, since it

> is known to be the correct one for electrons in an atom, and one would expect molecules to resemble electrons more closely than light-quanta.[12]

Other great men were not at once clear either about this issue, Einstein, Fermi, Heisenberg, and Pauli among them.[16]

Having obtained his doctorate, Dirac was free to travel and, in September 1926, he went to Copenhagen. 'I admired Bohr very much. We had long talks together, long talks in which Bohr did practically all the talking.'[4] It was there that he worked out the theory of canonical transformations in quantum mechanics, since known as the transformation theory.[17] 'I think that is the piece of work which has most pleased me of all the works that I've done in my life . . . The transformation theory (became) my darling.'[7] In this paper, Dirac introduced an important tool of modern physics, the δ-function, about which he remarked right away:

> Strictly, of course, $\delta(x)$ is not a proper function of x, but can be regarded as the limit of a certain sequence of functions. All the same, one can use $\delta(x)$ as though it were a proper function for practically all the purposes of quantum mechanics without getting incorrect results.[18]

Dirac's stay in Copenhagen – lasting till February 1927 – is also highly memorable, because it was there that he completed the first[19] of two papers in which he laid the foundations of quantum electrodynamics.The sequel[20] was written in Goettingen, the next important stop on his journey.

Preceding these two papers, Dirac had already given[12] a theory of induced radiative transitions by treating atoms quantum mechanically but still considering the Maxwell field as a classical system.[21] However, 'one cannot take spontaneous emission into account without a more elaborate theory.'[12] Here, Dirac echoed Einstein who, already in 1917, still the days of the old quantum theory, had stressed that spontaneous emission 'make[s] it almost inevitable to formulate a truly quantized theory of radiation.'[22] In his Copenhagen paper,[19] Dirac did just that. He proceeded to quantize the electromagnetic field, thereby giving the first rational description of light quanta, and then derived from first principles Einstein's phenomenological coefficient of spontaneous emission.[23]

The theory was not yet complete, however: 'Radiative processes . . . in which more than one light quantum take(s) part simultaneously are not allowed on the present theory.'[19] How young quantum mechanics still was. Early in 1927, Dirac did not yet know that these processes are perfectly well included in his theory. All one had to do was extend perturbation theory from first order (used by him in the treatment of spontaneous emission) to second order. So, in his Goettingen paper,[20] he developed[24] second order perturbation theory, which enabled him to give the quantum theory of dispersion.[25] He further noted[26] that the theory could now also be applied to the Compton effect, a subject that had interested him earlier.[27]

In Goettingen Dirac met Robert Oppenheimer (1904–67), who lived in the same pension and with whom he became close friends. Dirac found the catholic interests of Oppenheimer, who spent much time reading Dante in the original, very difficult to understand. It is said that Dirac once asked him: 'How can you do both physics and poetry? In physics we try to explain in simple terms something that nobody knew before. In poetry it is the exact opposite.'

In the year 1927, of which I speak, Dirac was elected Fellow of St John's College in Cambridge and began lecturing on quantum mechanics. In 1929 he was nominated Praelector in mathematics and physics, a post with only nominal duties. In 1930 he was elected Fellow of the Royal Society. As of September 30, 1932, he became the Lucasian Professor, a post he was to hold until 1969. Out of his lectures to students grew his book on quantum mechanics, the first edition of which appeared in 1930. I may note here that in all he published about 200 papers.

Dirac devoted only a small part of his duties to teaching and almost none to administration. He preferred to work by himself and created no school. It has been written of him that he was one of the few scientists who could work even on a deserted island.[28] While it lay therefore not in his nature to seek out research students, he nevertheless delivered a fair number of Ph.Ds.[29]

When Dirac wrote an article or gave a lecture he considered it unnecessary to change his carefully chosen phrases. When somebody in the audience asked him to explain a point he had not understood, Dirac would repeat exactly what he had said before, using the very same words.[30] Be that as it may, his style of lecturing was admirable, as I have been privileged to notice frequently. Some of his students have put it well:

> The delivery was always exceptionally clear and one was carried along in the unfolding of an argument which seemed as majestic and inevitable as the development of a Bach fugue.[31]

Nevertheless I tend to agree with Sir Nevill Mott (1905–96), who has said:

> I think I have to say his influence was not very great as a teacher . . . He never would advise a student to examine the experimental evidence and see what it means . . . He would never, between his great discoveries, do any sort of bread-and-butter problem. He would not be interested at all.[32]

I return to the year 1927, when I left Dirac in Goettingen. From there he went to Leiden and concluded his travels of that year by attending the Solvay conference in Brussels (in October), where he met Einstein for the first time. From discussions with Dirac, I know that he admired Einstein. The respect was mutual ('. . . Dirac, to whom, in my opinion, we owe the most logically perfect presentation of (quantum mechanics)'[33]). Yet, the contact between the two men remained minimal, largely, I would think, because it was not in Dirac's personality to seek father figures.

That 1927 Solvay conference marks the beginning of the well-known debate between Bohr and Einstein on the interpretation of quantum mechanics. Fifty years later Dirac said: 'This problem of getting the interpretation proved to be rather more difficult than just working out the equations.'[7] As time went by he expressed reservations not only regarding quantum field theory but also, though less strongly, in relation to ordinary quantum mechanics,[34,35] but never more clearly than in 1979 when he and I were both in Jerusalem to attend the Einstein centennial celebrations:

> In this discussion at the Solvay conference [of 1927] between Einstein and Bohr, I did not take much part. I listened to their arguments, but I did not join in them, essentially because I was not very much interested. I was more interested in getting the correct equations. It seemed to me that the foundation of the work of a mathematical physicist is to get the correct equations, that the interpretation of those equations was only of secondary importance . . . It seems clear that the present quantum mechanics is not in its final form . . . I think it is very likely, or at any rate quite possible, that in the long run Einstein will turn out to be correct, even though for the time being physicists have to accept the Bohr probability interpretation, especially if they have examinations in front of them.[36]

Later I shall comment further on Dirac's position.

Dirac has recalled a conversation with Bohr during the 1927 Solvay conference. Bohr: 'What are you working on?' Dirac: 'I'm trying to get a relativistic theory of the electron.' Bohr: 'But Klein has already solved that problem.'[4]

Dirac disagreed.

By the time of the 1927 Solvay conference, a relativistic wave equation was already known: the scalar wave equation, stated independently by at least six authors,[37] Klein and Schrödinger among them. One could not, it seemed, associate a positive definite probability density with that equation, however. That Dirac did not like at all, since the existence of such a density was (and is) central to his transformation theory. 'The transformation theory had become my darling. I was not interested in considering any theory which would not fit in with my darling . . . I just couldn't face giving up the transformation theory.'[7] That is why, as said, Dirac disagreed with Bohr. Accordingly, he began his own search for a relativistic wave equation that does have an associated positive probability density. Not only did he find it but, in the course of doing so, he also discovered the relativistic quantum mechanical treatment of spin.

That was a major novelty. In May 1927, Pauli had proposed[38] that the electron satisfy a two-component wave equation which does contain the electron spin, explicitly coupled to the electron's orbital angular momentum. Nothing determined the strength of that coupling, the 'Thomas factor,' which had to be inserted by hand 'without further justification.' This flaw, Pauli noted, was due to the fact that his equation did not satisfy the requirements of relativity. The theory was, in his words, provisional and approximate.

In his equation, Pauli described the spin by 2 × 2 matrices, since known as the Pauli matrices. It appears that Dirac had discovered these independently: 'I believe I got these (matrices) independently of Pauli, and possibly Pauli also got them independently from me.'[4] Always in quest of a relativistic wave equation with positive probability density, Dirac continued playing[39] with the spin matrices. 'It took me quite a while . . . before I suddenly realized that there was no need to stick to quantities . . . with just two rows and columns. Why not go to four rows and columns.'[4] Quite a while, actually, was only a few weeks. Toward the end of his life, Dirac reminisced: 'In retrospect, it seems strange that one can be so much held up over such an elementary point (!)'[40]

Thus, early in 1928, was born the Dirac equation[41,42] with the positive density its author had so fervently desired. To his great surprise, he had stumbled on much more, however.

> It was found that this equation gave the particle a spin of half a quantum. And also gave it a magnetic moment. It gave just the properties that one needed for an electron. That was really an unexpected bonus for me, completely unexpected.[7]

Spin was a necessary consequence, the magnetic moment and the Sommerfeld fine structure formula came out right, the Thomas factor appeared automatically, and for kinetic energies small compared to mc^2 (m = electron mass) all the results of the nonrelativistic Schrödinger theory were recovered. Dirac had played hard and played well. His discovery ('once you got the right road it jumps at you without any effort'[43]), ranking as it does among the highest achievements of twentieth century science, is all the more remarkable since it was made in pursuit of what eventually turned out to be a side issue, positive probabilities.[44]

Along with its spectacular successes, the Dirac equation was, for a few years, also a source of great trouble, however.

Pauli's wave functions have two components, corresponding to the options spin up and spin down. But Dirac's wave functions had four. The question: Why four? led to monumental confusion about which, in the 1960s, Heisenberg recalled: 'Up till that time [1928], I had the impression that, in quantum theory, we had come back into the harbor, into the port. Dirac's paper threw us out into the sea again.'[45]

From the outset,[41] Dirac had correctly diagnosed the cause for this doubling of the number of components. There are two with positive, two with negative energies, each pair with spin up/down. What to do with the negative energy solutions?

> One gets over the difficulty on the classical theory by arbitrarily excluding those solutions that have a negative energy. One cannot do this in the quantum theory, since, in general, a perturbation will cause transitions from states with E positive to states with E negative.[41]

He went on to speculate that negative energy solutions may be associated with particles whose charge is opposite to that of the electron. In that regard, Dirac did not yet know as clearly what he was talking about as he would one and a half years later. This undeveloped idea led him to take the problem lightly, initially: 'Half of the solutions must be rejected as referring to the charge $+e$ of the electron.'[41] In a talk given in Leipzig, in June 1928, he no longer spoke of rejection, however. Transitions to negative energy states simply could not be ignored. 'Consequently, the present theory is an approximation.'[46]

While in Leipzig, Dirac, of course, visited Heisenberg (recently appointed there), who must have been well aware of these difficulties. In May, he had written to Pauli: 'In order not to be forever irritated with Dirac, I have done something else for a change,'[47] the something else being his quantum theory of ferromagnetism. Dirac and Heisenberg discussed several aspects of the new theory.[48] Shortly thereafter, Heisenberg wrote again to Pauli: 'The saddest chapter of modern physics is and remains the Dirac theory,'[49] mentioned some of his own work, which demonstrated the difficulties, and added that the magnetic electron had made Jordan *trübsinnig* (melancholic). At about the same time, Dirac, not feeling so good either, wrote to Oskar Klein: 'I have not met with any success in my attempts to solve the $\pm e$ difficulty.

Heisenberg (whom I met in Leipzig) thinks the problem will not be solved until one has a theory of the proton and electron together.'[50]

Early in 1929, both Dirac and Heisenberg made their first trip to the United States, Dirac lecturing at the University of Wisconsin, Heisenberg at the University of Chicago. In August of that year, the two men boarded the steamer *Shinyo Maru* together in San Francisco, stopped over in Hawaii,[51] then went on to Japan, where they both lectured in Tokyo and Kyoto. I was curious whether they had discussed the problematics of the Dirac equation during their trip, so I asked Dirac. He replied:

> In 1929, Heisenberg and I crossed the Pacific and spent some time in Japan together. But we did not have any technical discussions together. We both just wanted a holiday and to get away from physics. We had no discussions of physics, except when we gave lectures in Japan and each of us attended the lectures of the other. I do not remember what was said on these occasions, but I believe there was essential agreement between us.[52]

Heisenberg has told a story of that trip which gives a rare glimpse of Dirac's attitudes towards the opposite sex:

> We were on the steamer from America to Japan, and I liked to take part in the social life on the steamer and, so, for instance, I took part in the dances in the evening. Paul, somehow, didn't like that too much but he would sit in a chair and look at the dances. Once I came back from a dance and took the chair beside him and he asked me, 'Heisenberg, why do you dance?' I said, 'Well, when there are nice girls it is a pleasure to dance.' He thought for a long time about it, and after about five minutes he said, 'Heisenberg, how do you know *beforehand* that the girls are nice?'[53]

In the meantime, Weyl had made[54] a new suggestion regarding the extra two components: 'It is plausible to anticipate that, of the two pairs of components of the Dirac quantity, one belongs to the electron, one to the proton.' In December 1929, Dirac (back in Cambridge) dissented[55]: 'One cannot simply assert that a negative energy electron *is* a proton, since this would violate charge conservation if an electron jumps from a positive to a negative energy state.'[56] Rather, 'Let us assume . . . that all the states of negative energy are occupied, except, perhaps, for a few of very small velocity,' this occupation being one electron per state, as the exclusion principle demands. Imagine that one such negative energy electron is removed, leaving a hole in the initial distribution. The result is a rise in energy and in charge by one unit. This hole, Dirac noted, acts like a particle with positive energy and positive charge. 'We are . . . led to the assumption that the holes in the distribution of negative energy electrons are the protons.'[56]

The identification of holes with particles is fine, but why protons? Dirac later remarked: 'At that time . . . everyone felt pretty sure that the electrons and the protons were the only elementary particles in Nature.'[57] (Recall that, in 1929, the atomic nucleus was still believed to be built up of protons and electrons![58])

Just prior to submitting his paper, Dirac wrote a letter[59] to Bohr which shows that he knew quite well that, at least in the absence of interactions, his holes should have the same mass as the electrons themselves. It was his hope (an idle one) that this equality would be violated by electromagnetic interactions:

> So long as one neglects interaction, one has complete symmetry between electrons and protons; one could regard the protons as the real particles and the electrons as the holes in the distribution of protons of negative energy. However, when the interaction between the electrons is taken into account, this symmetry is spoilt. I have not yet worked out mathematically the consequences of the interaction . . . One can hope, however, that a proper theory of this will enable one to calculate the ratio of the masses of protons and electrons.

Actually the 'complete symmetry' of which Dirac wrote, charge conjugation invariance, extends to the electromagnetic interactions as well. For want of a better procedure, Dirac briefly considered the mass *m* in his equation to be the average of the proton and the electron mass.[60]

The hole theory was in this fumbled state when Dirac reported on its status at a meeting of the British Association for the Advancement of Science in Bristol. According to the *New York Times*[61] he bewildered his audience – no wonder. 'Later Doctor Dirac was asked to discuss this theory but he shook his head, saying he could not express his meaning in simpler language without becoming inaccurate.'

The confusion lasted all through 1930 when first Oppenheimer,[62] then, independently, Tamm[63] noted that the proton proposal would make all atoms unstable because of the process: proton + electron → photons. In November 1930, Weyl took a new stand[64] in regard to the protons:

> However attractive this idea may seem at first, it is certainly impossible to hold without introducing other profound modifications . . . indeed, according to (the hole theory), the mass of the proton should be the same as the mass of the electron; furthermore . . . this hypothesis leads to the essential equivalence of positive and negative electricity under all circumstances . . . the dissimilarity of the two kinds of electricity thus seems to hide a secret of Nature which lies yet deeper than the dissimilarity between the past and future . . . I fear that the clouds hanging over this part of the subject will roll together to form a new crisis in quantum physics.

Then, in May 1931, Dirac bit the bullet[65] (or, in his words, he made 'a small step forward'[43]): 'A hole, if there were one, would be a new kind of particle, unknown to experimental physics, having the same mass and opposite charge of the electron.' Dirac eventually called the new particle anti-electron. Just before the year's end, Carl Anderson made the first announcement[66] of experimental evidence for the anti-electron. The name positron first appeared in print in one of his later papers.[67] The prediction and subsequent discovery of the positron rank among the great triumphs of modern physics.

That, however, was not at once obvious.

The detection of the positron was considered by nearly everyone as a vindication of Dirac's theory. Yet its basic idea, a positron as a hole in an infinite sea of negative electrons, remained unpalatable to some, and not without reason. Even the simplest state, the vacuum, was a complex consisting of infinitely many particles, the totally filled sea. Interactions between these particles left aside, the vacuum had a negative infinite 'zero point energy' and an infinite 'zero point charge.' Pauli did not like that. Even after the positron had been discovered, he wrote to Dirac: 'I do not believe in your perception of "holes" even if the "anti-electron" is proved.'[68] That was not all, however. Pauli to Heisenberg one month later: 'I do not believe in the hole theory, since I would like to have asymmetries between positive and negative electricity in the laws of nature (it does not satisfy me to shift the empirically established asymmetry to one of the initial state).'[69]

The zero point energy and charge are actually innocuous and can be eliminated by a simple reformulation of the theory.[70] Even thereafter, the theory is still riddled with infinities caused by interactions, however. To this day, the influence of interactions cannot be treated rigorously. Rather, one uses the fact that the fundamental charge e is small, more precisely that the dimensionless number $\alpha = e^2/\hbar c \simeq 1/137$ is small, and expands in α. To leading power in α, theoretical predictions were excellent for processes like photo-electron scattering, the creation and annihilation of electron–positron pairs, and many others. Contributions to these same processes stemming from higher powers in α are invariably infinitely large, however. One was faced with a crisis: how to cope with a theory which works very well approximately but which makes no sense rigorously. As Pauli put it in 1936 during a seminar given in Princeton: 'Success seems to have been on the side of Dirac rather than of logic.'[71] Or, as Heisenberg put it,[72] in a letter to Pauli (1935):

> In regard to quantum electrodynamics, we are still at the stage in which we were in 1922 with regard to quantum mechanics. We know that everything is wrong. But, in order to find the direction in which we should depart from what prevails, we must know the consequences of the prevailing formalism much better than we do.

Heisenberg was, in fact, one of that quite small band of theoretical physicists who had the courage to explore those aspects of quantum electrodynamics which remained in an uncertain state until the late 1940s, when renormalization would provide more systematic and more successful ways of handling the problem.

The first steps toward renormalization go back once again to Dirac. In August 1933, he had written[73] to Bohr:

> Peierls and I have been looking into the question of the change in the distribution of negative energy electrons produced by a static electric field. We find that this changed distribution causes a partial neutralization of the charge producing the field . . . If we neglect the disturbance that the field produces in negative energy electrons with energies less than $-137mc^2$, then the neutralization of charge produced by the other negative energy electrons is small and of the order 136/137 . . . The effective charges are what one measures in all low-energy experiments, and the experimentally determined value of *e* must be the effective charge on an electron, the real value being slightly bigger . . . One would expect some small alterations in the Rutherford scattering formula, the

> Klein–Nishina formula, the Sommerfeld fine structure formula, etc., when energies of the order mc^2 come into play.

Transcribed into the modern vernacular, Dirac's effective charge is our physical charge; his real charge our bare charge; his neutralization of charge our charge renormalization; and his disturbance that the field produces in negative energy electrons our vacuum polarization.[74]

In quantitative form, the results Dirac had mentioned to Bohr are found in his report[75] to the seventh Solvay conference (October 1933), the paper that marks the beginning of positron theory as a serious discipline. There, Dirac also gives the finite contribution to the vacuum polarization[76] which, in 1935, was to be evaluated by Uehling[77] for an electron moving in a hydrogen-like atom – a result which, in turn, was to provide the direct stimulus for the celebrated Lamb shift experiments of 1946.

With Dirac's Solvay report his exquisite burst of creativity at the outer frontiers of physics, spanning eight years, comes to an end.

The years 1925–33 are the heroic period in Dirac's life, during which he emerged as one of the principal figures in twentieth century science and changed the face of physics. He himself has called those years in his scientific career 'the exciting era.'[78] My foregoing sketch of that period is not, by any means, complete. For example, in 1931, Dirac produced[65] the first application of global topology to physics, his proof that the existence of magnetic monopoles implies, quantum mechanically, that electric charge is quantized. He returned to this subject some twenty years later[79] (he lectured upon it[80] at the Pocono conference, March 31–April 1, 1948) and, once again, nearly thirty years thereafter.[81] As these intervals illustrate, Dirac remained scientifically active for the fifty years following the developments that came to a close in 1933.

I shall turn shortly to a summary of those later undertakings by Dirac but will first make some comments on his personal life in the 1930s.

In 1933, Dirac received the Nobel Prize 'for his discovery of new fertile forms of the theory of atoms and for its applications,' sharing the award with Schrödinger. 'At first he was inclined to refuse the prize because he did not like publicity, but when Rutherford told him: "A refusal will get you much more publicity," he accepted.'[82] At that time he had ceased all contact with his father, so he only took his mother along to Stockholm, where he delivered his Nobel lecture.[83]

Much to his dismay, the Nobel Prize did make Dirac a public person. A London paper characterized him 'as shy as a gazelle and modest as a Victorian maid,' and called him 'The genius who fears all women.'[84]

Well, not quite all.

As mentioned before, Dirac was in Princeton during the academic year 1934–5. That autumn, Eugene Wigner (1902–95), professor of physics at Princeton University, received a visit from his sister, Margit Wigner Balasz (Manci to her friends), who lived in her native Budapest. Manci and Paul met. 'He spoke to me about his difficult, I should say very difficult, childhood, I told him about mine, which also left some sad memories about my unhappy marriage.'[85] In the summer of 1935 Paul visited Manci in Budapest. Manci has written a loving, tender account of their courtship.[85] They married on January 2, 1937. 'So started a very old-fashioned Victorian marriage.'[85] Paul gave up his bachelor quarters in St John's College. The couple moved into the house on Cavendish Avenue, where I first met them. They were joined by Manci's two children from her previous marriage, Judith and Gabriel (1925–84) – who became a mathematician of distinction – who both adopted the name Dirac. Paul and Manci had two daughters, Monica (b. 1940), and Florence (b. 1942). 'Paul, although not a domineering father, kept himself aloof from his children.'[85]

After Paul's father's death in 1936, Paul wrote to Manci: 'I feel much freer now.'[85] His mother became a frequent visitor to Cavendish Avenue. It was there that she died, in 1941.

As promised, I now continue with an account of Dirac's later work, and begin with some of his lesser known researches. First, in 1933 he collaborated with his good friend Pyotr Kapitza (1894–1984) on a theoretical study of the reflection of electrons from standing light waves.[86] This 'Kapitza–Dirac effect' was not experimentally observed until 1986.[87]

Secondly, also in 1933, Dirac invented a centrifugal method for separating gaseous isotope mixtures. Kapitza encouraged him to carry out the experiments himself, which Dirac did but did not complete. Dalitz has given a detailed account[88] of how, after 1940, construction projects of atomic bombs revived interest in that work, and how Dirac became an informal consultant for that project. He also contributed in a quite different way to the War effort, being a member of the small fire fighter team of St John's College, Cambridge, during the period when fire raids were expected (according to a letter dated April 28, 1993, from H. Peisir to R. Hovis, now in the Archives of St John's).

Interesting though these two topics are, they must be considered as digressions from Dirac's main later pursuits of fundamental issues, in which he continued to show his high mathematical inventiveness and craftsmanship but no longer that almost startling combination of novelty and simplicity that mark his heroic period.

Without pretence to completeness, and in fairly random order, here are some main themes which, as I see it, convey the flavor of his thinking in his later years.

Elaborations of Hamiltonian dynamics. These include studies of the special relativistic dynamical evolution of systems on various types of hypersurfaces, in classical theory[89] and in quantum mechanics.[90] Also, investigations of constrained Hamiltonian systems,[91, 92] leading to his Hamiltonian formulation of general relativity.[93] That work, in turn, aroused his interest in gravitational waves.[94] Did Dirac coin the name graviton? According to the *New York Times* of January 31, 1959, 'Professor Dirac proposed that the gravitational wave units be called gravitons.'

Related to Dirac's lifelong interest in general relativity are his papers on wave equations in conformal,[95] de Sitter,[96] and Riemannian spaces.[97] He lectured on general relativity until in his seventies.[98]

Cosmological issues, in which he had become interested already in his Goettingen days.[99] He did not publish on this subject until 1937.[100] From then on, until the end of his life, he was much intrigued by the possibility that the fundamental constants in nature actually are not constant but depend on time in a scale set by the cosmological epoch, the time interval between the big bang and the present.[101] It was his hope that simple relations should emerge between such extremely large but roughly comparable numbers as the ratio of epoch to atomic time intervals and the ratio of electric to gravitational forces between an electron and a proton.[102] No definitive advance was ever achieved. Others followed these exploits with more interest than enthusiasm.

The aether. A brief period (1951–3) of speculations to the effect that quantum mechanics allows for the existence of an aether.[103]

Quantum electrodynamics. One further contribution still belongs to the heroic period. In March 1932, Dirac proposed a 'many-time formalism' in which an individual time is assigned to each electron.[104] This new version of the theory, equivalent to earlier formulations,[105] marks an important first step toward the manifestly covarient procedures that were to play such a key role from the late 1940s on.

A few years later, Dirac turned highly critical of quantum electrodynamics. On the one hand, the work he produced as a result of this negative attitude has not in any way enhanced our understanding of fundamental issues. On the other hand, these later struggles are of prime importance for an understanding of Dirac himself. His radically modified position resulted from his work[75] on vacuum polarization in which he had encountered the infinities that, as said, constituted a crisis in the quantum field theory of the 1930s.

Dirac's drastic change in attitude is starkly expressed in a brief paper he wrote in 1936, his first publication following his involvement[75] with the implications of positron theory. I regard it as significant that this article followed a period during which he had not published at all for more than a year. The *a propos* was a fleeting experimental doubt about the validity of the theory of photo-electron scattering. Dirac reacted[106] as follows:

> The only important part (of theoretical physics) that we have to give up is quantum electrodynamics . . . we may give it up without regrets . . . in fact, on account of its extreme complexity, most physicists will be very glad to see the end of it.

At this point, it should be recalled that the germs of the difficulties with the infinities date back to the classical era. A classical electron considered as a point particle has an infinite energy due to the coupling to its own electrostatic field. With this in mind, Dirac adopted the strategy of attempting to modify the classical theory first, so as to rid it of *its* infinities, and thereupon to revisit the quantum theory in the hope that also there all would be well. At that time, that approach was followed also by others, Born, Kramers, and Wentzel among them. Even today, there remains a much needed understanding of what lies beyond the infinities.

