The Philosophy of Vacuum

Edited by

SIMON SAUNDERS and HARVEY R. BROWN

CLARENDON PRESS · OXFORD

The Philosophy of Vacuum

This book has been printed digitally and produced in a standard specification
in order to ensure its continuing availability

OXFORD
UNIVERSITY PRESS
Great Clarendon Street, Oxford OX2 6DP

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide in
Oxford New York
Auckland Bangkok Buenos Aires Cape Town Chennai
Dar es Salaam Delhi Hong Kong Istanbul Karachi Kolkata
Kuala Lumpur Madrid Melbourne Mexico City Mumbai Nairobi
São Paulo Shanghai Singapore Taipei Tokyo Toronto
with an associated company in Berlin

Oxford is a registered trade mark of Oxford University Press
in the UK and in certain other countries

Published in the United States
by Oxford University Press Inc., New York

Reprinted 2002

ISBN 0-19-824449-5

Acknowledgements

The Philosophy of Vacuum was conceived as an exercise in metaphysics; the diversity of viewpoints is a reflection of the science of our time. Many of the papers in this collection were presented in the conference at the Sub-Faculty of Philosophy at the University of Oxford, held in June 1987. We thank all those involved in the project, most particularly Rom Harré, without whom the conference would not have got off the ground, Angela Blackburn of Oxford University Press, and the Albert Einstein Archives of the Jewish National and University Library, the Hebrew University, Jerusalem, which holds the copyright over Professor Einstein's published and unpublished papers; their generosity is gratefully acknowledged. As regards the work of copy-editing the manuscript of this volume, we wish to thank Susan Hughes for carefully undertaking what was an unusual project for her. Finally, our thanks go to Timothy Budden for compiling the index, and to Frances Morphy of Oxford University Press for seeing the volume through to publication with such cheerful dedication.

S.W.S.
H.R.B.

Oxford
1991

Contents

Contributors

I. J. R. Aitchison
Department of Theoretical Physics, Oxford University

Michael Atiyah
Trinity College, Cambridge

Peter J. Braam
Department of Mathematics, University of Utah

Harvey R. Brown
Sub-Faculty of Philosophy, Oxford University

David Finkelstein
School of Physics, Georgia Institute of Technology

Gordon N. Fleming
Department of Physics, Pennsylvania State University

B. J. Hiley
Department of Physics, Birkbeck College, University of London

R. Penrose
The Mathematical Institute, Oxford University

Simon W. Saunders
Department of Philosophy, Harvard University

D. W. Sciama
International School for Advanced Studies, Trieste

Robert Weingard
Department of Philosophy, Rutgers University

Introduction

SIMON SAUNDERS

According to Aristotle, *vacuum* is τὸ κενόν, 'the empty'; it is 'space bereft of body'. What then is 'space' and 'body'? At once we have two of the central themes of metaphysics: the concept of vacuum is parasitic on the concept of space and the concept of substance. Most important of all, it rests on the distinction between the two.

The Ancient World knew many concepts of substance; some of them appeared to preclude the concept of 'the empty'. As for space, Aristotle used the word *topos*, 'place' or 'container'—a continuous, finite receptacle which persists through change and which is 'separable' from matter. What then is this distinction between matter and the 'container' of matter? It is only with the atomists that the concept of *emptiness* acquires an absolute character; the distinction is made out in terms of the existence or non-existence of atoms.

The atomists were surely correct in their basic tenet: *in some sense* the world is atomistic. Surely, then, *in some sense* it is possible to distinguish matter from void. What appears undeniable, however, is that in the present state of theoretical physics there are many levels to this notion of existence. Certain entities—particles—categorically exist. Others—virtual particles, energy, fluctuations—exist in some sense, perhaps in a relative sense (differences in energy, etc.). Others—negative-energy particles, Rindler quanta, wave-functions—perhaps do not exist. Equally, there are levels to the concept of space: there is the manifold; there is a topology and a differentiable structure; as we add an affine structure and a metric there is geometry. If the vacuum is the *least* that exists, where in our catalogue of realities is the least and most simple object?

Let me come to the concept of *force*. Action at a distance poses few problems to the atomist void; it is the reification of force, first the 'powers' and 'ethers' of eighteenth-century theories of electricity, magnetism, and heat, and later the forces associated with the wave theory of light—it is these that present difficulties. The elastic–solid ether of Fresnel is the fall-out of the material aspect of light, and it falls everywhere, filling the void. Is this ether 'empty enough' to count as vacuum? But then we must consider how this ether is used.

Empty space, whatever it is, now controls the dynamics of material bodies. It has a dynamic role in the organization of matter, because when we consider electromagnetism, it seems that matter, and functional relationships between particles of matter, are not able to do the job on their own. Of course space, too, may be thought of as an organization principle applied to matter. The vacuum is to causation what space is to geometric relationships.

What of geometry? Throughout all the debates of antiquity and through to the controversy over the elastic–solid ether, this strand to the concept of vacuum runs straight as a die: geometry is Euclidean. But it is not space so much as body that has geometric properties—primary properties, at that; *res extensa*, the Cartesian reality outside of the mental. Euclidean geometry is a set of positioning laws, 'Lagerungs-Gesetze', a phrase still used by Einstein in 1924. In the early nineteenth century this situation changed; at precisely the same time that the vacuum became ether, physical space became a *geometric space*.

Of course, from a contemporary standpoint, the 'Absolute Space' of Newton is necessarily a Euclidean space. So much is implicit in the laws of motion. Their invariance group makes of physical space and time a geometry—Newtonian space–time. But the recognition of this fact requires an understanding of the group concept and this came late, at the end of the century. The vacuum is now an organization not only of bodies, but also of itself.

This is just what we see in the electromagnetic ether; the organization in question was, however, to be provided by the Newtonian mechanics. At the same time, the ether was to act contiguously and in accordance with Maxwell's equations. With Lorentz, the material subject of this action becomes particulate. By the end of the century, the ether was conceived of in relationship to Absolute Space, Newtonian mechanics, atomism, Maxwell's electromagnetism, and contiguous action—a witch's brew. In retrospect, it is remarkable that so sophisticated and elaborate a theory as the electrodynamics of Lorentz could be constructed at all, let alone that it could claim the empirical success which it in fact did; in any event, it was the recognition of the role of space–time geometry that banished the Lorentz ether and restored to the concept of vacuum a certain austerity.

Just how this emerged I shall not try to summarize; I leave it to Albert Einstein's 'On the Ether' (Chapter 1) appearing for the first time in English translation. This paper was written in 1924, following the work of Bose and of Compton, but prior to the discovery of quantum mechanics. As an historical review of the theory of ether, most particularly from the point of view of general relativity, it is very nearly the last document of this kind which dates from the classical era. Strangely enough, it is also one of the only foundational surveys in which Einstein may be said to speak for his generation. In the year 1924 the light quantum hypothesis entered the mainstream, but by 1926 so too had quantum mechanics; Einstein bids us reluctant farewell.

However, in this paper Einstein makes only limited reference to the light quantum hypothesis; when he does, it is to call into question the independent reality of the electromagnetic field, not the metaphysical assumptions of his critique. In particular, he makes no connection between the light quantum and the concept of contiguous action.

Events over the next four years transformed this picture out of all recognition. The picture has scarcely stopped changing since. This collection is concerned with many different aspects of quantum theory and the fundamental forces, but there is one very simple feature of quantum field theory which has an immediate relevance to the ether, and the history with which Einstein is concerned. That is the discovery by Paul Dirac, as interpreted and elaborated by Pascual Jordan shortly after, that bosons and fermions may be described as excitations of quantum fields, whether they be photons, electrons, or protons.

We know what Einstein thought of quantum field theory (he was not interested). From this time his attention was focused exclusively on the foundations of quantum mechanics and classical field theory. But it is of interest to reassess the various transformations of ether from a perspective broad enough to include the quantum theory of fields; that, at any rate, is the programme of Simon Saunders and Harvey Brown in their 'Reflections on Ether' (Chapter 3). One other contribution is relevant. Roger Penrose, in 'The Mass of the Classical Vacuum' (Chapter 2), gives a brief but succinct account of an important gap in Einstein's critique: the question of the locality or otherwise of the energy momentum distribution within the gravitational field itself. Einstein was fully aware of this difficulty; as Saunders and Brown suggest, it may be that he sought the resolution of this (purely classical) non-locality in his projected classical synthesis of gravity and the quantum.

What of the concept of vacuum today? Relativity and quantum theory respectively define what is to count as *space* that is *empty*. In place of the classical critique, we have Riemannian spacetime, geodesic motion, quantum mechanics, quantum fields, and locality—the devil's own cauldron. We know that in some sense the vacuum should contain *least*—the least that is useful, or the least that is necessary, or the least that is operationally definable, which of these we do not know. But in keeping with the atomistic flavour of quantum theory, one intuition stands out: the vacuum should contain no particles. From a more general standpoint, the energy should be a minimum. In the non-relativistic case these two demands are not too hard to satisfy; a clear-cut definition of vacuum is possible. For example, the vacuum of non-relativistic quantum field theory does not exhibit zero-point fluctuations, in the sense that, there, no linear combinations of the field (or its canonical conjugate) can be considered observables (on pain of violating mass superselection); the uncertainty relationships between the fields may be considered purely mathematical and of no physical significance. No doubt the non-relativistic vacuum is tractable just because the dynamics is implemented by distance forces; in the relativistic case the situation is quite different. And the situation is far worse as soon as one goes to a more general (locally Lorentzian) space–time; it seems there is no criterion, based on the particle concept, of what is to count as 'empty' space.

In relativistic theory the situation is so difficult that one sympathizes with the radical approach, by Basil Hiley ('Vacuum or Holomovement', Chapter 9) and David Finkelstein ('Theory of Vacuum', Chapter 10) in particular, to abandon the traditional starting-point of the continuous space–time manifold. Not only have these authors come to a similar conclusion as to what must be abandoned in the conventional theory, but they adopt similar strategies: there must exist an object prior to space—'pre-space' in the language of Hiley, 'causal networks' according to Finkelstein—and in both cases this object is an algebraic structure. Further, the ideas of Grassman have an important influence. Grassman was concerned with the algebraic expression of 'activity', the relevant heuristic according to both Hiley and Finkelstein, and he was concerned with the algebraic expression of set theory, the prerequisite to a quantum theory of sets as envisaged by Finkelstein. But there the similarities end. In fact, Hiley and Finkelstein start from diametrically opposed positions; Hiley is concerned with hidden variable theory, Finkelstein with the extension of quantum principles to the most elementary mathematical categories. However, in this enterprise it is Finkelstein who maintains the principle of locality at the fundamental level, for it is built into the concept of a 'causal network'; indeed, the Lorentz group is to arise from the symmetric group of the 'binary node', the (in some sense favoured) nodal system of the network. What drives the Finkelstein programme is above all the demand for a locally *finite* theory. The algebra that underlies the quantum set theory, and thence the quantum topology (the basic subject of change), must be finite; therefore the algebraic operations cannot include negation. (The orthocomplement on the local logic takes one to the global logic; the complement of a finite subset of an infinite set is infinite.) Therefore one must give up unitarity in the representation of the local algebra defining the topology.

Hiley takes a different view; it is non-locality that is fundamental to quantum theory, and which governs the determinate motion of individual systems subject to the quantum potential. What is characteristic of 'pre-space' is that physical space (and locality) are encoded in the *implicate order*. The traditional expression of these concepts is the *explicate order*. The algebraic description of pre-space is intrinsically non-local, or 'a-local'. The various ways in which distinct explicate orders may be associated with an algebra, and therefore with distinct definitions of locality, are illustrated with a simple example: a finite Weyl algebra, corresponding to the representation of a symplectic geometry over a compact phase space. In this model the definition of locality is formulated in terms of a 'neighbourhood relation'. One speculation is this: a unique description of a local space–time manifold may arise only in the classical limit.