There are overwhelming reasons, however, why a return to the classical theory is the wrong way to go.[107]

Be that as it may, Dirac tried several times to reformulate the classical theory of the electron. His first attempt[108] dates from 1938. 'A new physical theory is needed which should be intelligible both in the classical and in the quantum theory and our easiest path of approach is to keep within the confines of the classical theory.' He started from the observations that Lorentz's classical theory of the electron's motion is not rigorously valid for high accelerations, since Lorentz's electron has a finite radius. Dirac, instead, started from a zero radius electron and was able to find a rigorous classical equation of motion for it which is free of the classical infinities but which exhibits new pathologies: it has solutions corresponding to accelerations even in the absence of external fields. He did find a not very palatable constraint that eliminates these unwanted solutions – but there was more trouble. New infinities arose upon quantizing the theory.[109] In order to eliminate these, Dirac introduced[110] what amounts to photons of negative energy. He attempted to eliminate the physical paradoxes resulting from this new postulate by introducing an indefinite metric in Hilbert space.[111] That, however, leads to still further difficulties, critically analyzed by Pauli.[112] These new postulates were never discussed in the context of positron theory.

Unable to find a satisfactory quantum version of his point electron, Dirac never mentioned this theory again in later years. By 1946, he tended to the view that the infinities are a mathematical artifact resulting from expansions in α that are actually invalid.[113]

Shortly thereafter, in the years 1947–78, quantum electrodynamics took a new turn when the renormalization program was systematically developed. That technique does not fully resolve the problem of the infinities. The electron's mass and charge unalterably remain finite. To a very large extent, these two infinities can be rendered harmless, however, in the sense that predictions to arbi-

FIGURE 1.1

Oppenheimer, Dirac and Pais in the common room, Institute for Advanced Studies.

(Photo by Alfred Eiderstadt, courtesy of P.A.M. Dirac and A. Pais)

trarily high orders in α can now be made for the scattering, creation, and annihilation processes mentioned earlier, where, before, the leading order in α had worked so well, but the higher orders had been intractable. As a result, quantum electrodynamics could now be confronted with experiment to vastly improved orders of magnitude. The results were spectacular. With good reason, Feynman has called[114] the new version of quantum electrodynamics 'the jewel of physics – our proudest possession.'

Dirac would have none of it.

In 1951, he wrote: 'Recent work by Lamb, Schwinger and Feynman and others has been very successful . . . but the resulting theory is an ugly and incomplete one.'[115] He had a deep aversion to

FIGURE 1.2

Dirac with Wolfgang Pauli and Rudolph Peierls, Birmingham, 1949.

(Photo courtesy of Margit Dirac)

the way infinite masses and charges are manipulated in the renormalization program. In that year, he started all over again for a second time in his search for a new, classical, point of departure. 'The troubles . . . should be ascribed . . . to our working from the wrong classical theory.'[115] His new suggestion may be considered as the extreme opposite of what he had proposed in 1938. This time, he began with a classical theory that does not contain discrete particles at all. 'The notion of electrons should be built up from a classical theory of the motion of a continuous stream of electricity rather than the motion of point charges. One then looks upon the discrete electrons as a quantum phenomenon.'[116, 117]

After 1954, this model, too, vanished from his writings without leaving a trace.

Thus, from the early 1950s on, Dirac went his own lonely way. He accepted the successes of the renormalization method. In fact, in the mid-1960s, he lectured on the anomalous magnetic moment and Lamb shift calculations.[118] He never wavered in his belief, however, that quantum electrodynamics needed a new starting point. In later years, he would occasionally seek new remedies in a reformulation not so much of classical as of quantum theory.[119] In 1970, he invented the last of the Dirac equations, a relativistic wave equation with positive energies only.[120]

From September 1970 to January 1971, Dirac was Visiting Professor at the Florida State University in Tallahassee. During that time he was offered a permanent position there, which he accepted. In 1972 he started a new life as Professor in Florida. One of his colleagues there has told me:

> At that time he was also courted by the New York State University at Stony Brook and by Miami. He declined those offers, principally because he could not go for walks there . . . In Tallahassee he walked about a mile to work . . . He was fond of swimming in nearby Silver Lake and Lost Lake, also sometimes at the seashore.
>
> Dirac was most happy in Tallahassee, he really changed. In Cambridge he only went to the University for classes and seminars but otherwise worked at home. In Tallahassee he came diligently all day, ate lunch with the boys, took a nap after lunch. His wife would pick him up in the late afternoon . . . We treated him like one of the boys . . . did not indulge in much red carpet treatment. He liked that.[121]

Dirac's writings in the Florida period are simply prolific. He published over 60 papers in those last 12 years of his life, most of them reviews of past events, including a short book on general relativity.[122] I cherish a 'Dear Bram' letter he wrote to me[52] in those days, thanking me for a copy of my scientific biography of Einstein. On the back flap of that volume one finds words of praise by Dirac for that book.

Dirac's last paper (1984), entitled 'The inadequacies of quantum field theory,'[123] contains his last judgment on quantum electrodynamics:

> These rules of renormalization give surprisingly, excessively good agreement with experiments. Most physicists say that these working rules are, therefore, correct. I feel that is not an adequate reason. Just because the results happen to be in agreement with observation does not prove that one's theory is correct.

The paper concludes with Dirac's final published scientific words:

> I have spent many years searching for a Hamiltonian to bring into the theory and have not yet found it. I shall continue to work on it as long as I can, and other people, I hope, will follow along such lines.

Dirac died on October 20, 1984, aged 82. He was buried in the Roselawn Cemetery in Tallahassee. It was his family's wish that he should rest where he left the world.

I have been charged to speak today of aspects of Dirac's life and work. It has been correctly said of him that his life was mostly science and his science was physics. That is reflected in what I have discussed so far: mostly his science with only brief digressions about other features of his life. I would be remiss, however, if I would not flesh out these latter aspects some more. This I do now, in my final comments.

Dirac's ascetic lifestyle, his indifference to discomfort or food has been likened to that of Gandhi.[124] He neither touched alcohol nor smoked. He shunned publicity and honors, of which he nevertheless received many.[125] Regarding religion, he tended towards atheism, as he has publicly expressed only once.[126] As Pauli once said: 'There is no God and Dirac is his prophet.'[127] Manci Dirac has written to me, however: 'Paul was no atheist. Many times did we kneel side by side in Chapel, praying. We all know, he was no hypocrite.'[127a]

Throughout his life, Dirac maintained a minimal, sparse (not terse), precise, and apoetically elegant style of speech and writing. Sample: his comment on the novel *Crime and Punishment*: 'It is nice, but in one of the chapters the author made a mistake. He describes the sun as rising twice on the same day.'[128] Once when Oppenheimer offered Dirac some books to read, he politely refused, saying that reading books interfered with thought.[129]

After his marriage Dirac became a keen gardener, and tried to deal with horticultural problems from first principles, which did not always lead to good results.[130]

I turn to my personal contacts with Dirac, mainly those at the Institute in Princeton, which began in the fall of 1946. At that time we would often have lunch together. It was on one of those occasions that I had my first exposure to the Dirac style of exhaustive inquiry. Because of a large appetite and a Dutch background, I would regularly eat three sandwiches at that time. One day, Dirac queried me. (Between each answer and the next question there was a half minute's pause.) D. Do you always eat three sandwiches for lunch? P. Yes. D. Do you always eat the same three sandwiches for lunch? P. No, it depends on my taste of the day. D. Do you eat your sandwiches in some fixed order? P. No. Some months later, when a young man named Salam visited me at the Institute, he said: I have regards for you from Professor Dirac in Cambridge. He wants to know if you still eat three sandwiches for lunch. Dirac and I often lunched together when he came back to the Institute for the

FIGURE 1.3

Dirac lifting a tree on his sister-in-law's farm.

(Photo courtesy of Margit Dirac)

FIGURE 1.4

Birthday celebration at Trieste, 1972.

(Photo courtesy of Margit Dirac)

academic year 1947–8. On an early occasion, Dirac looked at my plate and noted, triumphantly: 'Now you only eat two sandwiches for lunch.' Another recollection: A corridor conversation at the Institute. D. My wife wants to know if you can come for dinner tonight. P. I regret. I have another engagement. D. Goodbye. Nothing unfriendly implied. Nothing else said like 'Some other time perhaps.' The question had been posed and answered, the conversation was finished.

Everything had been arranged at the Institute for Dirac's next visit in the academic year 1954–5. It was not to be. The events of the troubled spring of 1954 were summarized in the News and Views column of *Physics Today*, July 1954, under two headings: The Oppenheimer Case; Dirac denied Visa. Dirac had been

FIGURE 1.5

With Sir George Thomson and von Laue, at Lindau.

(Photo courtesy of Margit Dirac)

informed by the American Consulate in London that he was ineligible for a visa under Section 212A of the Immigration and Naturalization Act, the infamous McCarran Act which (to quote *Physics Today*) 'Covers categories of undesirables ranging from vagrants to stowaways.' The reasons for this decision have never become quite clear, but it was believed that Dirac's seven pre-War visits to Russia, three in the course of his three trips around the world, and all for scientific purposes, had something to do with it.[131] The event, widely reported in the world press,[132] caused some American physicists to write to the *New York Times*: 'If this is what the McCarran Act means in practice, it seems to us a form of cultural suicide.'[133] It was a quite bad, yet by no means the worst, case of harm done during that period. It passed.

In 1988 I requested and received Dirac's FBI files, which contain only one line which I find moderately pertinent: 'The reason for Dirac's [1954] visit here was to discuss with Oppenheimer an invitation from Cambridge University to accept an offer as a professor. Dr Oppenheimer, bitter over the [security clearance] vote against him, will accept that British offer.' For the rest these documents are monumentally uninteresting.

Later Dirac was to spend two more academic years in Princeton. During all those visits I would draw him out, time and again, about his discontent with quantum electrodynamics. He would concede the successes of renormalization but forever was of the opinion that the remaining mass and charge infinities 'ought not to be there. They remove them artificially.'[123] This diagnosis may well be much better than the cures he proposed.

Other recollections: his evident pride at having invented the bra and ket notations, announced in a paper[134] specially written for this purpose. His reply to my question, posed in the early 1960s, why space reflexion and time reversal invariance do not appear in his book on quantum mechanics: 'Because I did not believe in them.' Indeed, in 1949, he had written: 'I do not believe there is any need for physical laws to be invariant under these reflections, although all the exact laws of nature so far known do have this invariance.'[135]

By far, the most revealing insight I gained from those discussions concerned the Dirac way of playing with equations, which can be summed up like this: first play with pretty mathematics for its own sake, then see whether this leads to new physics.

Throughout most of his life, that attitude is manifest in his writings. At age 28:

> There are, at present, fundamental problems in theoretical physics . . . the solution of which . . . will presumably require a more drastic revision of our fundamental concepts than any that have gone before. Quite likely, these changes will be so great that it

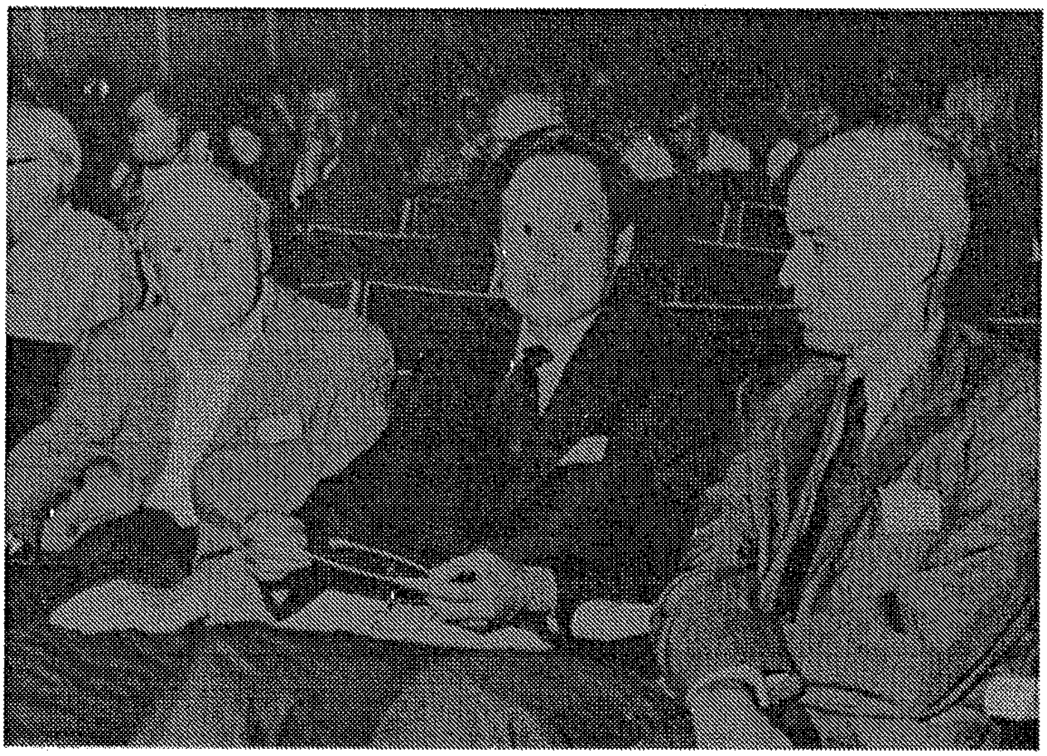


FIGURE 1.6

Dirac with C.N. Yang and W. Lamb.

(Photo courtesy of Margit Dirac)

> will be beyond the power of human intelligence to get the necessary new ideas by direct attempts to formulate the experimental data in mathematical terms. The theoretical worker in the future will, therefore, have to proceed in a more direct way. The most powerful method of advance that can be suggested at present is to employ all the resources of pure mathematics in attempts to perfect and generalize the mathematical formalism that forms the existing basis of theoretical physics, and after each success in this direction, to try to interpret the new mathematical features in terms of physical entities[65]

which is just what is happening these days. At age 36: 'As time goes on, it becomes increasingly evident that the rules which the mathematician finds interesting are the same as those which Nature has

FIGURE 1.7

With Kapitza, Wigner and daughter, in Lindau, 1982.

(Photo courtesy of Margit Dirac)

chosen.'[34] At age 60: 'I think it's a peculiarity of myself that I like to play about with equations, just looking for beautiful mathematical relations which maybe don't have any physical meaning at all. Sometimes they do.'[2] At age 78:

> A good deal of my research work in physics has consisted in not setting out to solve some particular problem, but simply examining mathematical quantities of a kind that physicists use and trying to fit them together in an interesting way, regardless of any application that the work may have. It is simply a search for pretty mathematics. It may turn out later that the work does have an application. Then one has good luck.[40]

In that last paper he gave three examples of the way he played: the Dirac equation, monopoles, and the last Dirac equation. His own judgment, at age 69: 'My own contributions since [the] early days have been of minor importance.'[135]

What kinds of mathematics did Dirac consider pretty?

> The research worker, in his efforts to express the fundamental laws of Nature in mathematical form, should strive mainly for mathematical beauty. He should take simplicity into consideration in a subordinate way to beauty . . . It often happens that the requirements of simplicity and beauty are the same, but where they clash the latter must take precedence.[34]

It is, of course, idle to argue about such subjective issues as the distinction between beauty and simplicity.

Dirac was a very private man, not much given to reminiscing about other personalities or past events. He would only rarely talk about himself. On a few occasions, he would reveal some of his emotions in his writings, however. I find it striking that, as mentioned, he would refer to the transformation theory as 'my darling.'[7] Equally notable are his rare utterances about anxiety. When, at age 60, he was asked about his feelings on discovering the Dirac equation, he replied: 'Well, in the first place, it leads to great anxiety as to whether it's going to be correct or not . . . I expect that's the dominating feeling. It gets to be rather a fever . . . '[2] At age 67: 'Hopes are always accompanied by fears, and, in scientific research, the fears are liable to become dominant.'[137] At age 69: 'I think it is a general rule that the originator of a new idea is not the most suitable person to develop it, because his fears of something going wrong are really too strong . . . '[136]

As my last example of Dirac talking about himself I quote from a letter[138] to me by a colleague: 'I had a conversation with him about a year and a half before his death . . . I asked him to come and talk at the University of Florida, and he said: "No! I have nothing to talk about. My life has been a failure . . . " And then went on to talk about the infinities [in quantum electrodynamics]!!' It is typical for many great men that failure outweighs success.

I should next like to add two Dirac stories to the immense lore about him.

One day Niels Bohr came into my office in Princeton, shaking his head while telling me of a discussion he had just had with Dirac. It was in the early 1950s, during the time of the Cold War. Bohr had expressed his dislike of the abusive language the American press was using in reference to the Russians. Dirac had replied that all this would come to an end in a few weeks' time. Bohr had asked why. Well, Dirac had remarked, by then the reporters will have used up all the invective in the English language, so therefore they will have to stop.

The other story is not about Dirac, but one that I have heard Dirac tell more than once, with relish. In a small village, a newly appointed priest went to call on his parishioners. On a visit to a quite modest home, he was received by the lady of the house. He could not fail to notice that her place was teeming with children and asked her how many the couple had. Ten, she replied, five pairs of twins. Astonished, the priest asked: You mean you always had twins? To which the woman replied: No, Father, sometimes we had nothing. Precision at that level had an immense appeal to Dirac.

My final story about Dirac concerns a letter to me by a friend of his and mine.[139] It concerns my very first encounter with the Diracs, in January 1946, during which Paul had queried me about my war experience. The letter says in part:

> It was about two weeks before his death . . . Margit and I were sitting at his bedside. He was pale, thin, and unusually talkative . . . He said that very near the end of the War you had been captured by the Germans and that you were about to be executed . . . The unusual thing about the situation was that he

> repeated the story in its entirety at least four times . . . Margit finally got through to him and made him stop . . . Some day perhaps you can tell me about it.

As I look back on the almost 40 years I knew Dirac, all memories are fond ones. I share Niels Bohr's opinion of him: 'Of all physicists, Dirac has the purest soul.'[140] In some, but only some, ways he reminds me of Einstein: one of the century's great contributors, always going his own way, not making a school, compelled by the need for beauty and simplicity in physical theory, in his later years more addicted to mathematics than was good for his physics, continuing his activities in pure research until close to his death. In other respects, I never knew anyone quite like him.

References

(In these notes D. stands for P. A.M. Dirac)

1 Parts of this paper were taken from my earlier writings about D., found in: *Aspects of quantum theory*, p. 79, A. Salam and E. P. Wigner, Eds., Cambridge University Press (1972); in *Inward Bound*, Oxford University Press (1986); and in *Reminiscences about a great physicist*, p. 93, B. Kursunoglu and E. P. Wigner, Eds., Cambridge University Press (1987). My principal secondary sources have been the fine writings by R. H. Dalitz and R. Peierls, *Biogr. Mem. Fell. R. Soc.* **32**, 139 (1986); and by H. S. Kragh, *Dirac*, Cambridge University Press (1990).

2 T. Kuhn, interview with D., May 7, 1963, Niels Bohr Archive, Copenhagen.

3 D., "A little 'prehistory,'" *The Old Cathamian*, p. 9 (1980).

4 D., in *History of twentieth century physics*, p. 109, C. Weiner, Ed., Academic Press, New York (1977).

5 D., interview in *Florida State University Bulletin*, Vol. **3**, February 1978.

6 For an account of D.'s Cambridge days, see R. J. Eden & J. C. Polkinghorne, in *Aspects of Quantum Theory*, p. 1, A. Salam & E. P. Wigner, Eds., Cambridge University Press (1972).

7 D., Report KFKI–1977–62, *Hung. Acad. of Sc.*

8 D., *Proc. R. Soc. Lond.* A**109**, 642 (1925).

9 M. Born, *My Life*, p. 226, Scribner, New York (1978).

10 *Reminiscences about I. E. Tamm*, E. Feinberg, Ed., Nauka, Moscow (1987).

11 Cf. D., *Proc. Camb. Phil. Soc.* **23**, 412 (1926).

12 D., *Proc. R. Soc. Lond.* A**112**, 661 (1926). For many more details about Dirac's early years and his contributions to quantum mechanics during 1925–6, see J. Mehra & H. Rechengberg, *The Historical Development of Quantum Theory*, Vol. **4**, part 1, Springer, New York (1982).

13 W. Heisenberg, *Zeitschr. f. Phys.* **38**, 411 (1926).

14 W. Heisenberg, *Zeitschr. f. Phys.* **39**, 499 (1926).

15 E. Fermi, *Rend. Lincei* **3**, 145 (1926); *Zeitschr. f. Phys.* **36**, 902 (1926); repr. in *Enrico Fermi, Collected Works*, Vol. **1**, pp. 181, 186, Univ. of Chicago Press (1962). In ref. 4, pp. 133, 134, Dirac has given a charming account of the time sequence of his and Fermi's contributions.

16 For the history of quantum statistics in the days of the old quantum theory, see A. Pais, *Inward Bound* (ref. 1), Chap. 13, section (d), Oxford University Press.

17 D., *Proc. R. Soc. Lond.* A**113**, 621 (1927).

18 Rigorous treatments lead to the theory of distributions, cf. I. Halperin & L. Schwartz, *Introduction to the Theory of Distributions*, Toronto University Press (1952).

19 D., *Proc. R. Soc. Lond.* A**114**, 243 (1927).

20 D., *Proc. R. Soc. Lond.* A**114**, 710 (1927).

21 This so-called semi-classical procedure (discussed in detail by W. Pauli, *Handbuch der Physick*, Vol. **24/1**, sections 15, 16, Springer,

Berlin (1933)) allows for a good approximate but not rigorous treatment of induced processes; radiative corrections are not properly accounted for.

22 A. Einstein, *Phys. Zeitschr.* **18**, 121 (1917). See further A. Pais *Subtle is the Lord*, Chap. 21, section (d), Oxford University Press, New York (1982).

23 D. was aware[19] that he missed a factor two in this coefficient because he had not yet treated polarization properly.

24 Independently of Schrödinger, *Ann. der Phys.* **81**, 109 (1926).

25 A far more detailed analysis of D.'s two founding papers on quantum electrodynamics has been given by R. Jost, in *Aspects of Quantum Theory*, p. 61, A. Salam & E. P. Wigner, Eds., Cambridge University Press (1972).

26 Ref. 20, p. 719.

27 D., *Proc. R. Soc. Lond.* **A111**, 405 (1926); *Proc. Camb. Phil. Soc.* **23**, 500 (1926).

28 L. Infeld, *Quest*, p . 203, 2nd Ed., Chelsea, New York (1980).

29 See especially the account of D.'s research students by Dalitz and Peierls, ref. 1, pp. 155–7.

30 H. B. G. Casimir, *Haphazard Reality*, p. 72, Harper and Row, New York (1983).

31 R. Eden and J. Polkinghorne, in *Tributes to Paul Dirac*, p. 5, J. C. Taylor, Ed., Hilger, Bristol (1987).

32 N. F. Mott, interviewed by T. S. Kuhn, March 1962, Niels Bohr Archive, Copenhagen.

33 A. Einstein, in *James Clerk Maxwell*, p. 66, New York (1931).

34 D., *Proc. R. Soc. Edinburgh* **59**, 122 (1939).

35 D., *Sci. Am.* **208**, 45 (May 1963).

36 D., in *Albert Einstein, historical and cultural perspectives*, p. 79, G. Holton and Y. Elkana, Eds., Princeton University Press (1982).

37 O. Klein, *Zeitschr. f. Phys.* **37**, 895 (1926); E. Schrödinger, *Ann. der Phys.* **81**, 109 (1926); V. Fock, *Zeitschr. f. Phys.* **38**, 242 (1926); Th. de Donder & H. van den Dungen, *Comptes Rendues* **183**, 22 (1926); J.

Kudar, *Ann. der Phys.* **81**, 632 (1926); W. Gordon, *Zeitschr. f. Phys.* **40**, 117 (1926).

38 W. Pauli, *Zeitschr. f. Phys.* **43**, 601 (1927).

39 He was looking for a four-dimensional generalization of $\sigma \times \boldsymbol{p}$. Later he was to play briefly with wave equations for higher spin, D., *Proc. R. Soc. Lond.* A**155**, 447 (1936).

40 D., *Int. J. Theor. Phys.* **21**, 603 (1982).

41 D., *Proc. R. Soc. Lond.* A**117**, 610 (1928).

42 D., *Proc. R. Soc. Lond.* A**118**, 351 (1928).

43 Ref. 2, interview May 14, 1963.

44 It was later shown by Pauli and Weisskopf (*Helv. Phys. Acta* **7**, 709 (1934)) that the scalar wave equation is amenable to a treatment compatible with the transformation theory.

45 W. Heisenberg, interviewed by T. Kuhn, July 12, 1963, Niels Bohr Library, American Institute of Physics, New York.

46 D., *Phys. Zeitschr.* **29**, 561, 712 (1928).

47 W. Heisenberg, letter to W. Pauli, May 3, 1928, reproduced in *Wolfang Pauli, Scientific Correspondence*, Vol. **1**, p. 443, Springer, New York (1979); referred to as PC below.

48 Ref. 46, p. 562, footnote 2.

49 W. Heisenberg, letter to W. Pauli, July 31, 1928, PC, Vol. **1**, p. 466.

50 D., letter to O. Klein, July 24, 1928, copy in Niels Bohr Library.

51 S. F. Tuan, *Dirac and Heisenberg in Hawaii*, unpublished manuscript.

52 D., letter to A. Pais, October 21, 1982.

53 W. Heisenberg, in *The Physicist's conception of nature*, p. 816, J. Mehra, Ed., Reidel, Dordrecht (1973).

54 H. Weyl, *Zeitschr. f. Phys.* **56**, 330 (1929).

55 Other pertinent developments which had meanwhile taken place include the derivations of the Klein–Nishina formula for Compton scattering; and of the Klein paradox. See further A. Pais, *Inward Bound* (ref. 1), Chap. 15, section (f).

56 D., *Proc. R. Soc. Lond.* A**126**, 360 (1929); also *Nature* **126**, 605 (1930).

57 D., ref. 4, p. 144.

58 See A. Pais, *Inward Bound* (ref. 1), Chap. 14.

59 D., letter to N. Bohr, November 26, 1929, copy in Niels Bohr Library.

60 D., *Proc. Camb. Phil. Soc.* **26**, 361 (1930).

61 *The New York Times*, September 9, 1930.