These attempts are speculative, and as yet there are no concrete returns in the physics and no fundamental innovations in the mathematics. In the latter respect, more promising is the astonishing development of a quantum theory of pure topology over 3- and 4-manifolds, and its connection with the classification of knots in

3-space and differentiable structures on four-dimensional manifolds. In this rapidly changing field, we are fortunate to have a brief account from one of its creators; in 'Topology of the Vacuum' (Chapter 11), by Michael Atiyah, and as elaborated by his student and co-worker Peter Braam in 'How Empty is the Vacuum?' (Chapter 12), the principal developments are summarized. Some remarks on the four-dimensional case: there Donaldson had shown that a certain class of solutions to the classical Yang–Mills equations—which in a certain sense correspond to 'vacuum' solutions, those that do not describe any particle fields—has itself an intrinsic topology which does not depend on the metric used on the original 4-space. In favourable cases (for example when the original space is compact), this topology simply determines the differentiable structure of the original space. The seminal advance of Witten is to show that the resulting invariants can be computed, using path-integral methods, as expectation-values of a certain kind of charge arising in the framework of quantum field theory, invariants that are, recall, independent of the metric. In this sense Witten has constructed a quantum theory of pure topology. There is the hope that quantum theory may be formulated on a manifold *prior to* the definition of any metrical structure; there should exist a more elemental vacuum, prior to the introduction of matter, force, or geometry. In this philosophy the vacuum has no spatial or temporal relationships, but it is still a topological manifold; according to Atiyah and Braam, that is all that it contains: it is a vacuum of 'pure space'.

It seems that the vacuum need not carry spatial and temporal relationships. Must it be defined in space–time terms at all, even a space–time devoid of metrical structure? Atiyah and Braam make of the vacuum something superficially more simple, because independent of the metric; but the physical structure of space–time is made more complex, because the topology is also subject to quantum law. But is it necessary that the vacuum be associated at all with the space–time manifold? There is the temptation of a logical retreat, that the concept of *nothingness*, 'the empty', should describe that which is empty of space and time, along with all other physical objects. Indeed, this is just the possibility offered by a closed universe; 'absolute nothingness' is what is not in the universe, it is what has no physical properties whatsoever.

This vacuum leads to unicorns and present kings of France. As Russell lamented long ago, 'how can a non-entity be the subject of a proposition?' As he subsequently remarked of his own theory of descriptions, what is wrong with such concepts, and a theory of meaning that gives them denotation, is the evident 'failure of that feeling for reality which ought to be preserved even in the most abstract studies'. Such a vacuum is much more like the empty set; it is, perhaps, the sense of vacuum considered by Descartes, the sense in which vacuum cannot exist between separated objects (for there would then be nothing that separates them). This vacuum does not exist, by definition.

Descartes, in the tradition of his time, believed that geometry could apply only to physical bodies. But what if geometry applies also to vacuum? This is Robert Weingard's proposal; in his 'Making Everything Out of Nothing' (Chapter 8), he stands Descartes on his head. If vacuum is extended, then it is *res extensa*; but this is the *only* non-mental reality, according to Descartes. Therefore vacuum is all that exists. To what extent is this a viable interpretation of contemporary physics?

Weingard is concerned not so much with the present situation in physics as with the question of principle: what might categorically prohibit an ontology of pure space? In a series of examples he demonstrates the versatility of geometric models; the issue is then whether any significant properties of 'matter' must be omitted. The theories that he considers are almost entirely classical or quasi-classical (*c*-number gauge theories); at this level he presents a convincing case. But they are also bosonic theories, and as Weingard admits, it is the fermion case that is most problematic.

From a phenomenological point of view, pure geometry is successful at the level of fields of force. (In my earlier terminology, one can geometrize the organization of the vacuum on itself.) The sources of these fields may perhaps be topological structures, but to date no such theory can claim a shred of empirical success. Ian Aitchison, in contrast, begins from the phenomenological theory, that is the standard model, and with the fundamental distinction between force and matter. In 'The Vacuum and Unification' (Chapter 7), we see the vacuum in all its splendour. The coherence and variety of phenomena and concepts presently exploited in the standard model are deeply impressive. The vacuum that emerges is rich: by turns a ferromagnet, a dielectric, a superconductor, and a thermodynamic phase. Increasingly, this vacuum is reminiscent of ether. Indeed, Aitchison is happy to draw parallels between the ether of the nineteenth century and the grand unified vacuum. He does this not out of any great fondness for the concept of ether, but through the perception of its unifying role in dynamics. The fundamental categories of matter, force, and space–time are most simply united through modification of the vacuum.

The analysis of the zero-point fluctuation due to Sciama in 'The Physical Significance of the Vacuum State of a Quantum Field' (Chapter 6) also demonstrates a methodology. These fluctuations provide a powerful heuristic, and even if many of the phenomena described by Sciama can be interpreted in other ways, they are seldom more simply or intuitively described. Further, the zero-point fluctuations, unlike the non-zero vacuum expectation values of the Higgs and Goldstone fields considered by Aitchison, seem to follow from basic principles of quantum theory. But as Sciama makes clear, the cosmological implications are all the more pressing; if the fluctuations are a reality, then so too is their associated energy density. But this energy density is infinite. It does not seem possible to eliminate this difficulty by appeal to the conventional philosophy of renormalization, where the discarded infinities are considered a symptom of the incompleteness of the theory.

I have emphasized a difficulty of Sciama's approach; let me also mention an important success. The existence of a non-zero particle distribution in the Minkowski vacuum, as described by an accelerating observer in an 'appropriate' coordinate system, is a remarkable and disturbing feature of relativistic quantum theory. Must we conclude that the concept of particle is observer-dependent? What of the energy associated with such a particle distribution? Unruh's idealized model of an accelerating particle detector shows further that such particles (the so-called Rindler quanta) should be experimentally detectable. (This model has unsatisfactory features, however.) But now, as Sciama shows, the zero-point fluctuations may be invoked. These can be considered a background 'noise-power' to the field. But its effect on a moving system is given by the Fourier transform of the auto-correlation function of the field evaluated along the world-line of the system; this will in general depend on the world-line. The Rindler quanta are only one more manifestation of the zero-point fluctuation, and they have energy supplied from this background.

Unity is fundamental to physical science, but so too is simplicity. In this empty space of Weingard, Sciama, and Aitchison, there is a great deal of activity. It may be that we can exploit the vacuum to describe observable phenomena—as did the ether theorists before us—but is the complexity of vacuum *forced* by the conventional theory?

The answer, it seems, is affirmative. Quite generally, the vacuum of a quantum field can have no unique definition in terms of particle number alone. As a system of infinitely many degrees of freedom, the Stone–Von Neumann theorem which ensures the uniqueness of representations (up to isomorphism) of the canonical commutation relationships fails; inequivalent representations of Fock type can be easily constructed, and there is a non-denumerable infinity of them. This fact underlies some of the difficulties of renormalization theory and plays a crucial role in circumventing Haag's theorem in any theory with a simple vacuum structure. Therefore the vacuum can never be specified by the algebra of creation and annihilation operators alone (unless one has pure kinematics).

But is the number operator the appropriate quantity? Are the canonical commutation relationships necessary to a relativistic theory? Dirac showed long ago that they are not. It is fundamental to quantum theory that the commutation relationships govern the algebraic structure of the generators of transformations which act on the initial data; it is not fundamental to quantum theory that the initial data are specified on a spacelike hyperplane. Dirac found that, if these initial data are specified on the 3-volume defined by the motion of a two-dimensional plane spacelike surface along a lightlike line (i.e. the propagating surface of a plane wave), the algebra of the commutation relationships is simplified. It emerged that there are no commutation relationships between the field and its canonical conjugate, which appears as a functional of the field; there is no vacuum fluctuation; and there is a unique and stable vacuum.

These remarkable but neglected properties of the 'front' (or 'null plane') form of quantization (in contrast to the 'instant' form, initial data defined at an instant in time over the whole of space) are systematically reviewed by Gordon Fleming ('The Vacuum on Null Planes', Chapter 5), in the context of $\lambda\phi^4$ theory. As Fleming remarks, the new vacuum is embarrassingly trivial, for it seems to imply that the inequivalent representations are in some sense illusionary.

What obstructs a clear resolution of these problems is the fact that an effective renormalization is made in the 'front' form of the theory, differing from the usual regularization of the 'instant' form. Its significance from the point of view of the 'instant' theory is unclear, and this is the source of the difficulty. As Fleming points out, it may be that this is trivial. It would be of interest to know different vacua; we would then have a different and perhaps more tractable approach to modelling the phenomena, in which the vacuum, empty space, appears truly empty, so long as by 'space' we consider the 3-volume swept out by a null 2-plane. On the other hand, it may be that the properties of $\lambda\phi^4$ are not a reliable guide; taken at face value, the uniqueness of the vacuum, together with Haag's theorem, imply that the theory is trivial, which we know to be true in $3+1$ dimensions. It would be of interest to know if the vacuum is unique in null-plane quantization in $2+1$ dimensions, where $\lambda\phi^4$ has consistent non-trivial models.

Like Fleming, Simon Saunders in 'The Negative-Energy Sea' (Chapter 4) considers alternative descriptions of familiar theories, by which the vacuum may be described in a simple and canonical way. The theory in question is quantum electrodynamics. This theory was created by Dirac on the basis of what is surely a fiction, the definition of vacuum as an infinite collection of negative-energy electrons, the negative-energy sea. No negative-energy state was to remain unoccupied in the Dirac vacuum. On this basis, antimatter and the processes of pair creation and annihilation were first discovered. The sea was soon eliminated from the theory, in a way that left no interpretation of antimatter at the level of the 1-particle theory, and which seemed to require new principles (gauge covariance, microcausality) foreign to the canonical basis of quantum mechanics. One does not know why the negative-energy sea led to effectively the same theory, because, although this vacuum motivates normal-ordering in a natural way, it seems that there is no connection with gauge covariance and microcausality.

Saunders offers an explanation of these and other puzzling features of the relationship between field and 1-particle theories. It turns out that it is possible to set up the relativistic kinematics in precise correspondence to the non-relativistic case, that is through a canonical second quantization, with no reference to the negative-energy sea. This construction makes sense only for kinematic observables; nevertheless, it may be possible to read back from the field theory to a 1-particle analogue of dynamical behaviour. In this sense the relativistic theory appears as a quantum mechanics with a difference: the Hilbert space complex numbers are linear transformations, and they do not commute with all the usual operators. In particular, it is precisely the 'naïve' (non-relativistic) interpretation of complex numbers which demands the representation of the vacuum as the negative-energy sea.

Whither relativistic quantum theory? Few would claim that its conceptual basis is fixed or even robust. In this volume we are concerned with a concept that is *a priori* the most simple but empirically complex; the concept of vacuum comes as close to the synthetic *a priori* as any in contemporary physics. It is surely just for this reason that it plays such a dominant role in the free construction of ideas: witness the theory of 'pure space' described by Atiyah and Braam. Issues of this kind have a definite philosophical flavour, but, it must be admitted, the language is excessively mathematical. In this respect the attempt to reach a wider audience has a real pay-off. There is something new to be learned, when concepts can be made intelligible to the non-specialist. The paper that follows sets a high standard; Einstein was a master of the genre.