62 J. R. Oppenheimer, *Phys. Rev.* **35**, 562 (1930).

63 I. Tamm, *Zeitschr. f. Phys.* **62**, 545 (1930).

64 H. Weyl, *The Theory of Groups and Quantum Mechanics*, pp. 263–4 and preface, Dover, New York.

65 D., *Proc. R. Soc. Lond.* **A133**, 60 (1931).

66 C. D. Anderson, *Science* **76**, 238 (1932).

67 C. D. Anderson, *Phys. Rev.* **43**, 491 (1933).

68 W. Pauli, letter to D., May 1, 1933, PC, Vol. **2**, p. 159.

69 W. Pauli, letter to W. Heisenberg, June 16, 1933, PC, Vol. **2**, p. 169.

70 Cf. A. Pais, *Inward bound* (ref. 1), Chap. 16, section (d).

71 *The Theory of the Positron and Related Topics*, report of a seminar conducted by W. Pauli, notes by B. Hoffman, Institute for Advanced Study, Princeton (1935–6), mimeographed notes.

72 W. Heisenberg, letter to W. Pauli, PC, Vol. **2**, p. 386.

73 D., letter to N. Bohr, August 10, 1933, copy in Niels Bohr Library.

74 The existence of vacuum polarization was also independently diagnosed by W. H. Furry & J. R. Oppenheimer, *Phys. Rev.* **45**, 245, 343 (1934).

75 D., in *Rapports du Septième Conseil de Physique*, p. 203, Gauthier-Villars, Paris (1934); cf. also D., *Proc. Camb. Phil. Soc.* **30**, 150 (1934).

76 A numerical error in the coefficient of that finite term was corrected by W. Heisenberg, *Zeitschr. f. Phys.* **90**, 209 (1934).

77 E. Uehling, *Phys. Rev.* **48**, 55 (1935).

78 Ref. 4, p. 140.

79 D., *Phys. Rev.* **74**, 817 (1948).

80 D., in dittoed notes of the Pocono conference, p. 72, unpublished.

81 D., in *New Pathways in Science*, Vol. **1**. A. Perlmutter, Ed., Plenum Press, New York (1976); see further E. Amaldi & N. Cabibbo, in *Aspects of Quantum Theory*, ref. 1, p. 183.

82 Dalitz and Peierls (ref. 1), p. 150.

83 D., 'Theory of electrons and positrons,' in *Nobel lectures in physics, 1922–1941*, p. 320, Elsevier, Amsterdam (1965).

84 *Sunday Dispatch*, November 19, 1933.

85 Margit Dirac, in Kursunoglu and Wigner (ref. 1), p. 3.

86 D., and P. Kapitza, *Proc. Camb. Phil. Soc.* **29**, 297 (1933).

87 P. Gould *et al.*, *Phys. Rev. Lett.* **56**, 827 (1986).

88 R. H. Dalitz, in Kursunoglu and Wigner (ref. 1), p. 69; also Dalitz and Peierls (ref. 1), p. 152.

89 D., *Rev. Mod. Phys.* **21**, 392 (1949).

90 D., *Phys. Rev.* **73**, 1092 (1948); *Proceedings of the Second Canadian Mathematical Congress 1949*, p. 10, University of Toronto Press, Toronto (1951).

91 D., *Can. J. Math.* **2**, 129 (1950); **3**, 1 (1951); *Proc. R. Soc. Lond.* A**246**, 326 (1958); *Proc. R. Irish Acad.* A**63**, 49 (1964).

92 See also F. Rohrlich, in *High Energy Physics*, p. 17, B. Kursunoglu & A. Perlmutter, Eds., Plenum Press, New York (1985).

93 D., *Proc. R. Soc. Lond.* A**246**, 333 (1958); *Phys. Rev.* **114**, 924 (1959); also in *Recent Developments in General Relativity*, p. 191, Pergamon Press, London (1962). See further D., *Proc. R. Soc. Lond.* A**270**, 354 (1962); *Gen. Rel. and Grav.* **5**, 741 (1974).

94 D., *Phys. Rev. Lett.* **2**, 368 (1959); *Proceedings of the Royaumont Conference 1959*, p. 385, Editions du CNRS, Paris (1962); *Phys. Bl.* **16**, 364 (1960).

95 D., *Ann. of Math.* **37**, 429 (1935).

96 D., *Ann. of Math.* **36**, 657 (1935).

97 D., in *Max Planck Festschrift 1958*, p. 339, Deutscher Verlag der Wissenschaften, Berlin (1958).

98 D., *General Theory of Relativity*, Wiley, New York (1975).

99 Ref. 4, p. 149.

100 D., *Nature* **139**, 323, 1001 (1937); also *ibid.* **192**, 441 (1961).

101 D., Report CTS-T. Phys. 69–1, Center for Theoretical Studies, Coral Gables, Florida (1969); *Comm. Pontif. Acad. of Sci* **2**, No. 46 (1973);

3, No. 6 (1975); *Proc. R. Soc. Lond.* A**338**, 446 (1974); *Nature* **254**, 273 (1975); in *Theories and Experiments in High Energy Physics*, p. 443, B. Kursunoglu *et al.*, Eds., Plenum Press, New York (1975); *New Frontiers in High Energy Physics*, p. 1, A. Perlmutter & L. Scott, Eds., Plenum Press, New York (1978); *Proc. R. Soc. Lond.* A**365**, 19 (1979).

102 See further F. J. Dyson, in *Aspects of Quantum Theory* (ref. 1), p. 213.

103 D., *Nature* **168**, 906 (1951); **169**, 146 (1952); *Physica* **19**, 888 (1953); *Sci. Monthly* **78**, 142 (1954).

104 D., *Proc. R. Soc. Lond.* A**136**, 453 (1932).

105 Cf. e.g. D., V. Fock & B. Podolsky, *Phys. Zeitschr. der Sowjetunion* **2**, 468 (1932).

106 D., *Nature* **137**, 298 (1936).

107 A. Pais, *Inward Bound*, (ref. 1), Chap. 16, section (c); Chap. 18, section (a).

108 D., *Proc. R. Soc. Lond.* A**167**, 148 (1938); see also ref. 92.

109 D., *Ann Inst. H. Poincaré* **9**, 13 (1939).

110 D., *Comm. Dublin Inst. Adv. Studies* A**1** (1943).

111 D., *Proc. R. Soc. Lond.* A**180**, 1 (1942).

112 D., *Rev. Mod. Phys.* **15**, 175 (1943).

113 D., *Comm. Dublin Inst. Adv. Studies* A**3** (1946); *Proceedings of the International Conference on Fundamental Particles and Low Temperatures, Cambridge, June 1946*, p. 10, Taylor and Francis, London (1946); *Proceedings of the 8th Solvay Conference 1948*, p. 282, R. Stoops, Ed., Coudenberg, Brussels (1950).

114 R. P. Feynman, *Quantum Electrodynamics, the Strange Story of Light and Matter*, Princeton University Press (1985).

115 D., *Proc. R. Soc. Lond.* A**209**, 251 (1951).

116 See also D., in *Deeper Pathways in High Energy Physics*, B. Kursunoglu *et al.*, Eds., Plenum Press, New York (1977).

117 See further D., *Proc. R. Soc. Lond.* A**212**, 330 (1952); **223**, 438 (1954); also D., *Proc. R. Soc. Lond.* A**257**, 32 (1960); **268**, 57 (1962).

118 D., *Lectures on Quantum Field Theory*, Belfer School of Science, Yeshiva University, New York (1966).

119 Cf. D., *Nuov. Cim. Suppl.* **6**, 322 (1957); *Nature* **203**, 115 (1964); **204**, 771 (1964); *Phys. Rev.* **139B**, 684 (1965).

120 D., *Proc. R. Soc. Lond.* A**322**, 435 (1971); **328**, 1 (1972); and in *Fundamental Interactions in Physics and Astrophysics*, p. 354, G. Iverson, Ed., Plenum Press, New York (1973).

121 Interview with Professor Joe Lannutti, January 30, 1986.

122 D., *General Theory of Relativity*, Wiley, New York (1975).

123 D., in *Proceedings of Loyola University Symposium, New Orleans, 1984*; reproduced in Kursunoglu and Wigner (ref. 1), p. 194.

124 N. F. Mott, *A life in science*, p. 42, Taylor and Francis, London (1986).

125 For a list see Kragh (ref. 1), p. 356, note 20.

126 D., *Chem. Zeitung* **95**, 880 (1971).

127 Quoted by Heisenberg in *Schritte und Grenzen*, Piper, Munich (1971).

127a Manci Dirac, letter to A. Pais, November 25, 1995.

128 G. Gamow, *Thirty Years that shook physics*, p. 121, Doubleday, New York (1966).

129 L. Alvarez, *Adventures of a physicist*, p. 87, Basic Books, New York (1987).

130 R. Peierls, in ref. 31, p. 36.

131 *Washington Post* and *Times Herald*, September 24, 1954.

132 E. G. *New York Times*, May 27, June 11, 1954; *New York Herald Tribune*, May 28, 1954; *The Times* (London), June 18, 1954; *The Financial Times* (London), August 6, 1954.

133 *New York Times*, June 3, 1954.

134 D., *Proc. Camb. Phil. Soc.* **35**, 416 (1939).

135 Ref. 89, p. 393.

136 D., *The Development of Quantum Theory*, Gordon and Breach, New York (1971).

137 D., *Eureka* No. 32, 2–4 (October 1969).

138 P. Ramon, letter to A. Pais, February 22, 1996.

139 J. Lannutti, letter to A. Pais, May 19, 1986.

140 Quoted by R. Peierls in ref. 130.

2 Antimatter

MAURICE JACOB
CERN, Geneva

Physical laws should have mathematical beauty
P. A. M. Dirac, 1955

I already have gray hair but I belong to a generation which grew up in physics calculating Feynman graphs and using the CPT invariance of Quantum Field Theory. The world would look very different if we could reverse the flow of time (an operation denoted by T), inverse all directions in space (an operation denoted by P) and change all particles into their antiparticle (an operation denoted by C). Yet the laws of physics would remain the same and all phenomena would occur in the same way. Our present understanding of physics implies the existence of antimatter, and all the properties of antimatter are predictable from the known properties of matter. All this looked so powerful, so beautiful and almost so natural to us, as we were learning modern physics in the late 1950s and early 1960s. The two ways to read the same simple Feynman graph, using it to describe, for instance, either the exchange of a photon between two electrons, or electron–positron annihilation and formation through one photon, looked like an obvious part of the calculation rules. This is shown in Figure 2.1. One can read it

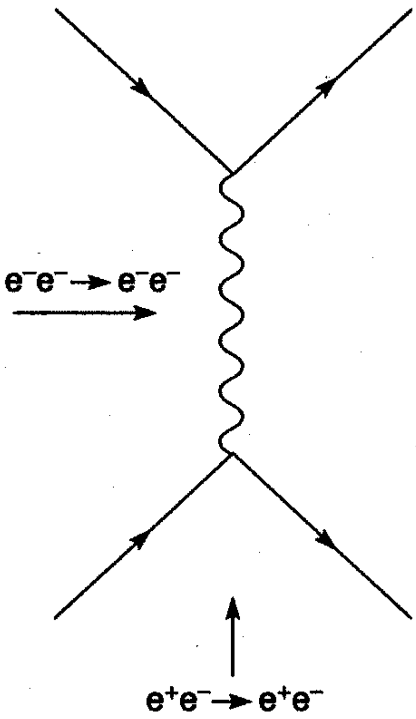


FIGURE 2.1

Feynman graph for lowest order electron–electron scattering (left–right) and for electron–positron annihilation and formation (down-up).

horizontally. This is scattering. One can also read it vertically. This is annihilation and pair formation.The same term can be used to describe both processes.

When the discovery of the antiproton was announced, in the mid-1950s, it was perceived by those of my generation more as an expected event than as a breakthrough. Antimatter had already lost its mysteries! To each known particle one has to associate an antiparticle with the same mass and the opposite internal quantum numbers -those not associated with kinematical properties- and the rules followed by the latter are fully determined by those followed by the former. Together with relativity and quantum mechanics, antimatter was part of the general framework in which to work when facing with enthusiasm the exciting perspectives of modern physics.

2.1. Antiparticles, the legacy of Dirac

The conception of antimatter

Conceiving antimatter was, however, not such an old achievement and, if I listed antimatter next to relativity and quantum mechanics, it is because Paul Dirac, in his masterful combination of quantum mechanics and relativity, came to the conclusion that it was an unavoidable necessity. This was in 1929–30. At that time it created a stir among physicists and the proton was even considered for a while as the candidate for the still elusive particle with a positive charge which had to be associated with the electron as its antiparticle. However, this idea had to be quickly disregarded. Indeed Dirac first considered that possibility, saying later: 'I just didn't dare to postulate a new particle at that stage, because the whole climate of opinion at that time was against new particles'. But, by 1931, it was clear that this was not tenable and he then summarized the situation, saying: 'This would be a new kind of particle, unknown to experimental physics, having the same mass and opposite charge as the electron. We may call such a particle an anti-electron'. And, he added at that time: 'We should not expect to find any of them in nature, on account of their rapid rate of recombination with electrons, but they would be produced experimentally and in high vacuum they would be quite stable and amenable to observation'. Dirac was right on all counts. The electron cannot exist without its positron counterpart. The discovery of the positron by Carl Anderson, in 1932, vindicated Dirac's electron theory. The positron, first seen in cosmic rays, by Anderson and soon afterward by Blackett and Occhialini, was there, as the anticipated antiparticle of the electron, with the opposite charge and the very same mass. Before long it was recognized as an active participant in the β decay of radioactive nuclei, together with Pauli's neutrino and its antineutrino. For instance the β decay of a neutron gives a proton, an electron and an antineutrino. I was fortunate to have Gian Carlo Wick as my thesis adviser. He was the one who showed that the β+ emission discovered by Joliot–Curie, whereby a positron is produced, was also included in Fermi Theory of β decay first formulated for the emission of electrons. One soon learned how to create large quantities of positrons, appearing in association with electrons in the interaction of radiation with matter. Radiation freely turns into matter and antimatter, and matter and antimatter freely annihilate into radiation.

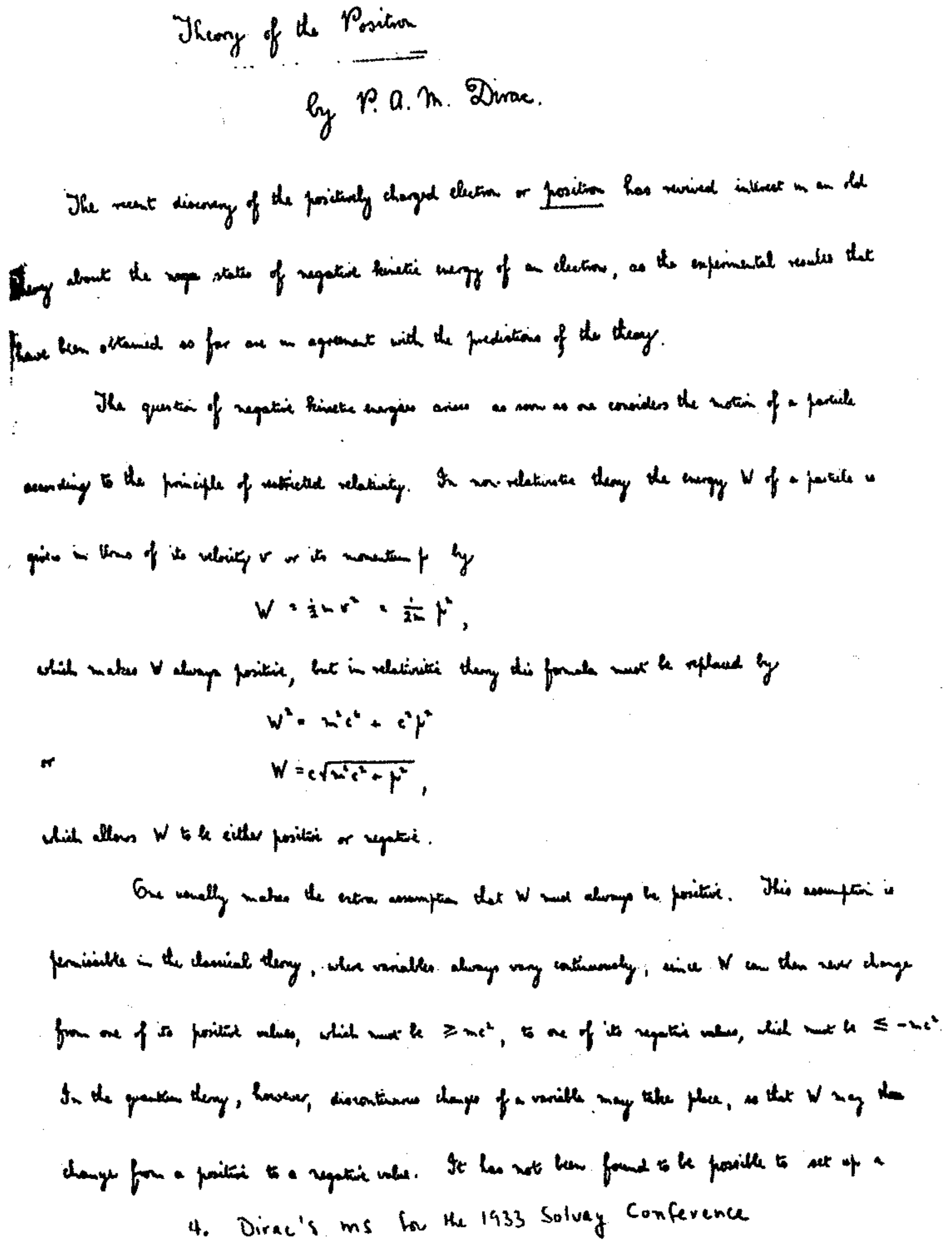
I wish to follow here the presentation given by Dirac at the 7th Solvay Council, in 1934. It was actually originally presented in a beautiful and precise French. It starts with a magnificent sentence:

> The recent discovery of the positively charged electron or positron has revived interest in an old theory about the states of negative kinetic energy of an electron, as the experimental results that have been obtained so far are in agreement with the predictions of the theory.

Figure 2.2 shows the first page of the draft (in English) which Dirac wrote for his Solvay article.

The theory was not that old!

The presentation goes on explaining that, in relativity, negative energies readily come into the picture since it is the square of the energy, together with the square of the momentum, which makes up an invariant. In classical physics, where energy varies in a continuous way, one can still separate the positive energy domain, where the full energy exceeds the mass energy, from the negative energy one, where the full energy is less than minus the mass energy, but, in quantum physics, jumps between the two domains are allowed and they can no longer be separated. Dirac goes on in his specific style saying: 'Under such circumstances two possibilities remain open: either there is a physical meaning for the negative energy states or we have to admit that the relativistic quantum



Theory of the Positron

by P. A. M. Dirac.

The recent discovery of the positively charged electron or positron has revived interest in an old theory about the [illegible] states of negative kinetic energy of an electron, as the experimental results that have been obtained so far are in agreement with the predictions of the theory.

The question of negative kinetic energies arises as soon as one considers the motion of a particle according to the principle of restricted relativity. In non-relativistic theory the energy W of a particle is given in terms of its velocity v or its momentum p by

$$W = \tfrac{1}{2}mv^2 = \frac{1}{2m}p^2,$$

which makes W always positive, but in relativistic theory this formula must be replaced by

$$W^2 = m^2c^4 + c^2p^2$$

or

$$W = c\sqrt{m^2c^2 + p^2},$$

which allows W to be either positive or negative.

One usually makes the extra assumption that W must always be positive. This assumption is permissible in the classical theory, where variables always vary continuously, since W can then never change from one of its positive values, which must be $\geq mc^2$, to one of its negative values, which must be $\leq -mc^2$. In the quantum theory, however, discontinuous changes of a variable may take place, so that W may then change from a positive to a negative value. It has not been found to be possible to set up a

4. Dirac's ms for the 1933 Solvay Conference

FIGURE 2.2

Dirac's draft (in English) for his presentation (in French) at the 1934 Solvay meeting.

theory is not correct'. He later goes on saying that 'A negative energy electron is an object foreign to our experience but, when considered in the framework of the electromagnetic theory, it behaves just as a positive energy electron having charge +e instead of −e. Yet one cannot identify it with a positron since positrons have positive energy'.

The solution which he then presents relies on the use of the exclusion principle of Pauli. He says

> Let us accept that in the Universe as we know it, almost all the negative energy states are occupied and that the resulting charge distribution is not detectable because of its homogeneity over space. In such a case any unoccupied state represents a disruption which breaks this uniformity. This appears as a hole and it is possible to admit that these holes are positrons. The exclusion principle of Pauli states that any dynamical state available to an electron can be occupied by at most one particle. An electron cannot therefore loose energy while falling into a lower energy state which is already occupied.

This resolves the difficulties associated with negative energies since a hole in the distribution of the negative energy electrons appears as having positive energy. The hole reacts to an electromagnetic field as a positively charge electron of positive energy and with the same mass as the electron.

The article goes on to discuss quantitatively all the consequences, including pair creation and electron-positron annihilation, comparing them with available experimental information, and in particular showing why a positron has a good chance to cross Anderson's chamber before annihilating against an electron. It continues with vacuum polarization effects.

The negative energy problem had been solved. A brilliant prediction had been made and verified. The price to pay was that the vacuum had become rather complicated. Following Dirac's approach the vacuum indeed behaves in many ways as a semiconductor, where electrons can be excited out of a filled valence band while leaving holes. This vacuum problem has now been solved but, as we shall see, other problems have come up. In any case, hole theory remained the standard way to calculate for many years and the complicated vacuum was not much of a philosophical problem.

In physics we are even getting used to complications of the vacuum. Today the vacuum of the electroweak theory behaves in many ways like a superconductor and so does the vacuum of chromodynamics, though in a different way. The electroweak theory and chromodynamics represent together our present understanding of the dynamics at the level of the fundamental particles. They globally constitute what is called the 'Standard Model'. The vacuum is defined as the lowest energy state of a system and has to be handled that way. Within our description of the dynamics it has a structure for which we find analogies in condensed matter physics.

From one problem to another

Pauli, in his 1945 Nobel lecture based on the exclusion principle, presented again the theory which Dirac had described earlier in his own Nobel lecture, explaining how Dirac could eliminate the problem of negative energies using the exclusion principle.

In his lecture Pauli describes Dirac theory where in the actual vacuum all the states of negative energy should be occupied and only deviations of this state of smallest energy, namely holes in the sea of these occupied states, are assumed to be observable and shows how it is the exclusion principle which guarantees the stability of the vacuum, in which all states of negative energy are occupied. He goes on to explain that the infinite 'zero charge' of the occupied states of negative energy is then formally analogous to the zero-point energy of the quantized one-valued fields, concluding that the former has no physical reality either and is not the source of an electromagnetic field.

Yet, at the end of his lecture, Pauli expresses his dissatisfaction and, following his own words, his 'critical opinion that a correct theory should neither lead to infinite zero-point energies nor to infinite zero charges, nor should it invent a "hypothetical world" which is only a mathematical fiction before it is able to formulate the correct interpretation of the actual world of physics'.

He then sets the goal very high saying that 'A theory should be established which will determine the value of the fine structure constant and will thus explain the atomistic nature of electricity'.

At present, the standard formulation of Quantum Electrodynamics brings back the vacuum to its state of expected emptiness while exhibiting perfect symmetry between matter and antimatter. Yet, it leaves the fine structure constant as a parameter.

I can but try to put in a nutshell the new features brought by the quantum field theory approach.

In quantum field theory the field which describes an electron is no longer a simple wave function but an operator which destroys and creates particles. It can excite and de-excite the states on which it acts. The 'negative' energies appear now as mere de-excitation energies and there is no longer anything puzzling about their appearance. The Dirac field destroys an electron and creates 'something', a particle of a new kind. Its adjoint creates that electron and destroys the same 'something'. The definition of a charge operator from the field and its adjoint, which, by definition, cannot change the charge, together with the more technical imposition of causality, shows that the 'something' must be a particle with the opposite charge and the same mass as the electron.

The quantum field approach has to give up the description of the electron as a single particle but it implies that one cannot describe the electron without describing also the positron. Matter and antimatter appear together and on the same footing while one deals only with the positive energy excitations of the most simple vacuum. The vacuum is back to emptiness. It contains neither electrons nor positrons.

Causality requires that the amplitude for emission of a particle is equal to the amplitude for the destruction of an antiparticle and vice versa. We call this crossing symmetry. Any process with an entering (emerging) particle is simply related to a process with an emerging (entering) antiparticle. This is illustrated by Figure 2.1.

The consequences of hole theory thus are more naturally expressed with a simple vacuum but the second quantized formalism is needed. This can be traced to the work of Majorana in 1937 but it took some time before the quantum field approach became part of the physicist's household.

The existence of antimatter, with this symmetry between matter and antimatter, is now seen as a direct consequence of the inner structure of quantum field theory. This is the formalism which combines relativity and quantum mechanics while requiring causality. The CPT symmetry of physics follows. Quantum electrodynamics turns out to be separately invariant under C (particle–antiparticle exchange), P (Parity) and T (Time reversal). Its formulation does not change when all particles are changed into antiparticles and vice versa. The system described may change but the equations which describe its dynamics remain the same!