A final remark by way of introduction to this paper. In 'On the Ether' Einstein speaks for his generation—almost. A minor point is his revival of the ether terminology; more important, he cannot resist making a novel and radical suggestion, one that turned out to be completely fallacious. This conjecture is unfamiliar and may appear naïve: in effect, Einstein groups, together with the light quantum hypothesis, difficulties in the explanation of the earth's magnetic field. Both are considered challenges to the traditional concept of field. He suggests that the magnetic field may couple directly to rotating mass distributions. It is true that at the time no satisfactory explanation existed for the earth's magnetic field, but, granted the rudimentary state of magnetohydrodynamics and geophysics in the twenties, Einstein's concern is surprising.

But there is a background to this story. It is a lesser-known aspect of Einstein's work that in the period 1910–16 he undertook difficult and elaborate experiments on the determination of atomic g-factors, that he was an expert in contemporary gyrocompass design, and that he was active in seeking a connection between the zero-point energy, atomic structure, and the magnetization of material media. The expected result, which follows from the Lorentz theory in application to the Rutherford model (or equally to that of Bohr), is a g-factor of unity. Einstein, in collaboration with de Haas, had obtained a result of this order; their experiments concerned rotating magnetic media, and the mechanical torque set up on magnetization, as the current loops (postulated by Ampère) are oriented parallel to one another. However, subsequent results suggested a g-factor closer to two, a fact that received detailed discussion at the Solvay Conference of 1921. Einstein followed the debates closely; by 1924 evidence for the anomaly was all but conclusive. As a result, Ampère's hypothesis is in crisis and with it the established theory of magnetization. By this time Barnett, one of the leading experimentalists involved, had speculated that the converse process to the Einstein–de Haas effect might have established the magnetic field of the earth. He now considered that the anomalous g-factor might be related to the anomalous Zeeman effect. Characteristically, Einstein looks to a new length-scale for clues for the needed modification of electrodynamics: he considered the hitherto unexplained magnetic field of the earth. That same year Goudsmit and Uhlenbeck, with the introduction of electron spin, made an innovation more radical still. The complete theory of atomic g-factors rests on Dirac's 1928 synthesis of relativity and quantum theory, constituting its earliest and most important success. (For details I refer to Peter Galison's beautiful account in *How Experiments End*, University of Chicago Press, 1987: 27–74; for the contemporary theory of the earth's magnetic field see Ronald Menil and Michael McElhinny, *The Earth's Magnetic Field: Its History, Origin, and Planetary Perspective*, Academic Press, 1983.)

1
On the Ether

ALBERT EINSTEIN

If we are here going to talk about the ether, we are not, of course, talking about the physical or material ether of the mechanical theory of undulations, which is subject to the laws of Newtonian mechanics, to the points of which are attributed a certain velocity. This theoretical edifice has, I am convinced, finally played out its role since the setting up of the special theory of relativity. It is rather more generally a question of those kinds of things that are considered as physically real, which play a role in the causal nexus of physics, apart from the ponderable matter that consists of electrical elementary particles. Therefore, instead of speaking of an ether, one could equally well speak of physical qualities of space. Now one could take the position that all physical objects fall under this category, because in the final analysis in a theory of fields the ponderable matter, or the elementary particles that constitute this matter, also have to be considered as 'fields' of a particular kind, or as particular 'states' of the space. But one would have to agree that, at the present state of physics, such a point of view would be premature, because up to now all efforts directed to this aim in theoretical physics have led to failure. In the present situation we are *de facto* forced to make a distinction between matter and fields, while we hope that later generations will be able to overcome this dualistic concept, and replace it with a unitary one, such as the field theory of today has sought in vain.

It is generally assumed that Newtonian physics does not recognize an ether, and that it is the undulatory theory of light that first introduced this ubiquitous medium able to influence physical phenomena. But this is not the case. Newtonian mechanics has its 'ether' in the suggested sense, which, however, is called 'absolute space'. In order to understand this clearly, and at the same time to

Originally published as 'Über den Äther', *Schweizerische naturforschende Gesellschaft, Verhanflungen* (1924), 105: 85–93. This translation © S. W. Saunders 1991.

render the ether concept more precise, we have to go back a little further.

We consider first of all a branch of physics that manages without an ether, namely Euclidean geometry, which is conceived as the science of the possible ways of bringing bodies that are effectively rigid into contact with one another. (We will disregard the light rays which might otherwise be involved in the origin of the concepts and laws of geometry.) The laws for the positioning of rigid bodies, excluding relative motion, temperature, and deforming influences, such as they are laid down in idealized form in Euclidean geometry, can make do with the concept of a rigid body. Environmental influences of any kind, which are present independent of the bodies, which act upon the bodies, and which are to be considered as influencing the laws of positioning, are unknown to Euclidean geometry. The same is true of non-Euclidean geometries of constant curvature, if these are conceived as (possible) laws of nature for the positioning of bodies. It would be another matter if one considered it necessary to assume a geometry with variable curvature. This would mean that the possible contiguous positions of effectively rigid bodies in various different cases would be determined by the environmental influences. In the sense considered here, in this case one would have to say that such a theory employs an ether hypothesis. This ether would be a physical reality, as good as matter. If the laws of positioning could not be influenced by physical factors, such as the clustering or state of motion of bodies in the environment and so on, and were given once and for all, such an ether would have to be described as absolute (i.e. independent of the influence of any other object).

Just as the (physically interpreted) Euclidean geometry has no need of an ether, in the same way the kinematics or phoronomics of classical mechanics does not require one either. These laws have a clear sense in physics as long as one supposes that the influences assumed in special relativity regarding rulers and clocks do not exist.

It is otherwise in the mechanics of Galileo and Newton. The law of motion, 'mass × acceleration = force', contains not only a statement regarding material systems, but something more—even when, as in Newton's fundamental law of astronomy, the force is expressed through distances, i.e. through magnitudes, the real definitions of which can be based upon measurements with rigid bodies. For the real definition of acceleration cannot be based entirely on observations with rigid bodies and clocks. It cannot be referred back to the measurable distances of the points that constitute the mechanical system. For its definition one needs in addition a system of coordinates, respectively a reference body, in a suitable state of motion. If the state of motion of the system of coordinates is chosen differently, then with respect to these the Newtonian equations of motion will not be valid. In these equations, the environment in which the bodies move appears somehow implicitly as a real factor in the law of motion, alongside the actual bodies themselves and their distances from one another, which are definable in terms of measuring bodies. In Newton's science of motion, space has a physical reality, and this is in strict contrast to geometry and kinematics. We are going to call this physical reality, which enters into Newton's law of motion alongside the observable ponderable bodies, the 'ether of mechanics'. The fact that centrifugal effects arise in a (rotating) body, the material points of which do not change their distances from one another, shows that this ether is not to be supposed a phantasy of the Newtonian theory, but that there corresponds to the concept a certain reality in nature.

We can see that, for Newton, space was a physical reality, in spite of the peculiarly indirect manner in which this reality enters our understanding. Ernst Mach, who was the first person after Newton to subject Newtonian mechanics to a deep and searching analysis, understood this quite clearly. He sought to escape the hypothesis of the 'ether of mechanics' by explaining inertia in terms of the immediate interaction between the piece of matter under investigation and all other matter in the universe. This idea is logically possible, but, as a theory involving action-at-a-distance, it does not today merit serious consideration. We therefore have to consider the mechanical ether which Newton called 'Absolute Space' as some kind of physical reality. The term 'ether', on the other hand, must not lead us to understand something similar to ponderable matter, as in the physics of the nineteenth century.

If Newton called the space of physics 'absolute', he was thinking of yet another property of that which we call 'ether'. Each physical object influences and in general is influenced in turn by others. The latter, however, is not true of the ether of Newtonian mechanics. The inertia-producing property of this ether, in accordance with classical mechanics, is precisely *not* to be influenced, either by the configuration of matter, or by anything else. For this reason, one may call it 'absolute'.

That something real has to be conceived as the cause for the preference of an inertial system over a non-inertial system is a fact that physicists have only come to understand in recent years. Historically, the ether hypothesis, in its present-day form, arose out of the mechanical ether hypothesis of optics by way of sublimation. After long and fruitless efforts, one came to the conviction that light could not be explained as the motion of an elastic medium with inertia, that the electromagnetic fields of the Maxwellian theory cannot in general be explained in a mechanical way. Under this burden of failure, the electromagnetic fields were gradually considered as final, irreducible physical realities, which are not to be further explained as states of the ether. The only thing that remained to the ether of the mechanical theory was its definite state of motion. It represented, so to speak, an 'absolute rest'. If all inertial systems are on a par in the Newtonian mechanics, therefore also in the Maxwell–Lorentz theory, the state of motion of the preferred frame of coordinates (at rest with respect to the ether) appeared to be fully determined. One tacitly assumed that this preferred system would, at the same time, be an inertial system, i.e. that the principle of inertia would hold in relation to the electromagnetic ether.

There is a second way in which the rising tide of the Maxwell–Lorentz theory shifted still further the fundamental concepts of physicists. Once the electromagnetic fields had been conceived of as fundamental, irreducible entities, it seemed they were entitled to rob ponderable inertial mass of its fundamental significance in mechanics. It was concluded from the Maxwell equations that an electrically charged body in motion would be surrounded by a magnetic field the energy of which would, to a first approximation, depend on the square of the velocity. What could be more obvious than to conceive of all kinetic energy as electromagnetic energy? In this way one could hope to reduce mechanics to electromagnetism, having failed to refer electromagnetic processes back to mechanical ones. This appeared to be all the more promising as it became more and more likely that all ponderable matter was constituted of electrical elementary particles. At the same time, there were two difficulties which one could not master. First, the Maxwell–Lorentz equations could not explain how the electrical charge that constitutes an electrical elementary particle could exist in equilibrium in spite of the electromagnetic forces of repulsion. Second, the electromagnetic theory could not explain gravitation in a reasonably natural and satisfactory manner. In spite of all this, the consequences of the electromagnetic theory were so important that it was considered an utterly secure possession of physics—indeed, as one of its best founded acquisitions.

In this way the Maxwell–Lorentz theory finally influenced our understanding of the theoretical foundations of physics to such an extent that it led to the founding of the special theory of relativity. It was realized that the electromagnetic equations do not in truth determine a particular state of motion, but that, in accordance with these equations—just as in classical mechanics—there is an infinite manifold of coordinate systems, moving uniformly with respect to each other, and all on a par, so long as one applies suitable transformation formulae for the space coordinates and the time. It is well known that this realization brought about a deep modification of kinematics and dynamics as a result. The ether of electrodynamics now no longer had any special or particular state of motion. It had the effect, like the ether of classical mechanics, of giving preference not to a particular state of motion, but only to a particular state of acceleration. Because it was no longer possible to speak of simultaneous states in different places in the ether in any absolute sense, the ether became, so to speak, four-dimensional, because there was no objective arrangement of its space in accordance with time alone. Also, following the special theory of relativity, the ether was absolute, because its influence on inertia and light propagation was thought to be independent of physical influences of any kind. While in classical physics the geometry of bodies is presumed to be independent of the state of motion, in accordance with the special theory of relativity, the laws of Euclidean geometry for the positioning of bodies at rest in relationship to one another are applicable only if these bodies are in a state of rest relative to an inertial system;[1] this can easily be concluded from the so-called Lorentz contraction. Therefore the geometry of bodies is influenced by the ether as well as the dynamics.