Antimatter and causality

The reason for antiparticles was beautifully addressed by Feynman in his Dirac Memorial lecture of 1986 and one may at this stage look at a picture of Dirac and Feynman discussing physics (Figure 2.3). Feynman illustrated his talk with clear and simple examples from which he extracted brilliant generalizations. He considered in particular two successive scatterings of an electron in an external field, showing that, if only positive energy states are allowed for the intermediate virtual electron which is propagating between the two

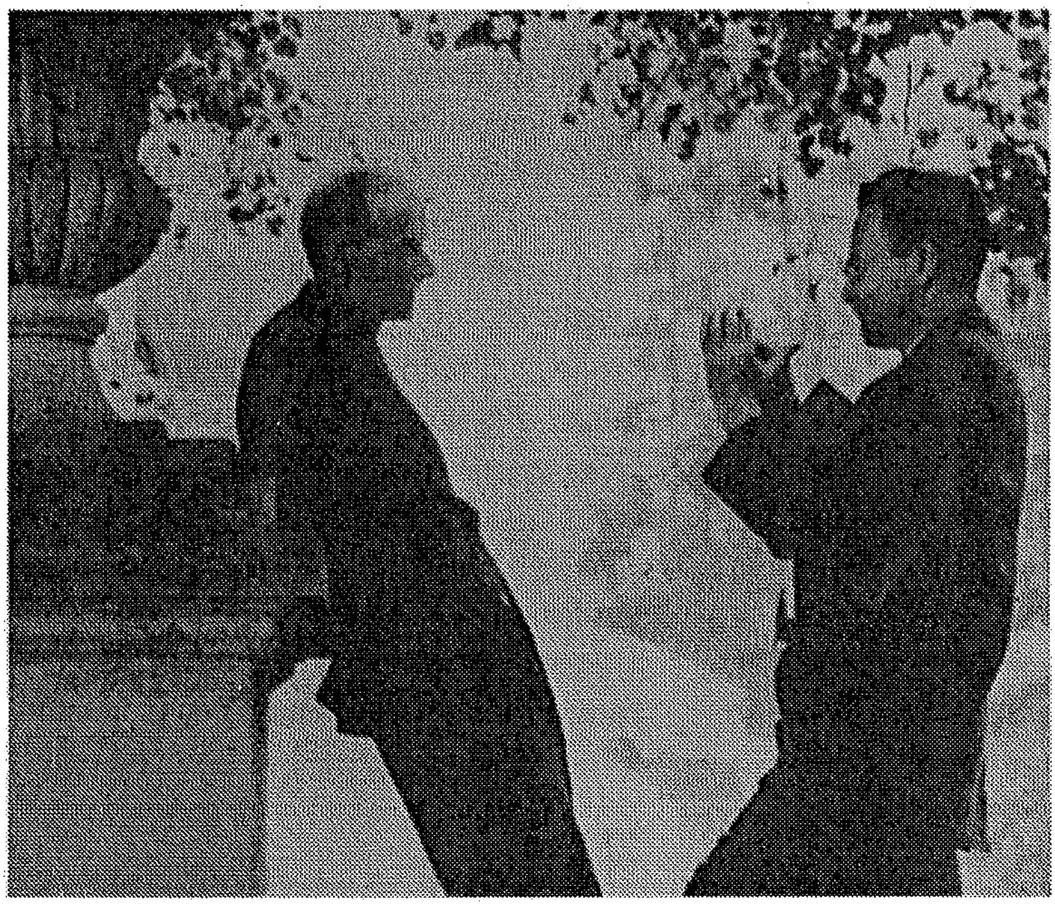


FIGURE 2.3
Dirac and Feynman discussing physics.

scatterings, the second event cannot be limited to the light cone of the first one. As a result, if one event is in the future of the other, in a particular system, it may appear as happening before it in another one. The consequence of relativity may then violate that of naïve causality, with the intermediate virtual particle now running backward in time. The presence of antiparticles is necessary to restore a causal structure to the process seen with the opposite sequence of time. The event now seen as appearing earlier is understood as a pair formation producing the final electron and, together with it, a positron. The positron moves forward in time, as it should, to annihilate the initial electron at the time now seen as coming later. As Feynman puts it 'the virtual particle of someone may be the virtual antiparticle of someone else'. Antiparticles appear as needed to maintain an apparent causal structure whatever the reference frame in which one describes the event may be. At a more general level, one may say that antimatter is the way nature enforces causality in a relativistic and quantum world. This is encoded in quantum field theory.

Feynman in his original approach to quantum electrodynamics in the late 1940s had introduced this new view of antimatter, first brought up by Wheeler, where positrons appear as electrons running backward in time. The quantum and relativistic description of the evolution of an electron between two events has indeed to sum over many paths including those for which the proper time appears for a while to run backward. At that stage the electron appears as a positron. Positrons have to be there because such configurations are needed. Positrons have to exist as bona fide particles.

Feynman proposed several metaphors for the appearance of the positron. One of them is that of a bombardier in a plane which follows a road at low altitude and who suddenly sees the road becoming apparently three roads, only to realize, looking at things more widely, that this was only a switch back on the first road. This is illustrated by Figure 2.4 which shows two versions of the same double scattering event. In the first case an electron 'travels' between the two scatterings. In the second case pair formation is followed by annihilation with a positron and two electrons 'travelling' in between. Including the positron is necessary to calculate amplitudes associated with the motion of electrons. Within a process, they look like electrons running backward in time. In Feynman's original conception, a vacuum filled with negative energy electrons was no longer needed to perform a calculation. Yet the expert knows that an important relative minus sign in his approach can be related to the exclusion principle in hole theory.

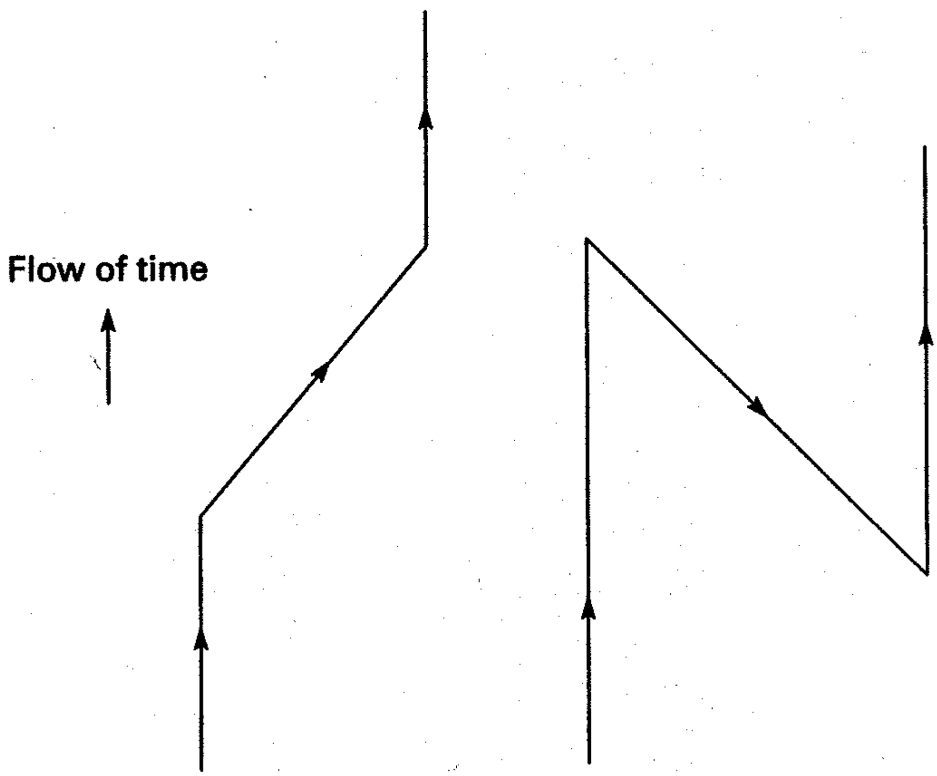


FIGURE 2.4

Double scattering in an external field seen in two different reference frames. Left: Two scatterings. Right: Pair formation and annihilation.

Theoretical physics as seen by Dirac

If the presence and the properties of antiparticles are now well understood, we can still reflect with admiration upon Dirac's achievement and try to benefit as much as possible from his masterful approach to physics. Didn't he arrive at antimatter because, as he said in his 1977 Varenna lectures: 'One must be prepared to follow up the consequences of theory, and feel that one just has to accept the consequences no matter where they lead'.

I would thus like to continue quoting from him as he describes the work of the theoretical physicists. This I shall borrow from the talk he gave during the ICTP Conference of 1968 organized by Abdus Salam on the theme 'From a life of physics', in order to express at the time, as Salam had put it later 'The sense of gratitude and adulation which everyone felt towards the great men of Physics still amongst us'. I was fortunate to attend this conference. Let us read what Dirac said:

> I shall attempt to give you some idea of how a theoretical physicist works – how he sets about trying to get a better understanding of the laws of nature.
>
> One can distinguish between two main procedures for a theoretical physicist. One of them is to work from the experimental basis. For this, one must keep in close touch with the experimental physicists. One reads about all the results they obtain and tries to fit them into a comprehensive and satisfying scheme.
>
> The other procedure is to work from the mathematical basis. One examines and criticizes the existing theory. One tries to pinpoint the faults in it and then tries to remove them. The difficulty here is to remove the faults without destroying the very great successes of the existing theory.
>
> There are the two general procedures, but of course the distinction between them is not hard-and-fast. There are all grades of procedure between the extremes.

He then goes on to say:

> With the mathematical procedure there are two main methods that one may follow, (i) to remove inconsistencies and (ii) to unite theories that were previously disjoint.

Then he hints at one of his successful methods:

> I would like to mention that I found the best ideas usually came, not when one was actively striving for them, but when one was in a more relaxed state. . . . I used to take long solitary walks on Sundays, during which I tended to review the current situation in a leisurely way. Such occasions often proved fruitful, even though,

> (or perhaps because,) the primary purpose of the walk was relaxation and not research.

This is in particular how he tried to reconcile Relativity and Quantum Mechanics facing the difficulties with what became known as the Klein–Gordon equation. As he said, 'Tensors were inadequate and one had to get away from them, introducing two-valued quantities, now called spinors. Those people who were too familiar with tensors were not fitted to get away from them and think up something more general, and I was able to do so only because I was more attached to the general principles of quantum mechanics than to tensors'.

He then comes to his electron theory and the appearance of antimatter, saying:

> The introduction of spinors provided a relativistic theory in agreement with the general principles of quantum mechanics, and also accounted for the spin of the electron, although this was not the original intention of the work. But then a new problem appeared, that of negative energies. The theory gives symmetry between positive and negative energies, while only positive energies occur in nature.

Solving this new problem brought up antimatter.
Indeed, as Dirac later said in the same talk:

> As frequently happens with the mathematical procedure in research, the solving of one difficulty leads to another. You may think that no real progress is then made, but this is not so, because the second difficulty was really there all the time, and was only brought into prominence by the removal of the first.
>
> This was the case with the negative energy difficulty. All relativistic theories give symmetry between positive and negative energies, but previously this difficulty had been overshadowed by more crude imperfections in the theory.

> The difficulty is removed by the assumption that in the vacuum all the negative energy states are filled. One is then lead to a theory of positrons together with electrons. Our knowledge is thereby advanced one stage, but again a new difficulty appears, this connected with the interaction between an electron and the electromagnetic field.

This is the difficulty met with divergences which was to always bother him. These divergences, or infinite quantities, are stumbling problems met in calculations going beyond the simplest processes.

The parameters of mass and charge associated with the electron in the formalism of electrodynamics are not yet quantities measured under ordinary conditions. A free electron is accompanied by an electromagnetic field which effectively alters the inertia of the system, and an electromagnetic field is accompanied by a current of electron–positron pairs which effectively alters the strength of the field and of all charges. Hence a process of renormalization must be carried out, in which the initial parameters are eliminated in favor of those with immediate physical significance.

Dirac recognized the power of renormalization theory describing it as a permitted change of the starting equations. But he said:

> You may think that the work of the theoretical physicist is easy if he can make any starting assumptions he likes, but the difficulty arises because he needs the same starting assumptions for all applications of the theory. This very strongly restricts his freedom. Renormalization is permitted because it is a simple change which can be applied universally whenever one has charged particles interacting with the electromagnetic field. The present quantum electrodynamics does not conform to the high standard of mathematical beauty that one would expect for a fundamental physical theory, and leads one to suspect that a drastic alteration of basic ideas is still needed.

We have gone a long way since in the exploration of the structure of matter. Yet it is impressive to see how the basic entities used in electrodynamics have been essentially merely generalized as new particles and new interactions have been discovered and studied. Quarks behave as electrons. It is said that after attending a lecture by Gell-Mann on quarks, Dirac, who had remained silent, eventually told Gell-Mann 'You know, I believe in quarks'. 'Wonderful', said Gell-Mann, overjoyed, 'but what is your main reason ?' 'This is because they have spin ½', was Dirac's answer.

We have so far discussed antimatter in connection with the description of spin ½ particles. In quantum field theory, half integer and integer spin fields are, however, treated on a similar footing using respectively anti-commutation and commutation rules. To each particle, whether a fermion or a boson, we are lead to associate an antiparticle with the same mass (a consequence of CPT symmetry) and the opposite internal quantum numbers, those which are not related to kinemetical properties.

2.2 Antimatter and present particle physics

Bound states of particles and antiparticles

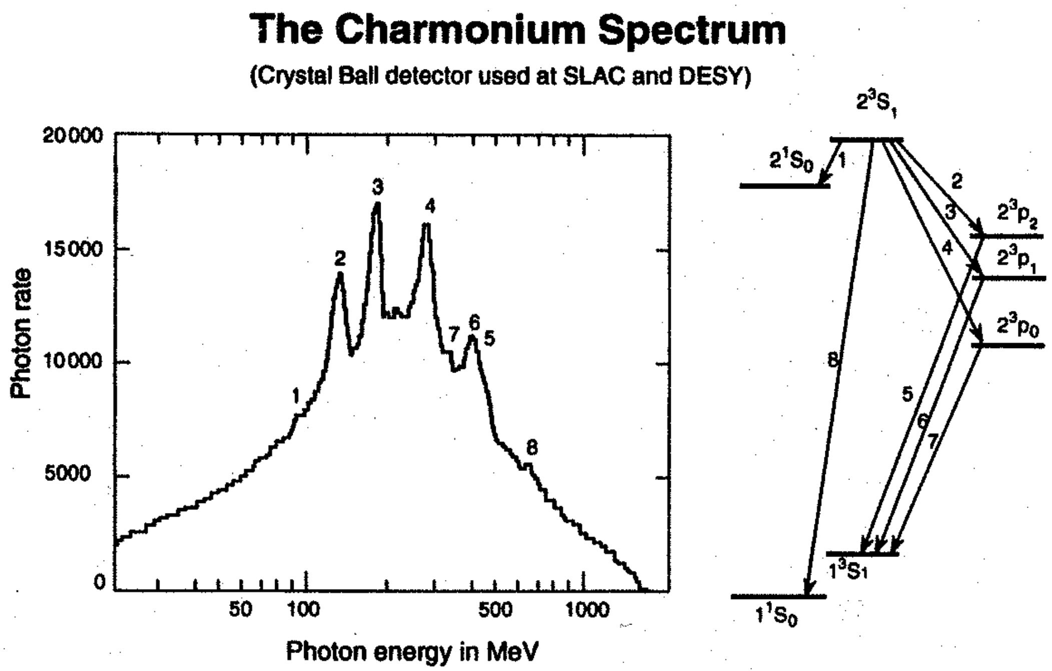
There is a usual feeling that matter and antimatter when brought together result in a violent explosion. Indeed, when considering protons and antiprotons, we have an energy per particle which is close to 200 times that available in a hydrogen bomb! Yet things are not always that violent. If chunks of matter and antimatter were to start to annihilate in large amounts, one would quickly have between the two a cushion of high pressure radiation which would slow down the process. The correct metaphor is that of droplets of water thrown on a hot stove. They run around a lot before evaporating, being protected from the intense heat by a cushion of vapor.

At present, when studying production and annihilation processes, we deal with antiparticles almost one by one. When meeting a particle, an antiparticle often makes a bound state with a decently long lifetime before annihilating. The system cascades down many levels before annihilation takes place. Positronium and muonium are examples which have been very extensively studied. Worth mentioning in connection with present day particle physics is charmonium, built with a charm-quark and its anti-charmed antiquark. The quark is here sufficiently heavy that one can follow the dynamics of the system in a non-relativistic way, predicting the energy levels in a potential which is coulombic at short distances and linear (confining) at larger ones. One can calculate the electromagnetic transition rates between the levels. This is the 'Hydrogen atom problem' at the quark level, namely a problem complicated enough to teach us something and yet simple enough to be handled in all details. In the case of the charmonium, the system is very tiny. Energy levels are separated by hundreds of Megaelectronvolts, not the mere electronvolts met with the positronium but the physics is very similar. There is hyperfine splitting as in usual atomic physics. The annihilation process is not so very fast by particle physics standards and can even be neglected in the calculation of the energy levels. Figure 2.5 shows the spectrum of photon emission from charmonium and the energy levels of the system.

More generally, all known mesons are seen as quark–antiquark systems and this has a great calculation value, as first shown by Dalitz. The quark and the antiquark may belong to different species. Changing the quark into its antiquark, and vice versa, we get the antiparticle of the meson.

Particle production

Trying to understand the deep structure of matter, we probe the structure and interactions of particles which may first appear as



The Charmonium Spectrum

(Crystal Ball detector used at SLAC and DESY)

FIGURE 2.5

The spectrum of charmonium and the energy levels of the systems formed by a charmed quark and its antiquark.

elementary by colliding them against each other with energy as high as we can get. Present usual collision energies (in the hundred of gigaelectronvolt range) are much greater than the mass energy of many known particles (the proton mass energy is of the order of 1 GeV). Nothing then forbids part of the incident kinetic energy to turn into mass energy and in general a high energy collision is associated with an abundant production of particles together with as many antiparticles. In the late 1940s, the energy of the Berkeley Bevatron had been chosen to make the production of proton–antiproton pairs possible and this is the way the antiproton was discovered.The appearance of antimatter has become bread and butter in the high energy collisions of present day particle physics. But it is not only its appearance which has become prominent, it is also its use!

Most of the produced particles and antiparticles are unstable but some of them are stable. In the vacuum, a positron is as stable as an electron and an antiproton is as stable as a proton. This is implied by CPT symmetry. One can then consider producing beams of positrons or beams of antiprotons, which can then be stirred and accelerated in an accelerator as it is the case for beams of electrons or beams of protons. A very good accelerator beam puts together hundreds of billions times fewer particles than the Avogadro number which gives us the order of magnitude for the number of particles in a gram of matter. Even though antiparticles are hard to get, we can now easily collect enough antiparticles from high energy collisions to make decent beams.

Particle–antiparticle colliders

Having a beam of antiparticles has an important advantage. A beam of positrons can be fed into an accelerator together with a beam of electrons. The machine can keep the two beams circulating in opposite directions and accelerate them at the same time in the same vacuum pipe. One can thus transform a particle accelerator into a particle–antiparticle collider. This has been an extremely important development in particle physics.

First came electron–positron colliders. Positrons are indeed easy to get in great numbers. Electron and positron bunches are accelerated and, as they cruise in opposite directions in the machine, they come into collision in certain areas where they can mutually annihilate, their full energy being available for a wide array of processes. The collision energy partly turns into matter and antimatter energy and more and more particles can be produced in a single collision as the energy increases. The great crop of data and results obtained at SPEAR, at SLAC, the Stanford Linear Accelerator Center in California, with the discovery of charm and of the tau lepton at a collision energy not much in

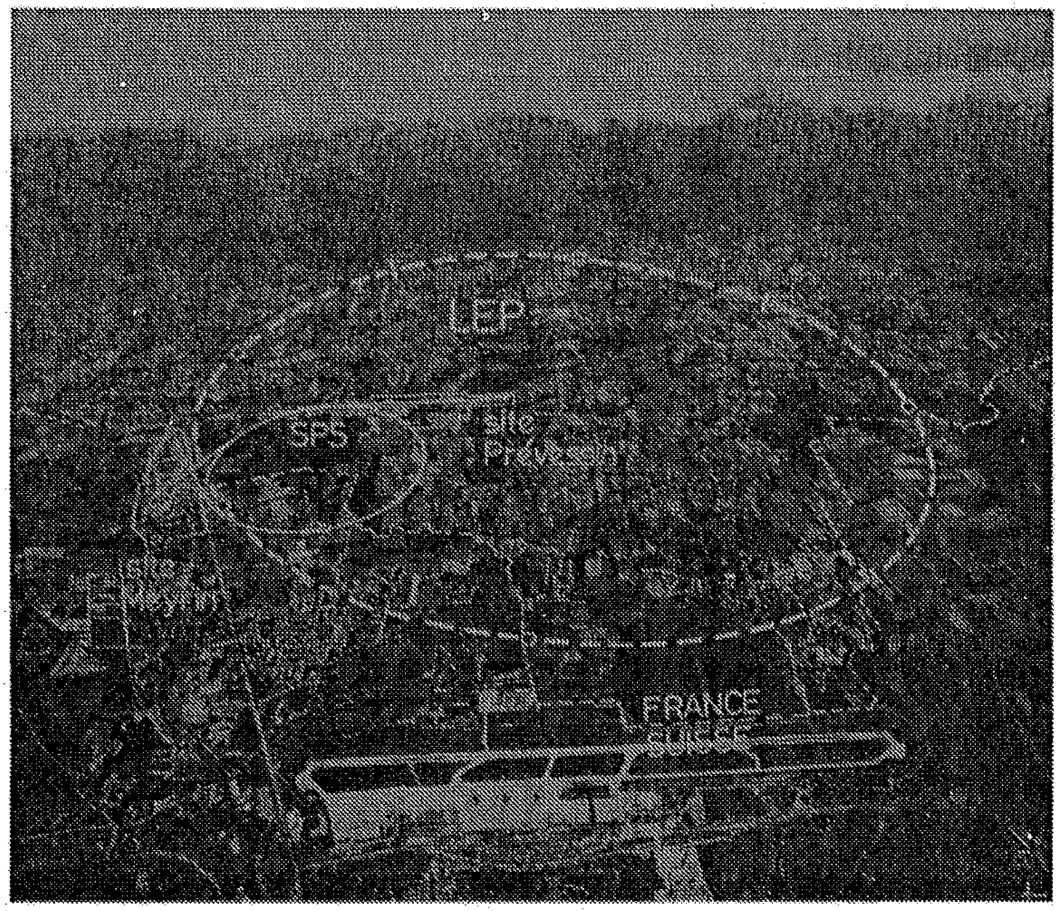


FIGURE 2.6
The LEP ring at CERN. The machine is in a tunnel about 100 m underground. It collides electrons and positrons accelerated in opposite directions in the same vacuum pipe. Also seen (to the left) is the smaller ring of the SPS which became the first proton-antiproton collider.
(Photo CERN)

excess of 4 GeV, has been at the origin of a brilliant series of circular machines with increasing energy. They are CESR at Cornell, PETRA at DESY, the German laboratory in Hamburg, PEP at SLAC, TRISTAN at KEK, the Japanese laboratory near Tokyo, culminating with LEP at CERN , the European centre in Geneva. Figure 2.6 shows the location of the LEP ring, which is about 100 meters underground, on the CERN site. LEP, with its present total collision energy of 100 GeV, now soon to be doubled, is the ideal machine to test in detail the consequences of the electroweak theory. Electroweak theory has so far come out of all tests with flying colors. So did chromodynamics, which can also be submitted to many tests at these collision energies.

A circular machine has a great advantage. Bunches cross many, many times since they coast at almost the speed of light and one can collect many events even if the probability to get one event in a bunch–bunch collision is very small. A particle bunch represents indeed an already good vacuum!

However, LEP will be the largest circular machine ever built. With higher energies, the synchrotron radiation of the accelerated beams gives a practically insurmountable problem trying to feed back to the coasting particles the energy which they radiate on every turn. On the other hand, linear colliders, whereby bunches of electrons and positrons are accelerated on two separate straight lines, do not give this problem. They should be the answer but the accelerated bunches now collide only once and they have to be made enormously small to enhance the probability of collisions between particles. At present very high energy linear colliders are considered on many laboratories' drawing boards, with pieces on test benches. This should eventually be the way to extend electron–positron collisions to much larger energies, and one of them, the SLC at SLAC, already works at 100 GeV of collision energy. In this ingenious device, electrons and positrons are separately accelerated in the same machine and brought into head on collision after bending through two separate arcs.

Antiprotons are harder to get since a proton–antiproton pair is about 2000 times more massive than an electron–positron pair. One needs a very high energy collision to have a decent chance to produce one. They have to be preciously collected and stored before a decent beam can be obtained. However, because of the heavy mass there is now no problem with synchrotron radiation. One can accelerate antiprotons to very high energy in a circular machine together with protons circulating in the opposite direction.

CERN, with the talent and drive of Carlo Rubbia and Simon van der Meer, was the pioneer in that field. Antiprotons could be collected, stored and cooled into a good beam, in a dedicated machine, and then accelerated together with protons in the SPS. Figure 2.7 shows the CERN antiproton accumulator where antiprotons are captured, cooled and accumulated. One can see the wave guides set across the machine. They transmit in a straight line signals associated with particles straying away too much from the mean energy so that this can be corrected after they have travelled along the corresponding arc at almost the speed of light. Antiprotons thus stacked and cooled to a circling beam of well defined energy are fed into the PS, accelerated and transferred to the SPS where they are further accelerated together with a beam of protons circling in the opposite direction. Collision energy of 600 GeV could be achieved between protons and antiprotons. This was necessary to produce the W and Z which were thus discovered, in 1983. The W and the Z are the carriers of the weak force in the electroweak theory. For instance, in neutron β decay, the neutron emits a (virtual) W as it transforms itself into a proton. The W fragments into an electron and an antineutrino. The mass energy of the W is about 80 times that of the proton. In β decay, it can appear only as a quantum fluctuation of very short time duration. With enough collision energy it can emerge as a bona fide particle. The same applies to the Z which has a mass energy of about 90 times that of the proton.

Protons (and antiprotons) are complicated objects. They contain quarks, antiquarks and gluons which all take part in the collisions. One may say that they correspond to broad band beams of quarks, antiquarks and gluons. As the energy increases, proton–antiproton and proton–proton collisions look more and

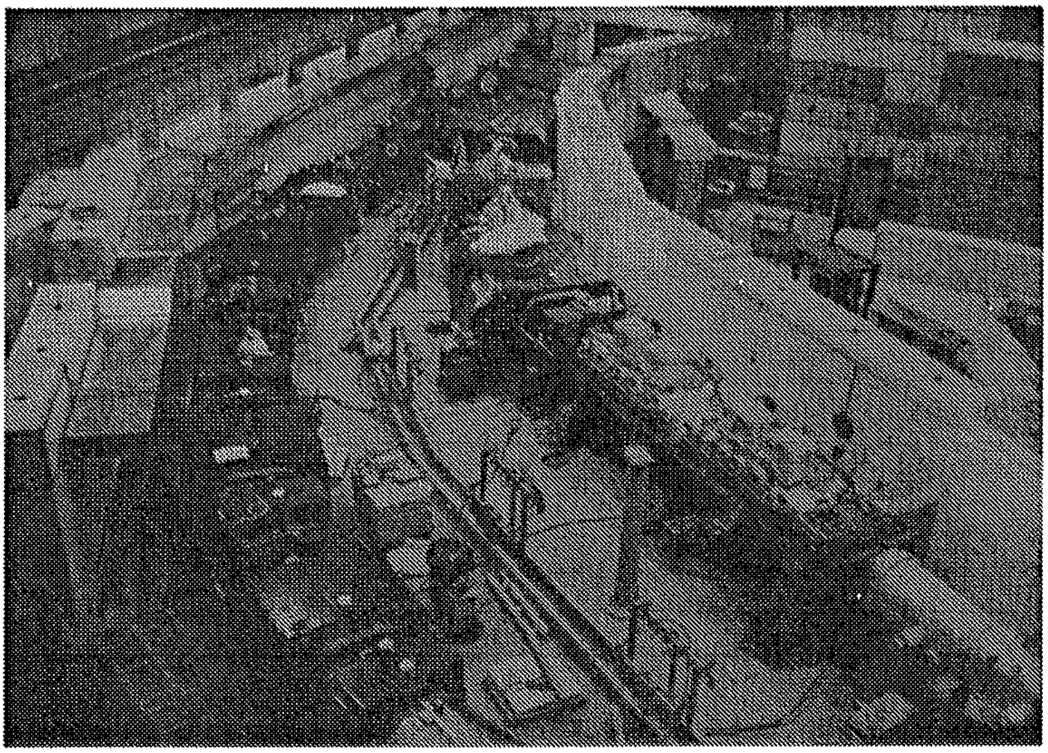


FIGURE 2.7

The CERN antiproton accumulator. One sees the lines set across the machine which transmit information about the beam as it circles around and control the cooling devices.

(Photo CERN)

more alike. The simple annihilation process of the colliding particles, so important in electron–positron encounters with its simple one photon process, loses its special prominence. What matters primarily is thus to reach very high energies while using a single accelerator already built to accelerate protons. The CERN proton–antiproton mode of operation was a great success in the 1980s. It has now been phased out. At present, Fermilab has a proton-antiproton collider of a much higher energy of 2000 GeV. This large energy was needed for the discovery of the very massive top quark, which is produced in top–antitop pairs (a mass energy of the order of 350 GeV).

Whereas one now produces intense beams of antiprotons, they are typically two to three orders of magnitude less dense compared to what one can do with protons. Since the luminosity of the machine, which controls the reaction rate, is also a very important parameter when searching for rare but most interesting processes, the new CERN collider, the LHC, with 14 000 GeV of collision energy, will be a proton–proton collider. One needs two separate beam pipes and magnetic structures but this is the accepted price to pay for luminosity.

One may illustrate the power of a proton–antiproton collider with two examples. One of them (Figure 2.8) is the production of a Z in quark–antiquark annihilation, with its subsequent decay into an electron–positron pair. This is the way the Z was discovered in 1983. The other one (Figure 2.9) is the production of two hadronic jets as point-like constituents within the colliding proton and antiproton (quark, antiquark or gluon) individually collide to give two particles shot at wide angle and which eventually appear as jets of hadrons. This is the modern aspect of the Rutherford experiment, giving evidence for hard point-like scatterers within the colliding particles.

The scattered constituents are ‘colored’. In chromodynamics the ‘color’ of a quark takes the role of the charge in electrodynamics. The scattered constituents cannot escape into the vacuum which is opaque to color. The penetration energy is of the order of 1 GeV/fermi and they thus don’t go very far before part of their energy turns into light hadrons and antihadrons, mainly π mesons, which emerge as a jet replacing the original particle. We do not see quarks and antiquarks but we see jets which are almost as spectacular.

Dealing with antimatter

With electron–positron colliders and proton–antiproton colliders, available antimatter has quickly become a very important tool in

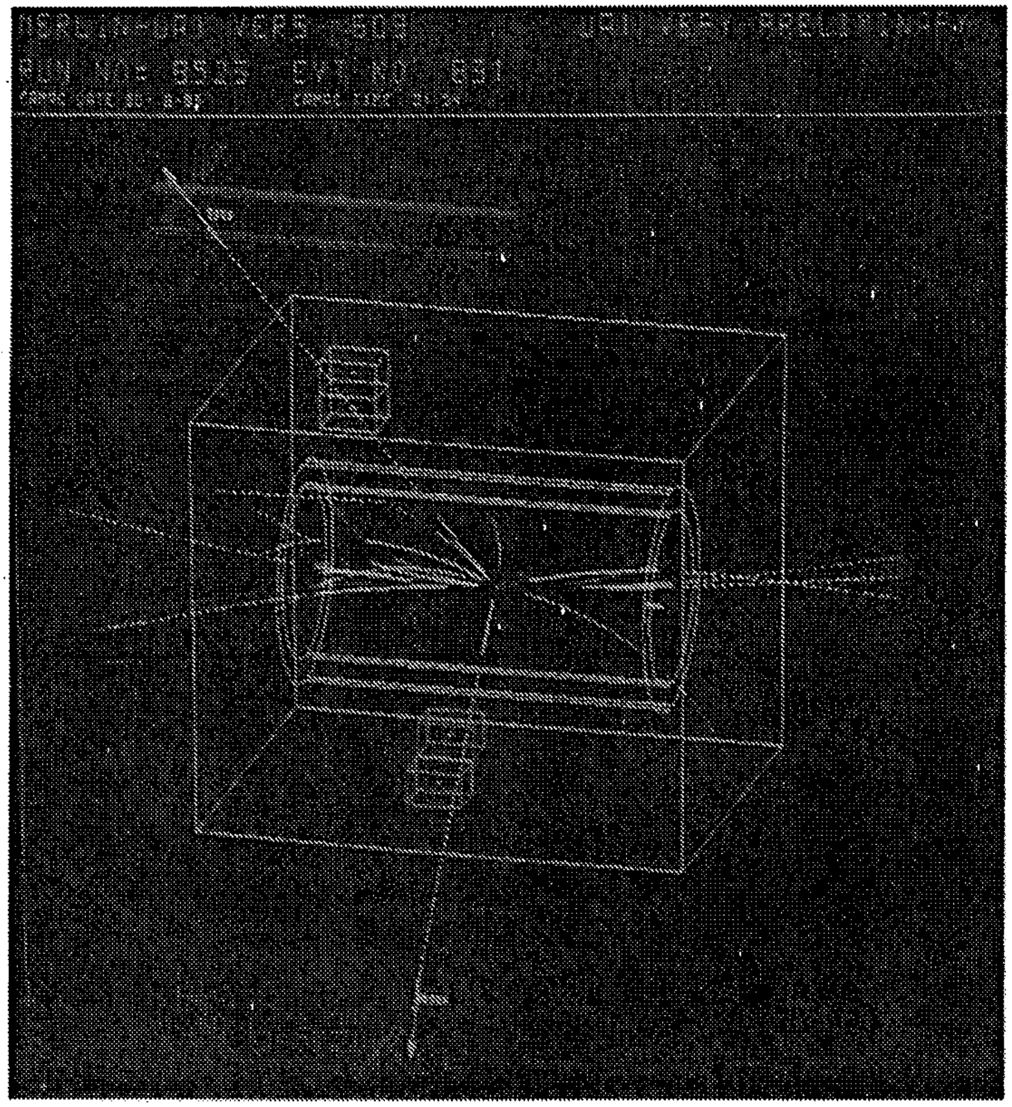


FIGURE 2.8

Production of a Z particle during a proton–antiproton collision (UA1 detector). The Z decays into an electron–positron pair which provides a clear signature.

(Photo CERN)

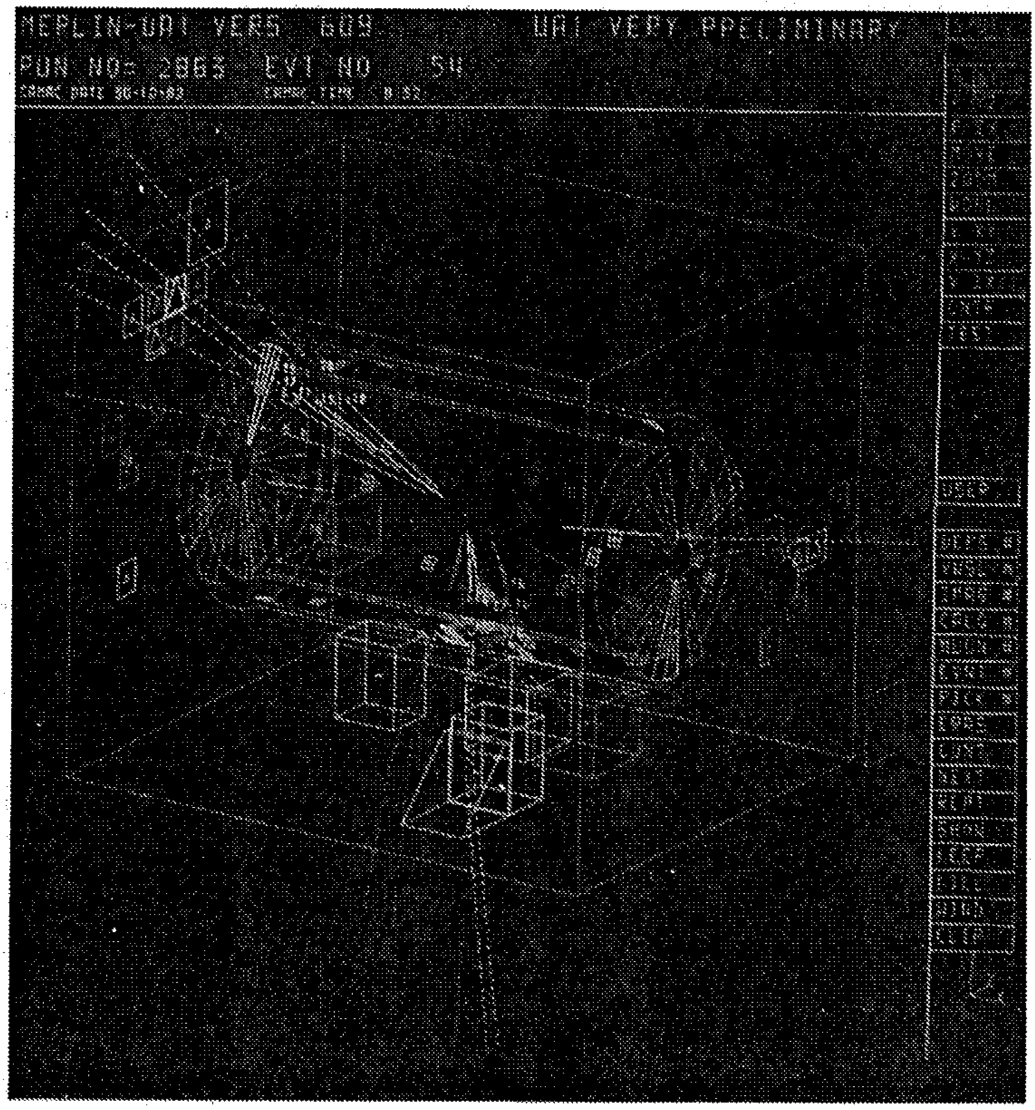


FIGURE 2.9

Production of two hadronic jets during a proton–antiproton collision (UA1 detector). This is the modern form of the Rutherford experiment whereby point-like constituents within the colliding particle hit each other head on and recoil at a wide angle.

(Photo CERN)

particle physics. There is, however, not only the high energy frontier. Worth studying also are specific features associated with the annihilation of proton and antiprotons at low energy, whereby a relatively large amount of mass energy turns into particles and antiparticles, allowing a detailed study of the spectroscopy of the objects formed. At CERN, the stored and cooled antiprotons can also be decelerated and stored in a dedicated machine, LEAR, the Low Energy Antiproton Ring. The machine is shown in Figure 2.10. It provides an intense source of antiprotons which can be extracted at will and only by themselves. This allows many scattering and annihilation studies, looking in particular for the properties of the many types of particles which are produced .

There are many other exciting things which one can do with an intense source of slow antiprotons. They can be captured in atoms where they replace an electron and orbit with specific atomic levels. One can study the spectroscopy of such compact antiprotonic atoms. Worth a special mention are the CERN recent results on antiproton helium atoms. One can also capture very low energy antiprotons in a trap. The present practical realization is offered by the trap invented by Hans Dehmelt. In the present Penning trap, the kinetic energy can be brought down to a thousandth of an electronvolt through collisions with electrons and the antiprotons can be held trapped in a magnetic field circling in a small 'bottle' for months. Figure 2.11 gives a schematic drawing of a Penning trap in use at CERN. It is 13 cm long. Within the trap, one can compare the motion of antiprotons in a magnetic field to that of protons. This is the best ever test of the expected identity between the proton and antiproton mass. The precision achieved is at the level of 10^{-9}. Another LEAR experiment is attempting to compare the gravitational pull on protons and antiprotons.

The next step would be to make real 'full' antimatter, making antihydrogen atoms from antiprotons and positrons. An antiproton would capture a positron created together with an electron in

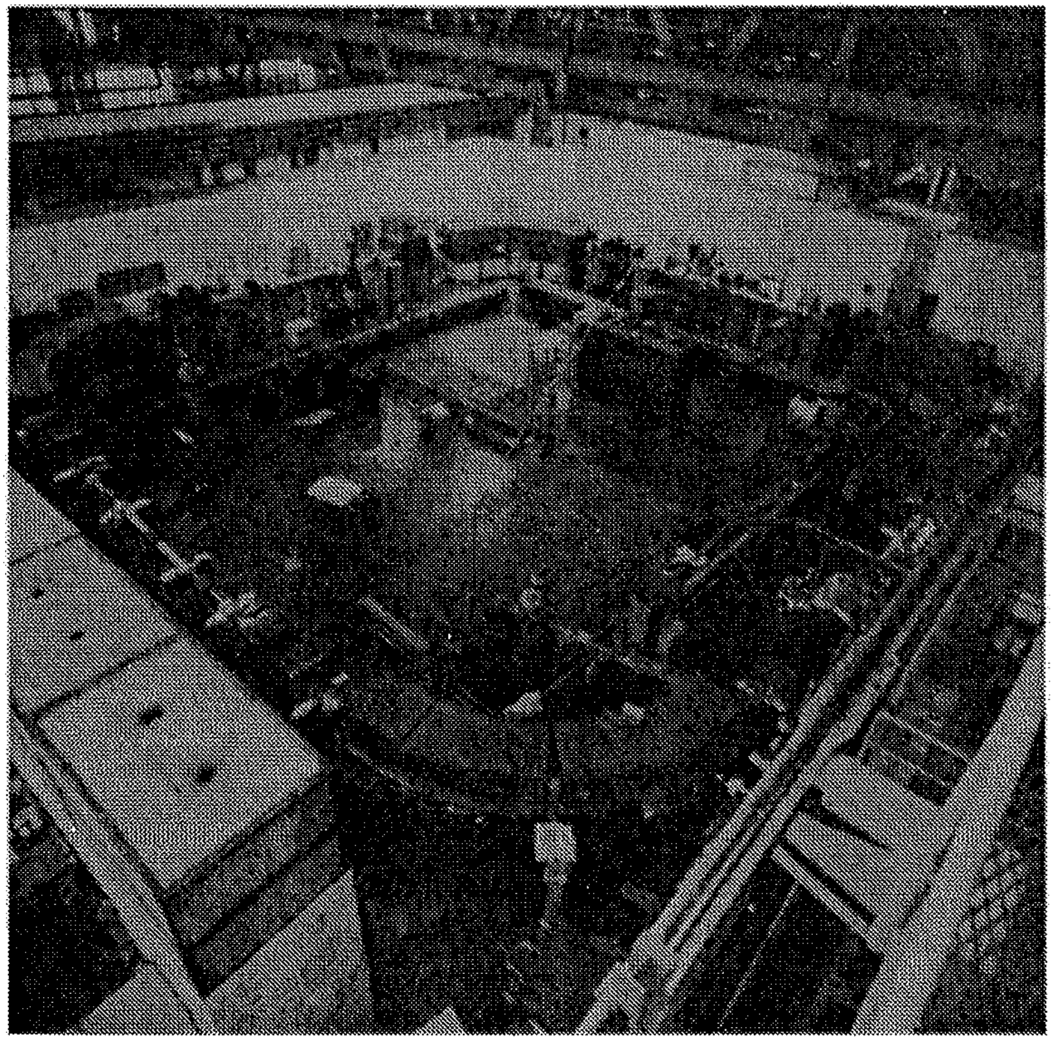


FIGURE 2.10

LEAR – the low energy antiproton ring, which stores a large quantity of antiprotons decelerated to low energies. They are held circling in a magnet ring and extracted for experiments.

(Photo CERN)

its collision with a heavy nucleus. Whereas this may provide evidence for antihydrogen, reaching decently large quantities would require combination between antiprotons and positrons stored in traps. This now seems possible but still looks a few years away with present techniques.

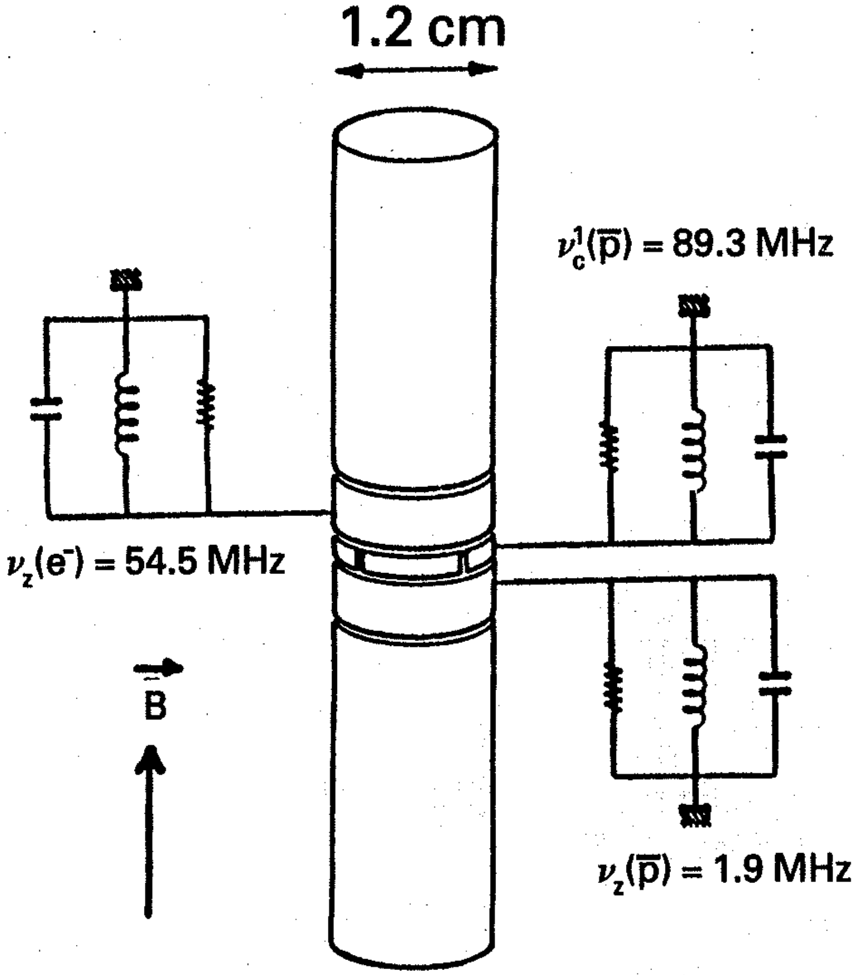

1.2 cm
$\nu_c^1(\overline{p}) = 89.3$ MHz
$\nu_z(e^-) = 54.5$ MHz
$\vec{B}$
$\nu_z(\overline{p}) = 1.9$ MHz

FIGURE 2.11

A schematic drawing of the Penning trap used in the LEAR experiment PS 196. In the trap, 13 cm long, antiprotons are cooled via repeated collisions with electrons. Antiprotons with energy 10 billion times lower than those in LEAR can be stored and studied over long periods of time (months), in a small apparatus.

We can do very exciting physics with tiny quantities of antiparticles, and in particular build accelerator beams. We would need a thousand billion accelerator bunches of antiprotons to reach one gram. The present price tag for antiprotons has been estimated at about £10^{16} per gram. Yet, as we shall see, antiprotons, even in relatively small numbers, have already found some uses outside of particle physics, and many more may be to come.

Particle–antiparticle oscillations

A very special study of antimatter in high energy physics is worth singling out. It is the analysis of the neutral K system, soon to be followed now by the analysis of the neutral B system, where Beauty replaces Strangeness (or is associated with Strangeness) in the meson structure. Strangeness and Beauty are two among the internal quantum numbers often referred to, which turn into their opposite when going from a particle to its antiparticle.

We saw how mesons are built of a quark and an antiquark. The neutral K meson is built out of a d-quark (charge −1/3) and a strange antiquark, which has antistrangeness (and charge +1/3). It is globally neutral but different from its antiparticle which is built out of an anti d-quark and a strange quark. It has the opposite strangeness. Strangeness is conserved in the production process which involves other final particles so that, depending on the event considered, one knows that either a K meson or an anti K meson has been produced. The meson flies off. Its decay is indeed mediated by the weak interaction, which violates strangeness conservation and 'takes some appreciable time' to act. As the meson flies, the weak interaction can also eliminate the strangeness from a quark which is transformed into a quark without strangeness. At the same time, it can put antistrangeness on an antiquark which did not have it in the first place. As a result a neutral K meson is turned into its antiparticle (the anti K meson) or vice versa. The two eigenstates of the mass matrix, which correspond to a specific evolution of the wave function with time (damped by decay at a particular rate) will therefore be mixtures of the K and anti K states. The probability of seeing the particle either as a K meson, or as an anti K meson oscillates with time according to the (tiny)

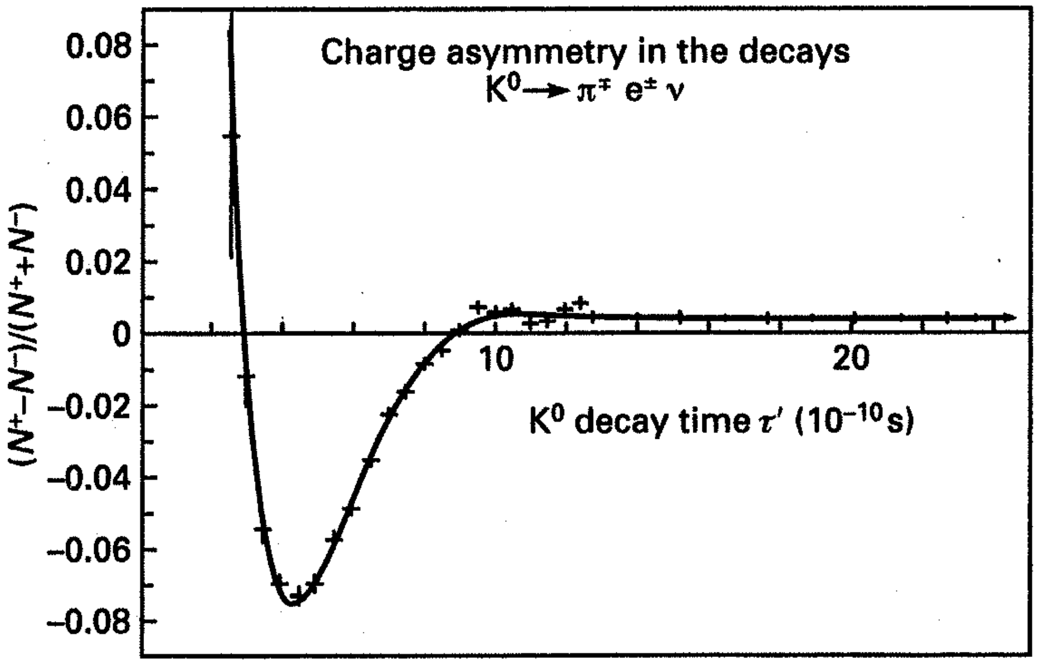


FIGURE 2.12

Oscillation between a neutral K meson and its antiparticle. This is seen through the appearance of either an electron or a positron among the decay particles in the electron (positron), pion antineutrino (neutrino) mode. The K° gives only a positron whereas the anti K° gives only an electron. The observed electron positron charged asymmetry as a function of time follows the oscillation between the K meson and its antiparticle. The left over asymmetry at longer times bears witness to the violation of CP symmetry. This decay mode can thus be used to define through the description of an experimental result what we call a particle (the electron) as opposed to its antiparticle (the positron).

mass difference between the two eigenstates of the mass matrix. The particle and the antiparticle continuously exchange their role. This is illustrated by Figure 2.12 which shows the positron (K decay) over electron (anti K decay) excess as a function of time.

Things are particularly simple in the neutral K meson case since the two eigenstates of the mass matrix have very different life times. One of them quickly disappears while the other one remains about a thousand times longer. The first one is almost an eigenstate of CP with value +1 and a favored decay into two π mesons. The second one is almost the CP eigenstate with value −1 for which the easy two π mode is forbidden. Not quite though, and the small admixture between the two CP eigenstates shows that CP invariance is violated in neutral K decay. This came as a big surprise in 1964. This is still a puzzle today and the experimental appearance of the violation of CP invariance is still limited to neutral K decay. It is possible that this could however be a natural feature in the Standard Model with six quark species, that is something which has no specific reason not to appear. It is therefore extremely tempting to collect similar evidence in other cases. The study of neutral B decay, where a b-quark is associated with either an anti d-quark or an anti strange quark (and the other way around for the anti B meson) looks particularly promising in the Standard Model. What we said about the neutral K system, with its oscillation between particle and antiparticle also holds for each of the neutral B systems. However, things are experimentally more complicated since the two eigenstates of the mass matrix now correspond to practically identical decay rates and remain on the same footing as the meson flies off and eventually decays. One can give special attention to decay modes which are CP eigenstates but they are now rather rare. Nevertheless, granting enough properly tagged B (and anti B) mesons, the detailed study of the evolution with time should be possible. Here, however, comes the next problem. The B meson being rather heavy, it is hard to produce and it is difficult to collect big enough a sample. This has motivated the construction of a dedicated b-factory at SLAC at Stanford and also at KEK in Japan, the construction of a special detector at HERA at DESY and it has been at the origin of a special proposal for the LHC. We may hope that the detailed study of neutral B decay will soon bring a new and valuable light to CP violation.

In any case CP violation is there and this could be at the origin of the excess of matter over antimatter in the early universe. In the explosive condition of the Big Bang, CPT symmetry could not restore an asymmetry between matter and antimatter brought by CP violation, as emphasized long ago by Zakharov.

2.3. Antimatter at the cosmic scale

Physics presents a beautiful symmetry between matter and antimatter, and what could be more natural than a universe where matter and antimatter would be both equally present, even if, in our surroundings, we have to happily acknowledge that matter is overwhelming. However, probing the cosmos we see no sign of the expected effects associated with a large amount of matter coming into contact with a large amount of antimatter with a large amount of radiation of specific signature. This even holds up to the super cluster level, which is at present the ultimate grouping scale. It seems that all the universe which we can see is made of matter. This actually tallies with our view of the universe originating from a Big Bang with densities and temperatures which are larger, the closer one tries to get to the beginning, extrapolating back in time from our present information. At the beginning of the universe the temperature falls as the inverse square root of its age. The density falls as the inverse square of its age. The physics which takes place is the one which we explore with particle physics. With the energy at LEP (100 GeV) we have the collision conditions which prevailed when the universe was 10^{-10} s old.