The general theory of relativity removes a defect of classical dynamics: in the latter, inertia and weight appear as totally different manifestations, quite independent of one another, in spite of the fact that they are determined by the same body-constant, i.e. the mass. The theory of relativity overcomes this deficiency by determining the

[1] For example, in accordance with the special theory of relativity, the Euclidean geometry does not apply to a system of bodies that are at rest relative to one another, but which in their totality rotate in relation to an inertial system.

dynamical behaviour of the electrically neutral mass-point by means of the law of the geodesic line, in which the inertia and weight effects can no longer be distinguished. Thereby it attributes to the ether, varying from point to point, the metric and the dynamical properties of the points of matter, which in their turn are determined by physical factors, to wit the distribution of mass or energy respectively. The ether of the general theory of relativity therefore differs from that of classical mechanics or the special theory of relativity respectively, in so far as it is not 'absolute', but is determined in its locally variable properties by ponderable matter. This determination is complete if the universe is closed and spatially finite. The fact that the general theory of relativity has no preferred space–time coordinates which stand in a determinate relation to the metric is more a characteristic of the mathematical form of the theory than of its physical content.

Even the application of the formal apparatus of the general theory of relativity was not able to reduce all mass-inertia to electromagnetic fields or fields in general. Furthermore, in my opinion, we have not as yet succeeded in going beyond a superficial integration of the electromagnetic forces into the general scheme of relativity. The metric tensor which determines both gravitational and inertial phenomena on the one hand, and the tensor of the electromagnetic field on the other, still appear as fundamentally different expressions of the state of the ether; but their logical independence is probably more to be attributed to the imperfection of our theoretical edifice than to a complex structure of reality itself.

I admit that Weyl and Eddington have, by means of a generalization of Riemann geometry, found a mathematical system that allows both types of field to appear as though united under one single point of view. But the simplest field equations that are yielded by that theory do not appear to me to lead to any progress in the understanding of physics. Altogether it would today appear that we are much further away from an understanding of the fundamental laws of electromagnetism than it appeared at the beginning of the century. To support this opinion, I would here like briefly to point out the problem of the magnetic fields of the earth and sun as well as the problem of light quanta, which problems concern, so to speak, the large-scale structure and the fine structure of the electromagnetic field.

The earth and the sun have magnetic fields, the orientation and sense of which stand in approximate relationship to the axes of rotation of these heavenly bodies. In accordance with the Maxwell theory these fields could be produced by electrical currents which flow in the opposite direction to the rotational movement around the axes of the heavenly bodies. The sunspots too, which for good reasons are looked upon as vortices, possess analogous and very strong magnetic fields. But it is hard to imagine that, in all these cases, electrical conduction or convection currents of sufficient magnitude are really present. It rather looks as if cyclic movements of neutral masses are producing magnetic fields. The Maxwell theory, neither in its original form, nor as extended by the general theory of relativity, does not allow us to anticipate field generation of this kind. It would appear here that nature is pointing to a fundamental process which is not yet theoretically understood.[2]

If we have just dealt with a case where the field theory in its present shape does not appear to be adequate, the facts and ideas that together make up the quantum theory threaten to blow up the edifice of field theory altogether. Indeed, the arguments are growing that the light quantum should be considered a physical reality, and that the electromagnetic field may not be looked upon as an ultimate reality by means of which other physical objects can be explained. The theory of the Planck formula has already shown that the transmission of energy and impulse by means of radiation takes place in such a manner as if the latter consisted of atoms moving with the velocity of light c and with the energy $h\nu$, and with an impulse $h\nu/c$; by means of experiments on the scattering of X-rays by matter, Compton now shows that scattering events occur in which light quanta collide with electrons and transmit part of their energy to the latter, whereby the light quanta change their energy and direction. So much is factually certain: the X-rays undergo such changes of frequency in their scattering as are required by the quantum hypothesis, as predicted by Debye and Compton.

[2] The electrodynamic analogy would suggest the assumption of a relationship of the form $d\mathbf{H} = -C\,dm\,\mathbf{v} \times \mathbf{r}/r^3$, in which dm is a mass moving with the velocity $\mathbf{v}$, and $\mathbf{r}$, respectively $r = |\mathbf{r}|$, is the distance of the origin from this mass. (This formula can, however, at best be considered only for cyclic motion, and then only as a first approximation.) The relationship between the magnetic fields of the earth and of the sun is in this way correctly given as far as the order of magnitude is concerned. The constant C has the dimension (gravitational constant)$^{1/2}$/(speed of light). From this one can estimate the order of magnitude of the constant C. If one puts this numerical magnitude into the above formula, it will, applied to the rotating earth, give the right order of magnitude for the magnetic field. These relationships deserve consideration, but could be accidental.

Furthermore, a paper has recently appeared by the Indian scientist Bose, regarding the derivation of the Planck formula, which is particularly important for our theoretical understanding for the following reason. Hitherto, all derivations of Planck's formula have somewhere made use of the hypothesis of the undulatory structure of radiation; for example, the factor $8\pi\nu^2/c^3$ of this formula, in the well-known derivation of Ehrenfest and Debye, was obtained by counting the number of eigenvibrations of the cavity that occur in the frequency range $d\nu$. This counting, which was based on the concepts of the wave theory, is replaced by Bose by a gas-theoretical calculation, which he applies to a light quantum situated in the cavity in the manner of a molecule. The question now arises whether it would not one day be possible to connect the diffraction and interference phenomena to quantum theory in such a way that the field-like concepts of the theory would represent only the expressions of the interactions between quanta, whereby no longer would an independent physical reality be ascribed to the field.

The important fact that, according to the theory of Bohr, the frequency of the radiation is not determined by electrical masses that undergo periodical processes of the same frequency can only increase our doubts as to the independent reality of the undulatory field.

But even if these possibilities should mature into genuine theories, we will not be able to do without the ether in theoretical physics, i.e. a continuum which is equipped with physical properties; for the general theory of relativity, whose basic points of view physicists surely will always maintain, excludes direct distant action. But every contiguous action theory presumes continuous fields, and therefore also the existence of an 'ether'.