The universe looks very quiet when observed with visible light. But, when looked at through radio waves or X-rays and gamma rays, it is rich in violent events in which antimatter (in any case positrons) comes readily into the picture.

Two events during the Big Bang

I shall here limit myself to singling out two periods in the early universe for which matter-antimatter symmetry is particularly relevant. They are selected for their being very important and of different kind.

The first period was when the universe was about 1 s old. This is the time when the temperature fell below 1 MeV. Up to that time electrons and positrons were continuously annihilating into two photons, but two photon collisions could produce at an equal rate electron and positron pairs. There was therefore an equilibrium between electrons, positrons and photons, which were practically equally numerous in the universe. When the temperature fell below 1 MeV, as the universe was expanding and cooling, electrons and positrons could still annihilate into photons but the photons soon did not have enough energy to make electron–positron pairs. The massacre of electrons and positrons was no longer compensated by a continuous production process. This is illustrated by Figure 2.13. All positrons annihilated against electrons. There remained only one electron survivor in a billion. Photons became by far the most numerous particles in the universe. This was the end of the so-called lepton era when electrons and positrons dominated.

The second period was when the universe was about 10 μs old, when the temperature was about 200 MeV. In the framework of quantum chromodynamics we expect that, at such a temperature, the vacuum becomes no longer transparent to the 'color' of the quarks. Up to that time the universe was a plasma of quarks, antiquarks and gluons which were freely roaming and crashing into each other. When the universe became opaque to color, quarks and antiquarks had to bind into globally colorless hadrons (protons, antiprotons and π mesons for instance) since only such particles could exist in the new vacuum. But the rest mass energy of the

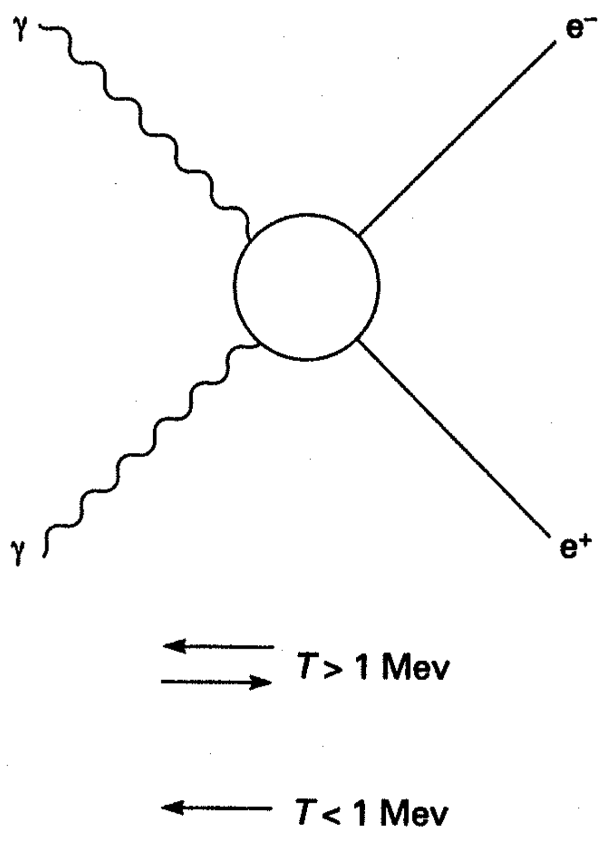


FIGURE 2.13

Electron–positron annihilation into two photons and pair formation in photon–photon collisions.

protons and antiprotons (about 1 GeV) which were created was already much larger than the surrounding temperature (200 MeV). Protons and antiprotons quickly annihilated against each other since the density was very high and they could not be formed again in pairs, through the collisions of the surrounding particles. There was therefore a massacre of protons and antiprotons. All antiprotons annihilate with protons, leaving only one proton in a billion as the survivors. This was the end of the quark era during which quarks and antiquarks had been the most abundant particles in the universe. These two periods are visible in Figure 2.14 which pre-

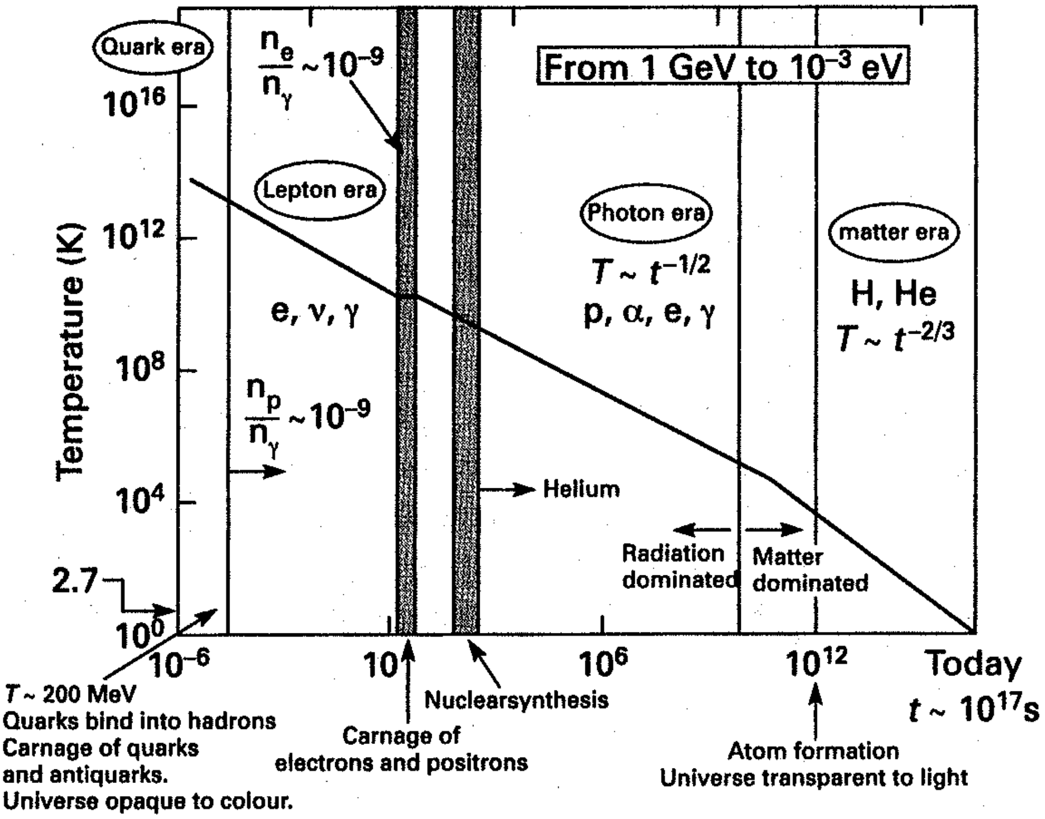


FIGURE 2.14

A piece of the thermal history of the universe showing in particular the two events which occurred when the temperature went through 200 MeV (carnage of quarks and antiquarks) and 1 MeV (carnage of electrons and positrons) respectively.

sents the thermal evolution of the universe from a microsecond to the present day.

One may then wonder why the number of surviving protons appears to be equal to that of the surviving electrons, both being a billion times less than the photons in the cosmic radiation background. Attempting to answer this question would lead us to CP violation in a Grand Unified Theory. Whereas the symmetry present in such a theory could have prevailed in the very early universe, its eventual breaking as the universe cooled down would have produced a differentiation between quarks and leptons appearing separately but with the proper relative numbers. This corresponds, however, to an energy domain where theoretical ideas are not yet supported by experimental information.

In the very early universe, matter and antimatter were almost on the same footing. This is certain from the overwhelming abundance of photons over protons and electrons in our present universe. Antimatter is likely to have fully disappeared from the cosmic scene, through the two successive massacres which we described. This is the prevailing view. Yet one can still entertain views in which antimatter could still prevail in some corners of the universe. This motivates searching for antinuclei in cosmic rays.

2.4. Other practical uses of antimatter

Next to the prominent use in making accelerator beams, other more practical applications are already worth noting. Positron beams are used in condensed matter and atomic physics but medical applications are worth a special description.

Positron tomography

Positrons are easy to make and they have already been put to an efficient medical use. Positron Emission Tomography (PET) uses positrons which originate from neutron deficient radioactive nuclei. The annihilation of the positrons against electrons produces pairs of back to back gamma rays of a well defined energy which can be detected in coincidence for a rather precise localization of the emitter. The positron mean free path is of the order of 1mm. There are positron emitters such as Carbon-11, Nitrogen-13, Oxygen-15 or Fluorine-18, which are easily made parts of biological substances used as tracers. The detection of their whereabouts inside the human body can be used not only to localize anomalies within specific organs but also to study biochemical changes as they take place. Pathological developments can thus be spotted long before any anatomical changes are detectable. The most commonly used radioactive isotopes, like those mentioned, have mean lives which correspond to typically 10 min. They have therefore to be made on the spot. This is done with dedicated cyclotrons which are now available as compact user friendly tools. The PET camera detecting the gamma rays provides an *in vivo* measurement of the localization of the tracer as a function of time. One gains over the long (and still much) used isotope imaging, both in sensitivity and in spatial resolution.

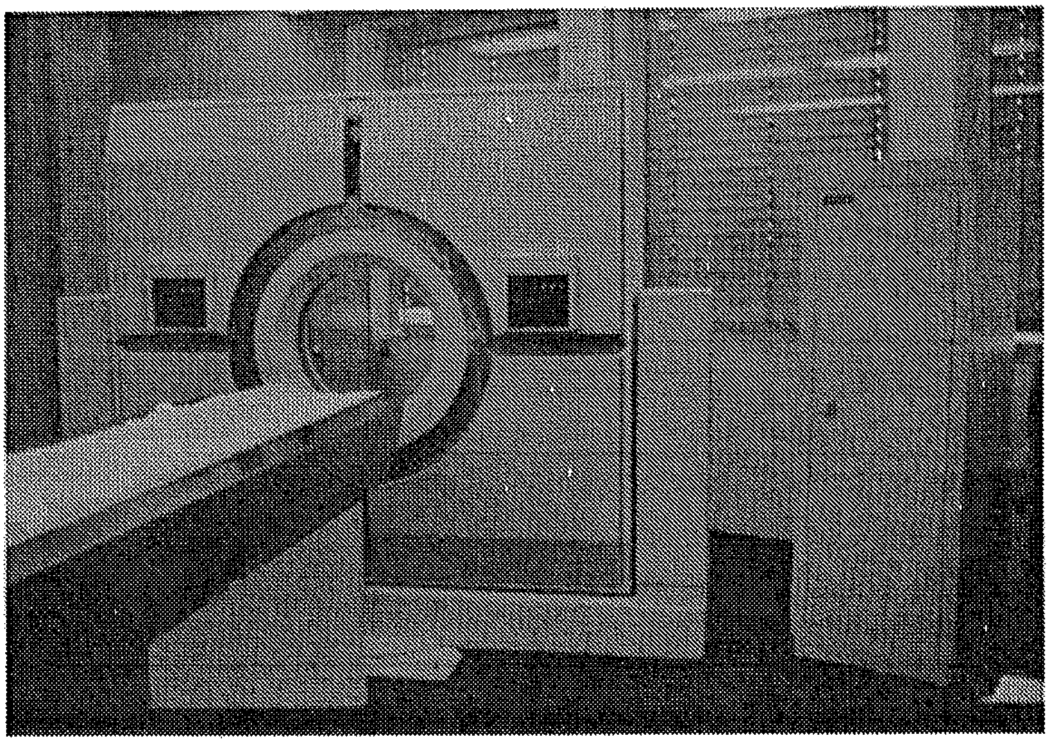
The whole equipment is now available in a highly automated form, suitable for hospital use. The PET scanner assembled at CERN and operational at Geneva's university hospital is shown in Figure 2.15.

At present there are already 140 PET centres in the world and their number is increasing at the level of close to 20 per year. There are many applications in the fields of oncology (tumor detection), neurology and cardiology. Clinical use is expected to grow rapidly.

Next to this beautiful medical use of positrons, one may now venture in the still speculative use of antiprotons.

The economics of antiprotons

We have seen how antiprotons can be produced in high energy collisions, collected and stored. Things are still not very efficient. The produced antiprotons are often lost when one tries to focus them with magnetic lenses and capture them in an accelerator ring. The best capture efficiency achieved so far is of the order of 1%. However, once antiprotons have been stored and cooled, being all brought into a beam of well defined low energy, handling them from one machine to an other is already possible with 90% efficiency. Present antiproton 'bottles' have been brought to hold



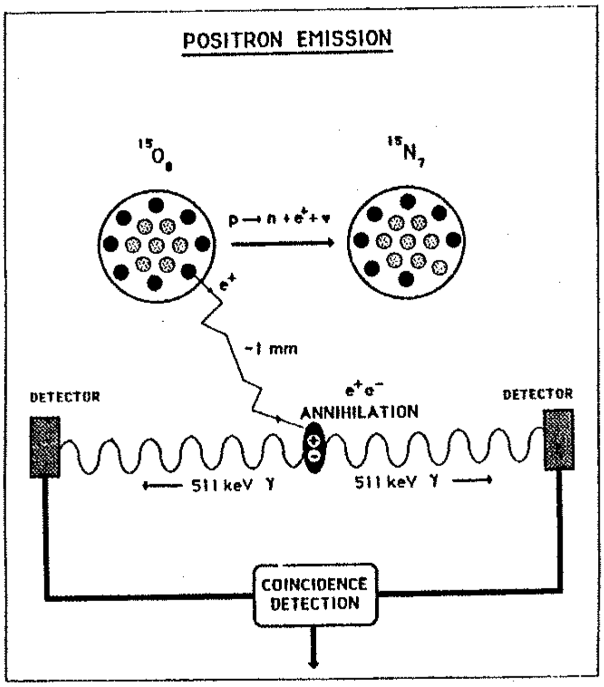


FIGURE 2.15

Positron emission tomography: the PET scanner assembled at CERN and operational at Geneva's University Hospital. Hundreds of patients have been scanned with this equipment (upper part of figure). The principle of the apparatus (lower part of figure): the annihilation of a positron with an electron produces two 511 KeV back to back photons and detecting them measures the position of the event immediately. *(Photo CERN)*

10^{12} antiprotons. This is still of the order of 10^{-12} g only. The energy associated with their annihilation would only be of the order of a few hundred Joule. The overall energy efficiency is therefore still of the order of one part in a hundred million.

A small magnetic trap can, however, be used for holding such an amount of antiprotons. We have already mentioned the Penning trap. It has been used to hold 10^5 antiprotons for months and holding the full load of 10^{12} appears possible. It is only 1 m long and the antiprotons are circling in the vacuum, away from the walls, in the strong magnetic field produced by superconducting magnets.

Antihydrogen has still to be produced but, once it has been (see note added), it can then also be stored despite its neutrality, using its diamagnetic property. If enough is produced, it could even be stored in dense solid form as antihydrogen-ice. This works for hydrogen-ice.

At present, it seems that antiprotons could be used for tomography in much the same way as proton beams have been used instead of X-rays, following the pioneering work of G. Charpak. Detection systems for charged particles, such as wire chambers, can be made very efficient and the probing intensity can thus be brought to a very low level with a very limited radiation exposure of the patient. With proton tomography one has, however, to use relatively rare wide-angle scattering events in order to locate with precision the target having been hit. The irradiation intensity has to be adjusted accordingly. But low angle scattering hits remain potentially harmful while providing little information. On the contrary, with antiprotons, each hit can be precisely located through an annihilation producing several particles independently detected. One can then much reduce the irradiation level. A good three-dimensional image of an organ could be obtained with a mere billion of antiprotons. The point where heavy particles come to rest and make the most damage can be relatively accurately defined and antiprotons, then used with higher density, could also be used to eliminate the tumors which they would allow one to detect. Efficient medical use could probably proceed with the tiny quantities presently available from high energy machines.

Looking to the more distant future one could be more speculative and consider rocket engines based on antimatter, but we are now talking about milligrams of antimatter (a billion times what is presently typically available) to a hundred milligrams depending on the ambition of the flight. The annihilation of antimatter would be used in heating a radiator at very high temperatures (3000 K). The propellant fuel would be shot through the radiator and exhaust at high velocity. With such high temperatures and exhaust velocities, a spacecraft may work with a much smaller quantity of fuel (or fuel to payload ratio) than is presently the case with conventional fuels.

Tiny quantities of antimatter (a thousandth of a billionth of a gram) are already very much in use. One can wonder with awe at what could be achieved when only milligrams would be in use!

But, besides considering these practical applications, at present and in the future, we can but admire Dirac's great achievement at predicting the existence of antimatter. In the 1960s, Heisenberg characterized the postulation of antimatter by Dirac 'as the most decisive discovery in connection with the properties or the nature of elementary particles'. The existence of antimatter is probably the revolutionary concept of physics in this century which had the strongest impact on the general public.

Note added.

At the beginning of 1996, the results of a group led by W. Oelert and M. Macri, working on LEAR at CERN, were announced. Antihydrogen had been formed. Nine antihydrogen atoms could be detected. In the experiment, low energy antiprotons traverse a Xenon jet. The electron–positron pairs produced by the scattering of the antiproton in the electric field of the heavy nucleus allow, a tiny part of the time, the capture of the positron by the antiproton. The neutral anti-atom escapes the binding of the guiding magnetic field of LEAR and flies off. Outside LEAR, a first detector signals in an unambiguous way the presence of a stripped positron in the neutral system. A spectrometer and a time of flight device signal in coincidence that the other element is an antiproton, before it annihilates. This is not the most efficient way to produce antihydrogen but this great discovery now fuels enthusiasm to do it with a much higher output ratio and to capture it, using antiproton and positron traps. The next step is to capture antihydrogen atoms standing almost still in a trap. Laser spectroscopy would then allow a comparison between hydrogen and antihydrogen levels (1s–2s) at the precision of 10^{-18}. This would provide a very valuable test of C and CPT symmetry.

3 The monopole

DAVID I. OLIVE
Department of Physics, University of Wales Swansea

It is a great privilege to speak to you on this occasion in which we commemorate Paul Dirac. In common with many of my friends in the audience, I enjoyed the good fortune of hearing the lectures on quantum mechanics delivered by him in Cambridge. Actually I was doubly fortunate as, in my year, 1959–60, he added a second course, extending beyond the material in his famous book.

Not only did we learn quantum mechanics as never before, but, very gently, we were shown a standard of logical presentation and clarity, indeed an aesthetic of logic, that was unforgettable. I believe I can say that this experience has affected many of us deeply and provided us with an ideal to which we struggle to aspire in our own research and teaching.

In Dirac's hands the beauty of mathematical logic and rational argument was not merely a tool for establishing sound proofs. Rather it could be a weapon of discovery that could lead to the most unexpected yet perfectly valid conclusions, which, once understood and assimilated, were unassailable in their beauty and rightness. The importance of this is that it is the discovery of the unexpected truth that changes the direction of scientific development, in both theoretical and experimental work. Dirac repeatedly showed how implacable yet beautiful logic could achieve these ends, as the other speakers also demonstrate.

In later life, 1977[1], he explained his attitude in a particularly vivid way when describing his personal affinity with Erwin Schrödinger (with whom he shared the Nobel prize):

> . . . Schrödinger and I both had a very strong appreciation of mathematical beauty, and this appreciation of mathematical beauty dominated all our work. It was a sort of act of faith with us that any equations which describe fundamental laws of Nature must have great mathematical beauty in them. It was like a religion with us. It was a very profitable religion to hold, and can be considered the basis of much of our success.

It has to be admitted that in most hands and at some times in the development of science this approach can be dangerous. Indeed even Dirac could be human and go wrong as he readily admitted in the same article:

> I think you can see here the effects of an engineering training. I just wanted to get results quickly, results which I felt one could have some confidence in, even though they did not follow from strict logic, and I was using the mathematics of engineers, rather than rigorous mathematics . . .
>
> It was perhaps the most suitable attitude to take for a quick development of the theory, but it did lead me to make mistakes.

He goes on to describe in detail some of the mistakes he made in setting up his general formulation of quantum mechanics. With hindsight these mistakes are rather shocking. Nevertheless they were eliminated, and what we learn from this is to appreciate another of Dirac's qualities, his transparent honesty, which informs all his writings. Besides this, his lucidity and his aesthetic sense, he also possessed another crucial quality, fearlessness. Dirac often voiced his worries concerning the fear generated by discovery. In the same article he states:

> You may wonder why I did not go on to consider higher approximations, but the reason is that I was really scared to do so. I was afraid . . . the results might not come out right. The originator of a new idea is always rather scared that some developments may happen which will kill it, while an independent person can proceed without fear, and can venture more boldly into new domains.

In his recent book[2], Murray Gell-Mann has reported that, when he asked Dirac why he had not predicted the positron immediately, the short reply was 'pure cowardice'. Nevertheless it is as a supremely intrepid, honest and logical creator of a new science that we honour Dirac.

Perhaps the single contribution that best illustrates Dirac's fearlessness is his investigation of the magnetic monopole. It certainly required courage to initiate a theory of an undetected particle. Because of what Dirac found, that theory has continued to intrigue researchers and continues to develop, as I shall explain. Now it seems that we can explain the invisibility of the monopole while unravelling another mystery, that whereby quarks, the undoubted constituents of nuclear particles, remain confined inside and unable to escape. The story of the magnetic monopole, nearly 65 years on, is still far from complete and indeed promises more revelations.

Magnetic fields are easy to observe. Everyone is familiar with the magnetic field of a bar magnet, as depicted in Figure 3.1.

The lines of force curve round to connect the two magnetic poles, north and south. The earth itself is a large magnet whose field can be detected and exploited by compasses to aid navigation. The magnetic field will be denoted by $\boldsymbol{B}$.

Although similar, electric fields occur differently in nature in the sense that it is possible to charge up individual objects separately so

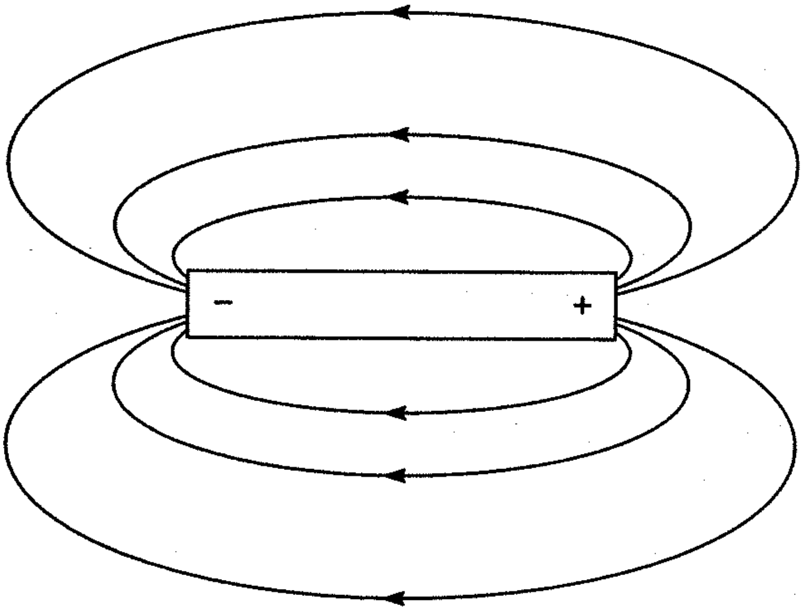


FIGURE 3.1

The magnetic field of a bar magnet, ***B***.

that they carry positive and negative charges respectively. If these are equal and opposite, we can produce an electric field, ***E***, similar to the magnetic field of the bar magnet as shown in Figure 3.2.

This distinction extends to the atomic scale and smaller. Electrons and protons carry electric charges which are equal and opposite in value while all other observed elementary particles carry the same values of the electric charge, or, if not, integer multiples thereof, but never any magnetic charge. This regularity, known as the 'quantization' of electric charge, is one of the most obvious and striking of the patterns seen amongst the elementary particles, and one of the most persistently difficult to explain.

Despite these facts, there is a tantalizing resemblance between the electric and magnetic fields and the way in which they behave. This was made more precise once Maxwell succeeded in formulating his celebrated equations governing their behaviour. This similarity is inherent in his equations and intimately connected to the success of the modern technologies based on them, such as

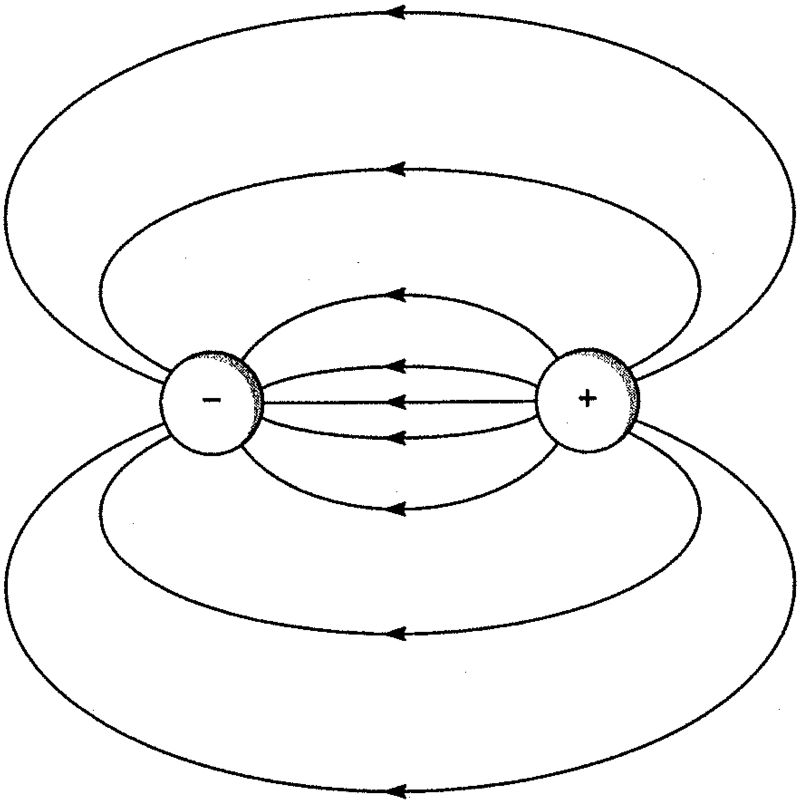


FIGURE 3.2

The electric field of two charged bodies, *E*.

electromagnetic power generation and radio wave transmission. The resemblances can be summarized by saying that the governing equations are unaffected if we make the interchanges

$$\boldsymbol{E} \to \boldsymbol{B} \text{ and } \boldsymbol{B} \to -\boldsymbol{E},$$

or better, if we replace the fields by rotated versions

$$\boldsymbol{E} \to \boldsymbol{E}\cos\theta + \boldsymbol{B}\sin\theta,$$
$$\boldsymbol{B} \to -\boldsymbol{E}\sin\theta + \boldsymbol{B}\cos\theta,$$

Thus the previous interchanges correspond to a rotation through a right angle. Physically observable quantities such as the density of energy and of momentum are indeed unchanged by these substitutions, at least if we forget the question of charge. Taking both elec-

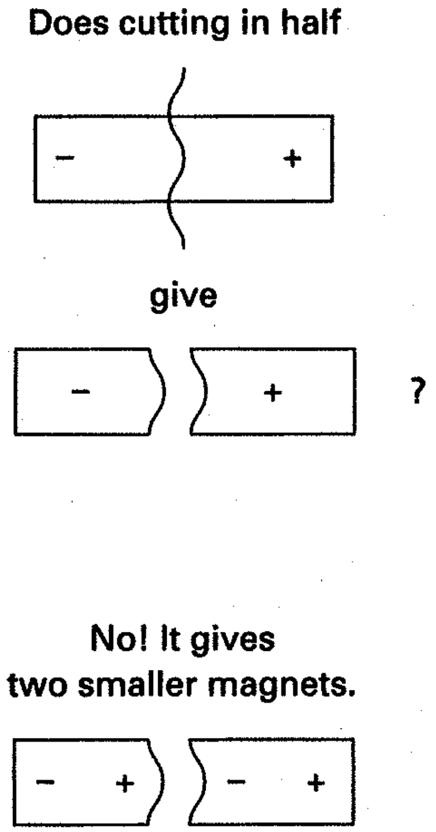


FIGURE 3.3

tric and magnetic charges into account, it is straightforward to write down perfectly satisfactory equations which include them and preserve the rotational symmetry, providing we also agree to the simultaneous replacement

$$q \to q\cos\theta + g\sin\theta,$$
$$g \to - q\sin\theta + g\cos\theta,$$

where q and g denote the electric and magnetic charges of a given particle. Unfortunately this requires that magnetic charge can be isolated in the same way as electric charge. The obvious way to do this macroscopically is to try to split the bar magnet into two opposite magnetic poles but the inevitable result is simply two smaller bar magnets, as Figure 3.3 illustrates.

Since the absence of isolated magnetic charge spoils the possibility of a beautiful symmetry we have to seek a reason. The natural step is to assume the existence of a magnetic charge or monopole and investigate whether this leads to any contradiction with known principles. This is what Dirac did[3]. He considered the possible effects of quantum theory. This requires that each electrically charged particle should have an associated complex wave function

$$\psi(x,y,z) = \psi(\boldsymbol{x})$$

whose phase is related to the electric charge q of the particle. This is made clear by the 'gauge principle' which requires that it is impossible to detect the phase of the wave function at any point of space. More precisely, the following change in the wave function,

$$\psi(\boldsymbol{x}) \to e^{-\frac{iq\chi(\boldsymbol{x})}{\hbar}}\psi(\boldsymbol{x}),$$

has no physical effect, whatever function $\chi(\boldsymbol{x})$ is used. Notice the appearance of the electric charge q and Planck's constant, $\hbar$. This principle requires the existence of a magnetic gauge potential $\boldsymbol{A}(\boldsymbol{x})$ which simultaneously alters as

$$\boldsymbol{A}(\boldsymbol{x}) \to \boldsymbol{A}(\boldsymbol{x}) + \nabla\chi(\boldsymbol{x}),$$

where χ is the same quantity as above. The magnetic field $\boldsymbol{B}$ is simply given by

$$\boldsymbol{B} = \nabla \wedge \boldsymbol{A},$$

at least in the static situation we have been considering on the grounds of simplicity, and is unaffected by the gauge transformation determined by χ.

Now Stokes' theorem tells us that, when the magnetic field, $\boldsymbol{B}$, can be expressed as in the last relation, its flux out of any closed surface must vanish. As a consequence, the enclosed magnetic charge must vanish. Thus it seems that it is quantum theory which forbids the existence of isolated magnetic charge. In his 1931 paper, Dirac examined this argument carefully, and found a fallacy owing to a mathematical subtlety. He revealed that, instead of vanishing, the hypothetical magnetic charge g had to satisfy

$$qg=2n\pi\hbar, \quad n=0, \pm 1, \pm 2, \pm 3, \ldots$$

Since this applies to the electrical charge q, occurring on any particle whatever, Dirac saw that this result was very exciting. It meant that the existence of an unseen magnetic charge g, sheltering somewhere in the universe, implied that q must satisfy

$$q=\frac{2n\pi\hbar}{g},$$

that is, be quantized in terms of units $\pm\frac{2\pi\hbar}{g}$, thereby following the mysterious pattern that is observed and otherwise unexplained.

Assuming this smallest unit of electric charge is that carried by the electron, which, in dimensionless units is $\sqrt{\frac{1}{137}}$, it follows that the corresponding dimensionless unit for the magnetic charge is the inverse, namely $\sqrt{137}$, and, therefore, large. Dirac argued that this disparity might explain the shyness of the magnetic monopole.

This is how Dirac left the subject in 1931. He returned to it in 1948[4], clarifying the dynamical behaviour. However many questions were left unanswered, including the following, for example.

1 What about the wave function of the magnetic monopole?

2 Why does the Dirac quantization condition above not respect the proposed rotational symmetry between q and g?

The best and most complete theoretical framework available for describing the quantum behaviour of elementary particles is known as 'quantum field theory'. Therefore a more sophisticated version of the first question would ask the following.

3 Can a quantum field theory accommodate particles with both electric and magnetic charges, and if so, how?

4 If this is possible, how heavy would the magnetic monopole be relative to the electron, say?

I want to tell you something about the partial answers to these questions that have emerged from the variety of modern developments in elementary particle theory. These new ideas include non-abelian gauge theories, spontaneous symmetry breaking, supersymmetry and supergravity, string theories, the soliton concept and so on. The exciting aspect of this is that these ideas, initially so disparate, all conspire to play a role in the further understanding of the monopole.

The most important new idea was that the gauge principle mentioned above could be extended[5]:

$$\psi(x) \to D(g(x))\psi(x),$$

where now $\psi(x)$ is a column vector with n complex entries, and $D(g(x))$ now an $n \times n$ unitary matrix. When $n = 1$ we have the previously considered situation leading to Maxwell's theory. When n is larger than one, the matrices need no longer commute, that is their product depends on the order in which they are multiplied,

$$D(g_1(x))D(g_2(x)) \neq D(g_2(x))D(g_1(x)).$$

This means the gauge group is no longer abelian and hence is open to a wider range of possibilities including ones close to the reality of the particle data as revealed by the large particle accelerators. Furthermore the new versions of Maxwell's equations are no longer linear and hence may support soliton solutions. Solitons are non-linear waves which can describe particles with structure when the theory respects the principle of relativity, as this does. It is important to realise that this mechanism for the appearance of particles in the theory is independent of quantum mechanics and so is quite distinct from that provided by the procedure of 'second quantization'. This latter mechanism interprets 'photons' as quanta of the electromagnetic field and has its origins in Planck's quanta of electromagnetic radiation which were the starting point of the quantum theory at the turn of the century.

In the non-abelian gauge theories we are considering, the soliton also requires a Higgs field[6], that is, a scalar field which fails to vanish in the vacuum, thereby 'spontaneously breaking' the gauge symmetry to a subgroup. We would like this to include an abelian factor describing electromagnetism, as above. This can be achieved by choosing the Higgs field to transform in a similar manner to the gauge potentials under the gauge transformations. This Higgs field then picks out the direction to be identified with electric charge. Furthermore the soliton that arises is automatically a magnetic monopole with charge satisfying Dirac's condition above[7]. Notice that the soliton possesses a type of charge, the magnetic charge, not possessed by any of the particles initially fed into the theory. Hence it cannot be regarded as a bound state of any of them. Rather, it is a new and unexpected excitation which is described as being 'topological', since there is a well-defined sense in which it corresponds to a knot tied in the Higgs field.

The importance of the Higgs field is that it provides a means of ascribing mass to the gauge particles other than the photon while preserving many of the highly desirable features of the original theory. In this theory, the generous Higgs field also provides the same service for the monopole, and furnishes its mass. Indeed, in the simplest version of the theory, the mass of any particle is given in terms of its electric and magnetic charges q and g by

$$\text{Mass} = a\sqrt{q^2+g^2} = a|q+ig|,$$

where the constant a is the non-vanishing value of the Higgs field in the vacuum.

For particles with vanishing magnetic charge g, this mass formula reduces to what is known as the Higgs mechanism, whereas, when the magnetic charge does not vanish, the particles are monopole soliton states whose mass is evaluated simply by integrating the energy density of the stationary solution over three dimensional space. This means that all particles are unified within this mass formula, irrespective of whether they are field quanta or solitons. Furthermore the anticipated rotational symmetry between q and g has been recovered. The theory described has the special feature that it can easily be made supersymmetric so that the above mass formula becomes exact in the full quantum theory, as we shall henceforth suppose. Supersymmetry is that rather special symmetry that can mix fermions and bosons, but there is no space for an explanation of the technicalities of it here, even though it plays an important role behind the scenes in what I have to say.

The simplest situation is when the matrices D above have just two rows and columns and, not surprisingly, it is the best understood. The particle states mentioned can be plotted in the (q,g) plane as shown in Figure 3.4.

M^+ denotes the monopole and M^- its antiparticle, the antimonopole. $W^\pm$ denote the heavy gauge particles, while the photon and the Higgs particle, bereft of either kind of charge, are massless and so situated at the origin.

The figure has the beautiful feature that distance from the origin measures the mass of the particle by virtue of the mass formula just quoted. Moreover the picture is extraordinarily symmetrical. This suggests an answer to the question posed earlier concerning the wave function of the monopole, namely that the theory could be reformulated exchanging the roles of the monopoles $M^\pm$ and the vector bosons (or gauge particles), $W^\pm$. Thus, in this reformulation, the monopoles would be quanta of the fields whose equations of motion would possess soliton solutions describing $W^\pm$. The reformulation would automatically exchange

$$\text{Mass} = a\sqrt{q^2 + g^2} = a\,|q + ig|$$

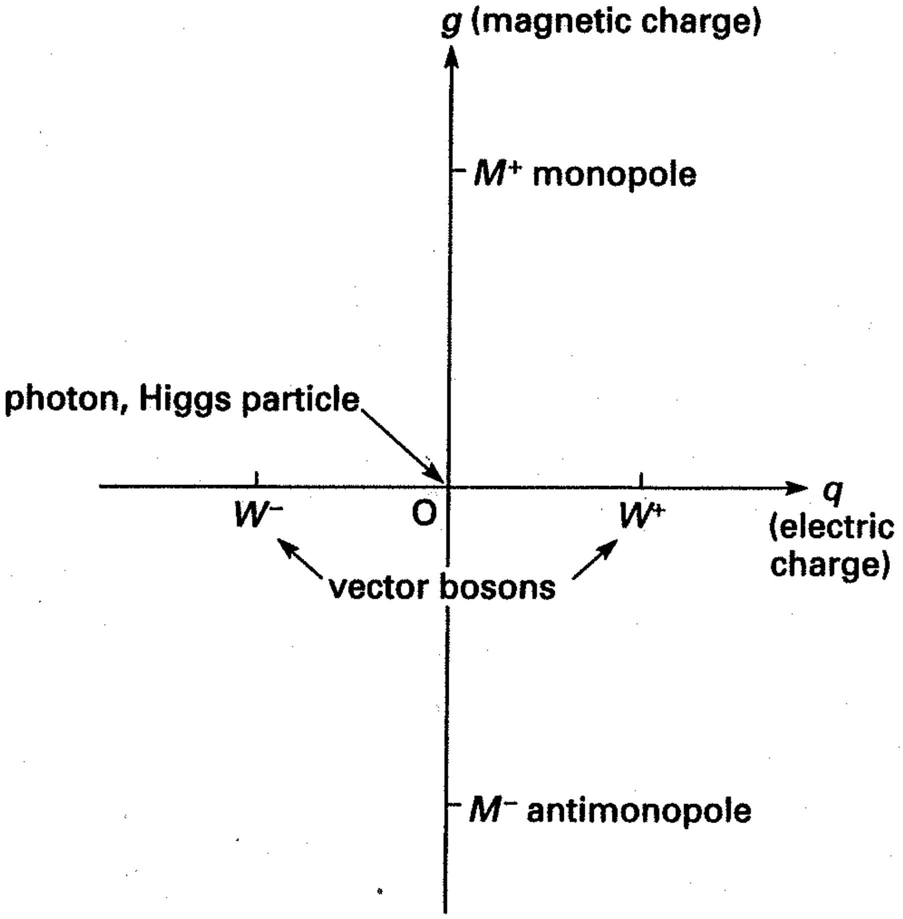


FIGURE 3.4

electric and magnetic fields, in line with our original intention. There would thus be two alternate starting points from which to calculate physical quantities, such as the masses already mentioned. There should, of course, be agreement, as Claus Montonen and I partially checked when we originally made the proposal[8]. However, a proof of the idea remains impossibly difficult to achieve as it would require the solution of a quantum field theory in four space-time dimensions, a goal well beyond reach, even now, 18 years later. Nevertheless, the understanding of this theory was substantially advanced in 1994 owing to the work of an Indian physicist, Ashoke Sen, working at the Tata Institute[9]. Consequently, the validity of the idea has become much more widely accepted. Indeed, it has become the basis for considerable further speculation whose justification is taken to be the confirmation of many non-trivial predictions.

The new work resulted in an improved picture of the charge plane just depicted. As well as the particles previously mentioned, there are also states christened 'dyons' by Schwinger[10] because they carry both electric and magnetic charge. The ones occurring with plus or minus one unit of magnetic charge are soliton solutions very much like the monopole solitons already mentioned. Sen was the first to understand how to construct the dyons with higher magnetic charges and did this explicitly for the doubly charged ones using a result of Atiyah and Hitchin[11], itself based on an idea of Manton[12]. In Figure 3.5 you see a lattice of points, sometimes denoted by crosses and sometimes by circles.

I shall call this the Sen pattern. These lattice points have coordinates (q,g) corresponding to the charges for any possible combination of particles. The lattice structure reflects two features, the fact that the electric and magnetic charges are separately conserved in any physical process, and the fact that the coordinates of any pair of points have to satisfy the Schwinger–Zwanziger quantization condition[10,13]. This is an extension of the Dirac condition mentioned earlier, taking into account the possibility of dyons. This condition does respect the symmetry under rotations between the electric and magnetic charges and so answers question (2) above.

The equations of motion of the original theory contain fields whose quanta can only carry electric charge (if at all) and not magnetic charge. These quanta therefore correspond just to the three circles on the electric, q, axis. The dyons and the monopoles correspond to the remaining circles off this axis. Thus I am saying that single particles can only correspond to the circles, whereas the

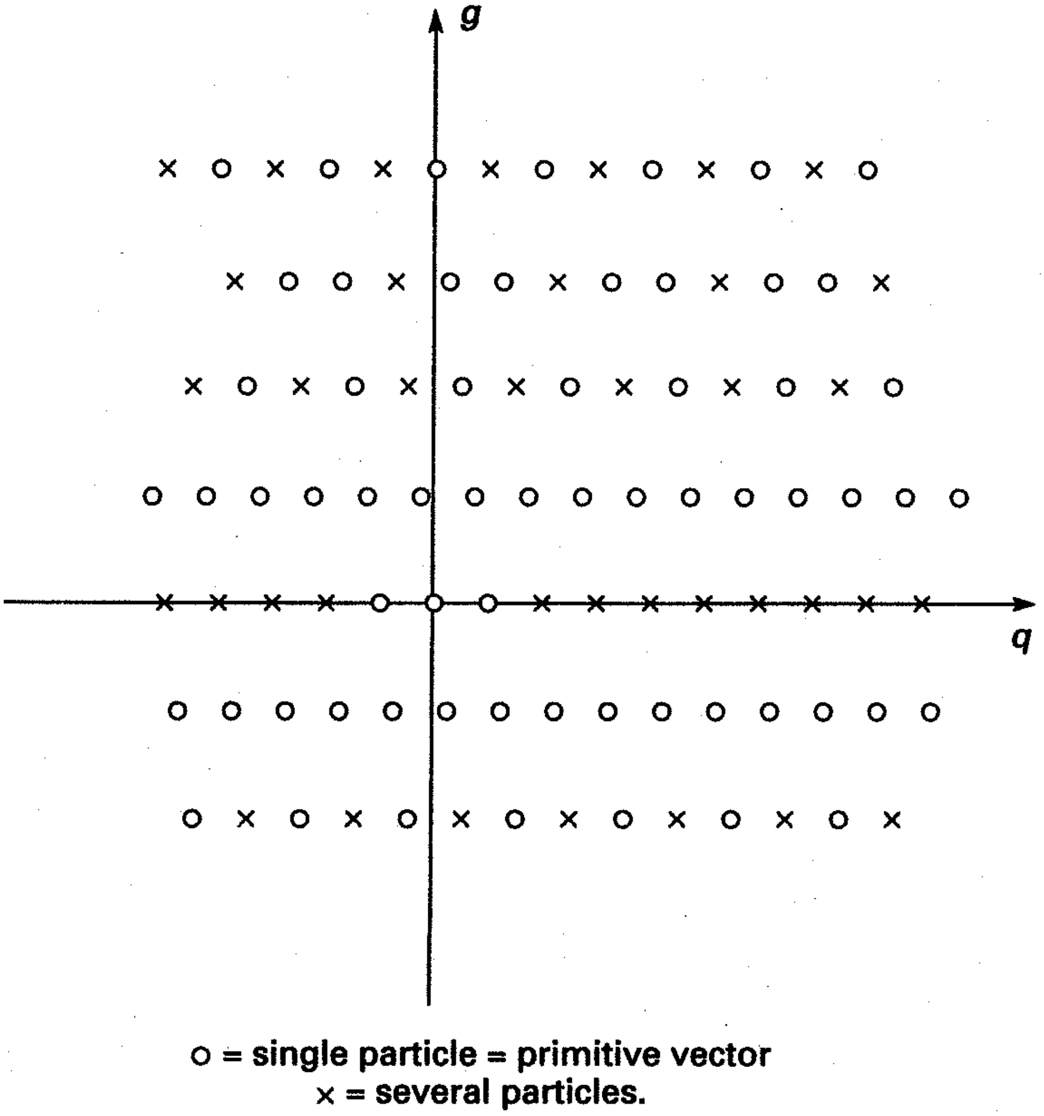


FIGURE 3.5
The Sen pattern.

crosses correspond to states describing several particles at once. There is a simple reason for thinking this, based on the formula for mass given above, as I shall explain. First note that the pattern of circles, at first sight very odd, is determined by a simple principle apparent from the figure. The circles are the only points of the lattice that can be joined to the origin by a straight line avoiding any other points of the lattice. Such points are called primitive vectors.

Consider the possibility that a particle corresponding to any point *P* of the lattice disintegrates into two pieces.These pieces must correspond to two further points of the lattice, *A* and *B*, say. During the process of disintegration the total charges, both electric and magnetic, must remain the same. That is, the charges of *A* and *B* must add up to that of *P*. This means that the vectorial sum of *OA* and *OB* must equal *OP*. Now, because of this, *OB* equals *AP* in magnitude and direction. By the formula for the masses, they are respectively proportional to the lengths *OP*, *OA* and *OB*, or equivalently *OP*, *OA* and *AP*. These form the three sides of a triangle. The length of any side is shorter than the sum of the lengths of the other two sides unless the triangle has collapsed to a line. Thus, providing *OP* is a primitive vector, particle *P* is definitely lighter than its two constituent pieces. This means the disintegration cannot actually occur as there is insufficient energy available. Thus, if *P* is a primitive vector, any single particle corresponding to it must be stable in the sense that it cannot disintegrate. On the other hand, the argument breaks down if *P* is is not a primitive vector. Then there are two points *A* and *B* of the lattice on *OP* such that the lengths *OA* and *OB* add up to *OP*. In this case it is just possible for *P* to disintegrate into *A* and *B*. Thus points not corresponding to primitive vectors of the lattice, that is crosses in the figure, are unlikely to describe single particles.

This argument is attractively general and simple, but it does not take into account unexpected extra rules forbidding decays. Therefore the conclusion has to be checked, using the details of the specific equations of motion of the theory we have been considering. Sen did this and his construction of the dyon states with higher magnetic charge explicitly produced this pattern with circles corresponding to single particles. The pattern has some other extraordinary consequences. For example, as there are circles at all points with unit magnetic charge, there exist indefinitely heavy stable single particle states. The single particle states corresponding to the circles all have different masses, depending on their distance from the origin, but are otherwise very similar, possessing exactly the same intrinsic spins, because of the nature of the supersymmetry assumed to hold.

This suggests that we could select any of the primitive vectors and construct equations of motion for the theory involving the fields whose quanta are the particles at that point, its negative, and the origin. Thus we have an infinite number of descriptions of the same theory and we expect these to be equivalent. This is the generalized electromagnetic duality conjecture. These different equations of motion must have the same structure and can only differ in terms of the parameters such as the dimensionless charge unit, physically the square root of the fine structure constant, $\sqrt{\alpha}$, and another parameter known as the vacuum angle, θ. Actually the geometry of the charge lattice also depends on parameters. A change in the description of the theory is accompanied by a change in these parameters. Because the change of description actually involves a new choice of two primitive vectors rather than one, it corresponds to what mathematicians call a modular transformation. These transformations form an infinite discrete group called the modular group. The dependence of physical quantities on α and θ should be invariant under the action of this group acting on the combination $\frac{\theta}{2\pi}+\frac{i}{\alpha}$ as a fractional integer linear transformation.

I have drawn your attention to this picture because it is radically at odds with the conventional idea of associating fields with particles accepted since the beginning of the century. Although a proper proof is probably out of reach of present knowledge, an increasing number of non-trivial tests bolster confidence. Thus the real advance is a sudden and dramatic realisation that there is much more to be understood about the structures of the quantum field theories that unify relativity and quantum theory, underpinning all theories of particle physics. This new perspective owes much to Dirac's vision and courage in considering the magnetic monopole so seriously in 1931.

Before I close, I would like to summarize the progress of the last 18 months, based upon the picture just outlined.

For a variety of reasons, the theory just sketched does not seem to describe the real world as we know it, even though it possesses some of the ingredients currently thought relevant. For example, the idea of a massless Higgs particle would raise a few experimental eyebrows. Nevertheless there are some reasons for thinking that this theory, or a similar one, will play a special role in our understanding of reality. The theory I have talked about is known technically as the $N = 4$ supersymmetric gauge theory. It appears to be the most symmetric theory possible when gravity is excluded. One of its features is that it requires none of the infinite renormalizations of the type that made Dirac so unhappy in the context of conventional quantum electrodynamics, as the other speakers have mentioned. The surfeit of symmetry which makes this possible, and which leads to the electromagnetic duality features I have discussed, is also its undoing. It is just too symmetrical to be true. Nevertheless this extreme symmetry should make it soluble. If so, it would be the first known soluble quantum field theory in the physical space-time of four dimensions.

A reasonable hope is that the hypothetical realistic theory may be an approximation to it, in which some of the symmetry features have been wiped out in a cunning way that hides them from view, while preserving some of their advantages. This sounds strange, but in fact the Higgs field works like this. The last ten years have revealed other (but related) versions of such phenomena, explaining such puzzles as phase transitions in matter. This picture would replace the conventional view that realistic theories are approximations to free theories and hence calculable when the parameter representing the 'strength' is small. An advantage of the new picture is that the modular transformations described will exchange weak and strong values of the coupling parameters. This means that the regime of strong coupling, which is usually so intractable, may now be calculable.

A very specific example of a natural modification of the theory described has been proposed by Witten and Seiberg[14]. This has the effect of making the masses of the monopole particles imaginary. This means that they will then condense, that is, smear out in the vacuum. If the electromagnetic duality I have described holds, then an analogy with the BCS theory of superconductivity could explain the confinement of electric charge by a dual Meissner effect as advocated by 't Hooft[15] and others. (The Meissner effect refers to the expulsion of magnetic flux from a superconductor.) Thanks to Witten and Seiberg there is now an explicit mechanism for this. If the relevant gauge group is interpreted as the colour gauge group, the electrically charged particles would be the quarks and it would be the mystery of their confinement that was explained.

In the transition between real and imaginary monopole mass mentioned above, there is an intermediate 'critical point' when the monopole masses vanish. Witten found special equations which the monopole fields should then satisfy[16]. Study of the solutions to these equations on different four-dimensional space-time manifolds has led to new ways of classifying such manifolds. These new results simplify and extend the celebrated results of Simon Donaldson[17]. Their power and validity supply the strongest pieces of evidence for the larger picture I have described.

The supersymmetric gauge theory described above was originally found as a limiting case of the superstring theory and it is therefore quite possible that the duality ideas can be extended to string theories, and hence include gravity. Indeed, many examples of duality in the various versions of string theory have been emerging. These involve ever more unexpected mathematical miracles.

As these often refer to the interpretation of the extra space dimensions that the strings require for sustenance, the results have profound implications for the phenomenology of particle physics. Thus there is abroad a real feeling that we are witnessing the opening of a new chapter in our understanding of the universe.

References

1 P. A. M. Dirac, 'Recollections of an exciting era', pp. 109–146, in *History of twentieth century physics: Proceedings of the international school of physics 'Enrico Fermi', Course 57, Varenna*, edited by C. Weiner, Academic Press (1977).

2 M. Gell-Mann, *The Quark and the Jaguar*, Little, Brown and Company (1994).

3 P. A. M. Dirac, *Proc. R. Soc. Lond.* A**33** (31) 60–72, 'Quantised singularities in the electromagnetic field'.

4 P. A. M. Dirac, *Phys. Rev.* **74** (48) 817–30, 'The theory of magnetic monopoles'.

5 C. N. Yang and R. L. Mills, *Phys. Rev.* **96** (54) 191–5, 'Conservation of isotopic spin and isotopic gauge invariance'; R. Shaw, University of Cambridge thesis (55), *The problem of particle types and other contributions to the theory of elementary particles.*

6 P. Higgs, *Phys. Rev.* **145** (66) 1156–63, 'Spontaneous symmetry breakdown without massless bosons'.

7 G. 't Hooft, *Nucl. Phys.* **B79** (74) 276–84, 'Magnetic monopoles in unified gauge theories'; A. M. Polyakov, *JETP Lett.* **20** (74) 194–5, 'Particle spectrum in quantum field theory'.

8 C. Montonen and D. I. Olive, *Phys. Lett.* **72B** (77) 117–20, 'Magnetic monopoles as gauge particles?'

9 A. Sen, *Phys. Lett.* **329B** (94) 217–21, 'Dyon -monopole bound states, self-dual harmonic forms on the multi-monopole moduli space, and $SL(2, \mathbb{Z})$ invariance in string theory'.

10 J. Schwinger, *Science* **165** (69) 757–61, 'A magnetic model of matter'.

11 M. F. Atiyah and N. Hitchin *Phys. Lett.* **107**A (85) 21–5, 'Low energy scattering of non-abelian monopoles'.

12 N. S. Manton, *Phys. Lett.* **110B** (82) 54–6, 'A remark on the scattering of BPS monopoles'.

13 D. Zwanziger, *Phys. Rev.* **176** (68) 1489–95, 'Quantum field theory of particles with both electric and magnetic charges'.

14 N. Seiberg and E. Witten, *Nucl. Phys.* **B426** (94) 19–52, *Erratum* **B430** (94) 485–6, 'Electromagnetic duality, monopole condensation and confinement in $N = 2$ supersymmetric Yang-Mills theory'.

15 G. 't Hooft, *Nucl. Phys.* **B190** (81) 455–78, 'Topology of the gauge condition and new confinement phases in nonabelian gauge theories'.

16 E. Witten, *Int. Math. Res. Notes* **1** (95) 769–96, 'Monopoles and four-manifolds'.

17 S. K. Donaldson, *Topology* **29** (90) 257–315, 'Polynomial invariants for smooth 4-manifolds'.

4 The Dirac equation and geometry

MICHAEL F. ATIYAH
Trinity College, Cambridge, England

Introduction

The differential operator introduced by Dirac in his study of the quantum theory of the electron has turned out to be of fundamental importance both for physics and for mathematics. Essentially the operator is a formal square-root of the wave operator or, with a different signature, of the Laplacian. In this lecture I will attempt to survey its role in mathematics. I will begin with the algebraic underpinnings, which go back to Hamilton and Clifford, and I will then go on to the role of the Dirac operator in Riemannian geometry. In particular I shall discuss various aspects of the index theorem. Finally I will briefly allude to the very recent results on four-dimensional manifolds, arising from new physical ideas of Seiberg and Witten.

4.1 Algebraic background

Let us begin by recalling Hamilton's quaternions. These are generated, over the real numbers, by three 'imaginary' quantities i, j, k, satisfying the relation:

$$\begin{aligned}&\mathrm{i}^2 = \mathrm{j}^2 = \mathrm{k}^2 = -1\\&\mathrm{ij} = -\mathrm{ji} = \mathrm{k}\\&\mathrm{jk} = -\mathrm{kj} = \mathrm{i}\\&\mathrm{ki} = -\mathrm{ik} = \mathrm{j}.\end{aligned}$$

For a general quaternion

$$x = x_0 + \mathrm{i}x_1 + \mathrm{j}x_2 + \mathrm{k}x_3.$$

We define its conjugate by

$$\bar{x} = x_0 - \mathrm{i}x_1 - \mathrm{j}x_2 - \mathrm{k}x_3.$$

Then the norm-squared of x is given by

$$x\bar{x} = \sum_{r=0}^{3} x_r^{\,2} = |x|^2.$$

The quaternions of unit norm form a group (the 3-sphere S^3) with the inverse given by

$$x^{-1} = \bar{x}|x|^2.$$

This group acts by left and right multiplication on $R^4 = C^2$, giving the two 'spin representations' of S^3. The group also acts by conjugation

$$x \rightarrow uxu^{-1}$$

on R^3 (the imaginary quaternions, where $x_0 = 0$). This gives a double covering

$$S^3 \rightarrow SO(3)$$

with kernel ± 1.

For three and four dimensions the quaternions provide all one needs to understand spinors (and in due course the corresponding Dirac operator). However, for higher dimensions we need to introduce the following.

Clifford Algebras[1]

These are generated by symbols $e_1, ..., e_n$ satisfying

$$e_r^2 = -1, e_r e_s = -e_s e_r \text{ (for } r \neq s).$$

If $x = \Sigma e_r x_r$, with real coefficients x_r, then the Clifford Algebra identities are clearly equivalent to the assertion

$$x^2 = -\Sigma x_r^2.$$

When $n=3$, if we put

$$i = e_1 e_2, j = e_2 e_3, k = e_3 e_1,$$

we recover the quaternion identities.

The cases n even or n odd are slightly different so let us concentrate on the even case and put $n = 2l$. Then it turns out that the Clifford Algebra (generated by $e_1, ..., e_n$) has an essentially unique minimal representation S of dimension 2^l. Moreover, under the action of the even part of the Clifford Algebra (involving even numbers of products of the e_r) S splits up as

$$S = S^+ \oplus S^-$$

with the e_r interchanging S^+ and S^-. These are called the ± spin representations. The choice of + versus − depends on an orientation of R^n, i.e. on an ordering of $e_1, ..., e_n$.

The group Spin(n) is defined as a certain subgroup of the invertible elements of the even part of the Clifford Algebra[1]. It acts therefore on the spaces S^+, S^-, its two spin representations. Conjugation action on R^n, the space spanned by the e_r, gives a double covering

$$\text{Spin}\,(2l) \rightarrow SO(2l).$$

There is an important general relation between spinors and complex structures which we now explain. Fix a complex structure on R^{2l}, identifying it with C^l. Consider the *exterior algebra* Λ^* of C^l. This decomposes according to degrees:

$$\begin{aligned}\Lambda^* &= \Lambda^0 \oplus \Lambda^1 \oplus \ldots \oplus \Lambda^l \\ &= \Lambda^+ \oplus \Lambda^-\end{aligned}$$

where Λ^+ is the sum of the even degrees and Λ^- the sum of the odd degrees. Note that the dimension of Λ^* is 2^l.

Λ^0 and Λ^l are both one-dimensional. The action of the unitary group U(l) on Λ^0 is trivial but its action on Λ^l is given by the determinant.

U(l) is embedded in SO($2l$) and so we get a double covering $\tilde{U}(l) \rightarrow U(l)$ induced from the double covering Spin $(2l) \rightarrow SO(2l)$. We can therefore ask how the representations S^+ and S^- behave when restricted to $\tilde{U}(l)$. The answer is that[1]

$$S \rightarrow \Lambda^* \otimes (\text{Det})^{-1/2}$$

with $S^\pm$ corresponding to $\Lambda^\pm$. The square root of the determinant representation makes sense on the double covering $\tilde{U}(l)$.

With these algebraic preliminaries out of the way we now move on to the analysis and geometry.

4.2. The Dirac operator on Riemannian manifolds

First consider Euclidean R^{2l} and define the Dirac operator

$$D=\sum_{r=1}^{2l} e_r \frac{\partial}{\partial x_r}$$

then $$D^2=-\sum_{r=1}^{2l} \frac{\partial^2}{\partial x_r{}^2}$$

showing that D is the square-root of the Laplacian $-\Delta$. We consider D as acting on spinor-valued functions on R^{2l}, so that the e_r act as matrices on the spinor space S. Moreover D interchanges S^+ and S^-. Formally D is self-adjoint and the adjoint of the operator from S^+ to S^- is the operator from S^- to S^+.

When $l=1$, D is essentially the Cauchy–Riemann operator. In general, if we fix a complex structure on R^{2l}, the Dirac operator can be expressed in terms of the Cauchy–Riemann operator acting on differential forms, together with its adjoint. This depends on the algebraic relation between S and Λ^* described in §4.1.

Now let us move from flat space to a general oriented Riemannian manifold. Using covariant derivatives instead of ordinary derivatives we can (at least locally) define the Dirac operator on spinor-valued fields. Because of the double-covering involved in going to the spinor group there is a potential global obstruction to defining spinor fields (and hence the Dirac operator). When this obstruction vanishes we call the manifold a spin manifold. As an example to indicate the nature of this restriction consider an algebraic surface X_d of degree d in the complex projective three-space. Then

$$X_d \text{ is spin} \Leftrightarrow d \text{ is even.}$$

Having introduced the general Dirac operator let us digress for a moment for some historical reflections. Maxwell's equations for electro-magnetism provided one of the main motivations in the 1930s for Hodge's theory of harmonic forms[7] (the other motivation coming from algebraic geometry). Formally, in Hodge's notation, we now write Maxwell's equations *in vacuo* in the succinct form:

$$dw = 0 \quad d^{*}w = 0$$

where w is the differential 2-form (or skew-symmetric tensor) representing the electro-magnetic field. Here d is the exterior derivative and d^{*} its formal adjoint. Of course in the Riemannian case, studied by Hodge, the metric is positive definite (not Lorentzian) and not necessarily flat. Hodge's Laplacian on differential forms, of any degree, is

$$(d+d^{*})^{2} = dd^{*} + d^{*}d.$$

This operator is elliptic and so the analysis is quite different from the Lorentzian case relevant to physics. Nevertheless the formal structure of Maxwell's equations was an important stimulus for Hodge.

Let us now look at the corresponding history of the Dirac operator and its subsequent use in geometry. The first significant appearance of the Dirac operator in geometry was in the 1960s in connection with the index theorem[3] (which will be discussed in more detail in the next section). Thus it took more than 30 years for the physics to influence geometry in this case. Admittedly this was less than the time from Maxwell to Hodge, but one might have thought that a lesson would have been learnt and progress speeded up, particularly since Hodge and Dirac were, for over 30 years, professors in the same department in Cambridge and from the same College (St John's).

However, there are good reasons why Hodge and Dirac did not collaborate, or at least exchange ideas on spinors. The reason is that the geometrical significance of spinors is still very mysterious. Unlike differential forms, which are related to areas and volumes, spinors have no such simple explanation. They appear out of some slick algebra, but the geometrical meaning is obscure as is illustrated by the global obstruction referred to earlier.

As we shall see in the next section, algebraic geometry again provided the key motivation in a way which built on Hodge's theory, but this had to await the new developments in algebraic geometry which flourished in the 1950s.

4.3. The index theorem

We now need to digress to review briefly some standard material in global differential geometry, the theory of *characteristic classes*.

The key local data of a Riemannian manifold is contained in its Riemann curvature tensor R_{ijkl}. In particular there are certain polynomial functions in R which give differential forms p_j of degree 4j. If dim M = 4*s* then any integer polynomial $f(p)$ of total degree 4*s* can be integrated over M to give a real number. A basic theorem asserts that this number is independent of the Riemannian metric and is therefore an *invariant* of M. These numbers are called characteristic numbers (or Pontrjagin numbers). The p_j represent cohomology classes which are also invariants of M: these are the Pontrjagin classes.

Now let us turn to algebraic geometry which provides so much of the motivation in modern differential geometry. A complex projective algebraic manifold of dimension *l* gives an underlying real manifold of dimension 2*l*. Important objects associated to M are the spaces H^r of holomorphic differential forms of degree *r*. Their dimensions h^r are interesting numerical invariants of the complex structure of M, generalizing the genus of an algebraic curve. The alternating sum

$$\chi(M)=\sum_{r=0}^{l}(-1)^r h^r$$

is called the arithmetic genus. A famous problem, outstanding for many years, was to express this in terms of topological invariants. This problem was solved by Hirzebruch[8] in the 1950s as a special case of his more general Riemann–Roch theorem. Hirzebruch's formula can be written as

$$\chi(\mathrm{M}) = \int_{\mathrm{M}} T(p,K)$$

where the p are the Pontrjagin forms, K is a form representing the canonical class (dual to the cycle of zeros of a holomorphic l-form) and T is a certain universal polynomial (named after J. A. Todd).

An interesting aspect of the Hirzebruch formula is that, although the Todd polynomial T has rational coefficients, the value of the integral is an integer (equal to $\chi(\mathrm{M})$). We can then ask if there are similar 'integrality theorems' for real manifolds which are not complex algebraic. In particular, dropping the dependence on the canonical class K, we can ask if

$$\int_{\mathrm{M}} T(p,0)$$

is always an integer. A clue in this direction is that the canonical class comes from the complex determinant and the Todd polynomial is related to complex exterior powers and hence to the characters of the spin representation. We are led therefore to consider spinors to see whether they may provide an answer.

Returning therefore to a spin manifold M of dimension $2l$, let D be the Dirac operator for some Riemannian metric. We define a *harmonic* spinor by the equation $\mathrm{D}s = 0$. This is reasonable because the equation $\mathrm{D}s = 0$ is (on a compact manifold) equivalent to $\mathrm{D}^2 s = 0$ and D^2 is an appropriate Laplacian. The space H of harmonic spinors decomposes as

$$\mathrm{H} = \mathrm{H}^+ \oplus \mathrm{H}^-$$

corresponding to the decomposition of the spinor-field. The quantity

$$\dim H^+ - \dim H^-$$

can be shown to be *independent* of the metric: it is called the *index* of the Dirac operator. Strictly speaking we should consider the restriction to positive spinors:

$$D^+ : S^+ \to S^-.$$

Its index is defined as

$$\text{index } D^+ = \text{Ker } D^+ - \text{Ker } (D^+)^*$$

but this coincides with our definition in view of the fact that $(D^+)^* = D^-$.

The index theorem for the Dirac operator[3] asserts that

$$\text{index } D = \int_M T(p,0)$$

and is therefore the appropriate replacement for the Hirzebruch theorem on the arithmetic genus of a complex algebraic manifold. Moreover, just as the Hirzebruch theorem is a special case of a more general Riemann–Roch theorem for arbitrary holomorphic vector bundles, so the index theorem generalises to spinor fields coupled to auxiliary vector bundles.

Remarks

1 When Singer and I were investigating these questions we 'rediscovered' for ourselves the Dirac operator. Had we been better educated in physics, or had there been the kind of dialogue with physicists that is now common, we would have got there much sooner.

2 We can now see why Hodge was unlikely to take up spinors

and the Dirac operator geometrically. The individual spaces of harmonic spinors have no obvious geometric meaning (and can vary with the metric). Only the index has invariant meaning and it was not until Hirzebruch's work on the Riemann–Roch theorem that the algebraic geometry was sufficiently advanced to provide the right motivation.

4.4. Further aspects of the index theorem

Dirac operators give basic examples of elliptic differential operators with interesting indices. In fact, in a certain sense, they generate all examples and so the index theorem for Dirac operators leads to the index theorem for arbitrary elliptic operators. To understand the reason why Dirac operators are so basic we must review some relevant topology.

Since the index is topological and since we are dealing with linear operators it is important to understand the topology of the linear or unitary groups. In fact such groups occur in the theory for two separate reasons, relating to the dependent and independent variables respectively. In terms of vector bundles this means we have to deal with the spin bundle (and auxiliary fields) and also with the tangent bundle of the manifold.

The key topological theorem is the famous periodicity theorem of Bott[4], which asserts that for large N the homotopy groups of the unitary groups are given by:

$$\begin{aligned}\pi_i(\mathrm{U}(N)) &= 0 \ (i \text{ even})\\ &= \text{integers } (i \text{ odd}).\end{aligned}$$

Moreover in the odd case there is a simple formula for the generator. For $N=1$ we know that

$$(x_1, x_2) \to (x_1 + \mathrm{i}x_2)$$

maps the unit circle $x_1{}^2+x_2{}^2=1$ with degree 1 onto the circle U(1). For higher dimensions the generator of $\pi_{2l-1}(U(2^{l-1}))$ is given by

$$x \to \sigma\left(\sum_{r=1}^{2l} x_r e_r\right)$$

where $x \in R^{2l}$, with $|x|=1$, and σ is the representation of the Clifford algebra acting from

$$S^+ \to S^-.$$

The fact that the Clifford algebra describes the generator of the homotopy group of U(N) explains in part why the Dirac operator is so fundamental.

The index theorem for the Dirac operator, and its generalizations, have many interesting geometrical applications. A particularly attractive example concerns manifolds with circular symmetry, i.e. which admit a non-trivial action of the circle group. One can prove the following theorem[2].

THEOREM. *If M^{4s} is a spin manifold with circular symmetry then the index of the Dirac operator is zero.*

Now we know that

$$\text{index D} = \int_M T(p)$$

so we deduce the vanishing of the integral. Conversely if we have a manifold M with

$$\int_M T(p) \neq 0$$

we deduce that M cannot have any circular symmetry. Note that this statement makes no direct reference to the Dirac operator or spinors, they appear only in the proof. However, the spin condition on M is essential. For example the complex projective plane has non-vanishing p_1 and admits a large symmetry group, namely SU(3). However, it is not a spin manifold.

The method of proof of the theorem is elegant but somewhat unusual. Let me outline the idea. If a circle action exists we can define a refined index as a character of the circle, i.e. as a finite Fourier series

$$f(z) = \Sigma a_n z^n \quad (a_n \text{ integers}).$$

For $|z| = 1$ but $z \neq 1$ this can be computed as a sum of integrals over the fixed points of z. Studying the behaviour of $f(z)$ shows that

$$f(z) \to 0 \text{ as } z \to 0$$
$$f(z) \to 0 \text{ as } z \to \infty.$$

For a finite Fourier series this can only happen if $f(z) \equiv 0$. But the ordinary index is the value

$$f(1) = \Sigma a_n$$

and so this must be zero.

Let me conclude this section with a few brief remarks about the relation of the index theorem to physics.

In the first place the relevant physics is quantum and not classical. Because quantum calculations are frequently performed formally via the Feynman integral and by continuation to imaginary time we encounter the Euclidian signature and hence the elliptic rather than the hyperbolic Dirac operator. The index of the Dirac operator then turns up in the guise of an 'anomaly', something at the quantum level which violates a classical symmetry. In this case the symmetry is 'parity' or orientation change, which switches positive and negative spinors. A non-zero index indicates that, in global geometry, and in quantum physics this local and classical symmetry is not preserved.

Next we note that there is a string theory version of the Dirac operator. Geometrically this can be viewed as an index theorem on the infinite dimensional loop space of M. This space has a natural circle action given by internal rotation of each loop and one has to use this action to give a refined index. Moreover, if there is a further circle action on M itself, we get two independent circle actions on the loop space. Using the appropriate index theorem in this setting Witten[18] derived further 'vanishing theorems' for manifolds with circular symmetry. These have been given rigorous mathematical proofs by Bott and Taubes[5]. Moreover this whole area has stimulated a new branch of topology called 'elliptic cohomology'[11].

4.5. Four-dimensional geometry

Although the Dirac operator has been used in all dimensions, the most recent and perhaps deepest applications have been in four dimensions, the dimension of usual space-time and precisely where the physics began.

It has been known for a decade, through the work of Simon Donaldson[6], that four-dimensional geometry exhibited special phenomena of an unexpected kind. The tools that Donaldson introduced to study these features of four dimensions were the Yang–Mills equations and their instanton solutions. Using these, Donaldson introduced some subtle invariants of four-manifolds which were polynomials functions on the second homology of the four-manifold.

Much work has been done on these Donaldson invariants and they have proved to be very powerful, enabling one to distinguish differentiably between four-manifolds which are topologically equivalent. On the theoretical level there have been three major discoveries.

First, Witten[17] showed that the Donaldson invariants could be interpreted as the output (as correlation functions) of a *topological* quantum field theory. Moreover this theory, while not itself physical, was a 'twisted' version of a conventional physical theory, namely $N = 2$ supersymmetric Yang–Mills theory. This brought Donaldson theory closer to physics and stimulated a general interest in other topological quantum field theories.

The next discovery was made by Kronheimer and Mrowka[9] who showed (under certain conditions) that the Donaldson polynomials, summed over all degrees, could be re-expressed in terms of the exponentials of a finite number of two-dimensional cohomology classes. These classes were called basic classes and they contained all the information in the totality of Donaldson polynomials. Unfortunately their existence was deduced indirectly, through recurrence relations, and a direct interpretation of their geometric significance was lacking.

Meanwhile physicists were re-examining some old ideas on duality which had been proposed by Montonen and Olive. The duality between electricity and magnetism exhibited by Maxwell's equations had always been intriguing and was, of course, behind Dirac's introduction of his magnetic monopole. With the appearance of non-abelian gauge theories the situation became even more interesting. On the one hand 't Hooft[14] and Polyakov[13] discovered the existence of smooth magnetic monopoles which were soliton solutions of the relevant non-linear equations. Montonen and Olive[12] proposed a phenomenological duality at the quantum level in which solitons and elementary particle fields interchanged their roles. These ideas were taken up recently by Seiberg and Witten[15] who have shown that many super-symmetric theories exhibit such a duality. This is of great interest to physics because the duality switches weak and strong coupling and has potential implications for quark confinement.

As a mathematical version of these ideas Witten[19] has argued that Donaldson theory (as a supersymmetric QFT) should have a dual version based on mass-less monopoles. In particular the

Donaldson invariants should be calculable in terms of the classical solutions of a system of equations for an abelian gauge theory coupled to spinors. These equations are now called the Seiberg–Witten equations and have turned out to be remarkably useful[10]. Their definition in brief is as follows.

On a compact oriented four-manifold M we fix a complex line-bundle L and a U(1)-connection A. For simplicity assume M is a spin manifold and let S^+ be its bundle of positive spinors. Let φ be a section of $S^+ \otimes L$, then the Seiberg–Witten equations for the pair (A,φ) are:

$$D_A\varphi = 0 \quad \text{(Dirac equation)}$$
$$F_A{}^+ = [\varphi\bar{\phi}]^+$$

where $F_A{}^+ \in \Lambda^2{}_+$ is the self-dual part of the curvature of A and the term $[\varphi\bar{\phi}]^+$ is the component of $\Lambda^2{}_+$ in the decomposition

$$S^+ \otimes S^+ = \Lambda^0 + \Lambda^2{}_+.$$

The line-bundle L is determined topologically by its first Chern class $c_1(L)$ and we can ask, for which two-dimensional cohomology classes $x = c_1(L)$, do the Seiberg–Witten equations have solutions.

The beauty of the Seiberg–Witten equations is that the space of solutions is compact and in general consists of finitely-many points. Witten has argued, on physical grounds, that the Kronheimer–Mrowka basic classes are just those for which the Seiberg–Witten equation has solutions. Although this has yet to be established as a mathematical theorem the evidence is impressive. In particular using the Seiberg–Witten equations as a new tool it has been possible to reprove and, in many cases, improve results obtained by Donaldson. A notable example is the proof[10] of the old conjecture of René Thom which asserts that a compact oriented surface embedded smoothly in the complex projective plane has minimal genus *g*, for fixed degree *d*, when the surface is an algebraic curve. In other words

$$g \geq \frac{(d-1)(d-2)}{2}.$$

A remarkable aspect of the Seiberg–Witten duality, for geometers, is that spinors do not enter into the original Donaldson theory. Their magical appearance clearly provides some deep insight into their geometric meaning.

Although the Seiberg–Witten equations can be used *ab novo*, without reference to the Donaldson theory, this is clearly an unsatisfactory and temporary situation. Geometers should seek to understand the duality in their own terms. This should be a very enlightening process.

In conclusion therefore we see that spinors and the Dirac equation are still playing an important role in geometry. We are getting closer to a proper understanding of spinors. Perhaps I could close by a quotation from Hermann Wey[16] which has always intrigued and bemused me. I hope Weyl would have felt vindicated by recent developments.

> Only with spinors do we strike that level in the theory of representations [of the orthogonal group] on which Euclid himself, flourishing ruler and compass, so deftly moves in the realm of geometric figures. In some way Euclid's geometry must be deeply connected with the existence of the spin representation.

References

1. M. F. Atiyah, R. Bott and A. Shapiro. Clifford modules. *Topology* **3** (Suppl. 1) 1964, 3–38.
2. M. F. Atiyah and F. Hirzebruch. *Spin manifolds on group actions. Essays on topology and related topics. Memoires dédiés à Georges de Rham.* Springer-Verlag (1970), 18–28.
3. M. F. Atiyah and I. M. Singer. The index of elliptic operators on compact manifolds. *Bull. Am. Math. Soc.* **69** (1963), 422–33.
4. R. Bott. The stable homology of the classical groups. *Proc. Nat.*

Acad. Sci. USA **43** (1957), 933–5.

5. R. Bott and C. Taubes. On the rigidity theorems of Witten. *J. Am. Math. Soc.* (1989), 137–86.
6. S. Donaldson and P. Kronheimer. *The Geometry of Four-Manifolds*. Oxford University Press 1990.
7. W. V. D. Hodge. *The Theory and Application of Harmonic Integrals*. Cambridge University Press 1941.
8. F. Hirzebruch. *Topological Methods in Algebraic Geometry*. Springer-Verlag (1966).
9. P. Kronheimer and T. Mrowka. Embedded surfaces and the structure of Donaldson's polynomial invariant. *J. Diff. Geom.* **41** (1995), 573–734.
10. P. Kronheimer and T. Mrowka. The genus of embedded surfaces in the projective plane. *Math. Res. Lett.* **1** (1994), 797–808.
11. P. Landweber (Ed.) Elliptic Curves and Modular Forms in Algebraic Topology, *Lecture Notes in Mathematics 136*, Springer 1988.
12. C. Montonen and D. Olive. *Phys. Lett.* **B72** (1977), 117.
13. A. Polyakov. Particle spectrum in quantum field theory. *JETP Lett.* **20** (1974), 194.
14. G. 't Hooft. Magnetic charges in unified gauge theories. *Nucl. Phys.* **B79** (1974), 276.
15. N. Seiberg and E. Witten. Electric-magnetic duality, monopole condensation and confinement in N = 2 supersymmetric Yang-Mills theory. *Nucl. Phys.* **B426** (1994), 19.
16. H. Weyl. *The Classical Groups*. Princeton University Press 1946.
17. E. Witten. Topological quantum field theory. *Comm. Math. Phys.* **117** (1988), 353–86.
18. E. Witten. The index of the Dirac operator in loop space, in [11] 161–81
19. E. Witten. Monopoles and four-manifolds. *Math. Res. Lett.* **1** (1994), 764–96.