2

The Mass of the Classical Vacuum

R. PENROSE

There is something a bit paradoxical in the lessons classical physics has to teach us about the physical nature of matter. We may ask, What indeed *is* 'matter'? The commonsense reply might be that it is the real substance of which actual physical objects—the 'things' of this world—are composed. It is what you, I, and our houses are made of. How, then, does one actually quantify this substance? Our elementary physics textbooks provide us with Newton's clear answer: it is the *mass* of an object, or system of objects, that measures the quantity of matter that it contains. This, indeed, now seems right; there is no other physical quantity that can seriously compete with mass as the true measure of total substance. Moreover, it is *conserved*—so that the mass of any system whatever must be unchanging with time.

Yet Einstein's famous (1905) formula from special relativity,

$$E=mc^2,$$

tells us that mass and energy are interchangeable with one another. Mass is still conserved, but now it seems less clearly to be the true measure of actual substance. Energy, after all, depends upon the speed with which that substance is seen to be travelling. The energy of motion in an express train is considerable, but if we happen to be sitting in that train, then, according to our own reference frame, the train possesses no motion at all. The energy of that motion (though not the heat energy of the individual particles, nor the rest-mass energy of those particles) is now reduced to zero by our particular choice of frame. The total mass of the express train, being proportional to its energy, appears to be less to a traveller on the train than to someone who remains stationary on the ground as the train speeds by.

This article is based largely on a passage (the final section of ch. 5) from *The Emperor's New Mind*, by R. Penrose (Oxford University Press, 1989).

Here the discrepancy in the total mass remains relatively small because the train's speed is small compared with the speed of light. On the other hand, for a striking example where the effect of Einstein's mass–energy relation is at its most extreme, consider a π^0-meson. It is certainly a material particle, with a well-defined (positive) mass. After about 10^{-16} of a second, it decays—almost certainly into a pair of photons. In the frame of the π^0-meson, each photon carries away half the energy and, indeed, half of the π^0-meson's mass. Yet this mass is of the nebulous kind: pure energy. For if we were to travel rapidly in the direction of one or other of the photons, then we could reduce the mass–energy of that photon to as small a value as we please.

All this makes for a consistent picture of conserved mass, but it is not quite the one that we had before. Mass can still, in a sense, measure 'quantity of matter', but there has been a distinct change of viewpoint. The conserved quantity that takes over the role of mass is now the entire energy-momentum 4-vector. The time component,

$$p^0 = E/c = m_I c,$$

describes only the observer-dependent 'inertial mass' m_I (or, equivalently, the energy) of a particle or system relative to the observer.

This 4-vector has a (non-negative) Minkowskian norm m_R, defined by

$$c^2(m_R)^2 = E^2c^{-2} - \mathbf{p}\cdot\mathbf{p} = (p^0)^2 - (p^1)^2 - (p^2)^2 - (p^3)^2,$$

where $\mathbf{p} = (p^1, p^2, p^3)$ is the 3-momentum. The norm is the rest-mass m_R (or, equivalently, the rest energy $m_R c^2$). One might try to take the view that m_R would be a good measure of 'quantity of matter', since this is observer-independent. However, it is not additive: if a system splits into two, then the original rest-mass is not the sum of the resulting two rest-masses. Recall the π^0-meson decay considered above. The π^0-meson's rest-mass is non-zero, while the rest-masses of each of the two resulting photons is zero. (Any particle that moves with the speed of light must have rest-mass zero.) The additivity property, now in the sense of vector addition, holds for the entire 4-vector, and this vector must thus be our measure of 'quantity of matter'.

Think now of Maxwell's electromagnetic field. As Maxwell himself clearly pointed out (1865, 1873), this field carries energy. Thus, by $E = mc^2$, the field must also have mass. Thus, Maxwell's field is also matter! This must now certainly be accepted, since Maxwell's field is intimately involved in the forces that bind particles together to form atoms. There must be a substantial contribution to any body's mass from the electromagnetic fields within it. (Though, unfortunately, this contribution is incalculable, on present theory—which gives *infinity* as its provisional, but unhelpful, answer!)

The energy in an electromagnetic field can be described as an energy *density*, i.e. as energy per unit volume, which is the normal way of describing the energy for a continuous medium. It is the spatial *integral* of this density that provides the total energy of the system. The energy density is a component not of a 4-vector, but of a valence-2 tensor. If there are many continuous media present (e.g. quantum field descriptions of particles), then we have an energy density, and hence a corresponding tensor, for each one. We add all these tensors together to obtain a quantity

$$T_{ab},$$

referred to as the *energy momentum tensor* of the system.

What about Einstein's gravitational field? In many ways it resembles Maxwell's. As with Maxwell's theory, bodies in motion can emit waves, and like electromagnetic waves they travel with the speed of light and carry energy. Yet this energy is not measured in the standard way, which would be by the above energy momentum tensor. In Einstein's equation

$$R_{ab}-\tfrac{1}{2}Rg_{ab}=-8\pi GT_{ab}$$

(where G is Newton's gravitational constant, R_{ab} is the Ricci tensor, g_{ab} the metric tensor, and R the scalar curvature), the T_{ab} on the right is supposed to be describing the entire non-gravitational energy. In vacuum, which in Einstein's theory means in the absence of all physical fields except gravity, the energy-momentum tensor is zero, whence $R_{ab}=0$; but there can still be a gravitational field present. This gravitational (tidal distortion) field is described by the full *Riemann curvature tensor* R_{abcd}, which has a total of 20 components. The Ricci tensor has just ten components, and the remaining ten collect together in the form of another tensor,

$$C_{abcd},$$

called the *Weyl tensor*. In vacuum, such as inside a (pure) gravitational wave, the Weyl tensor still survives, and it can be thought of as describing the free gravitational field.

The Weyl tensor is not an appropriate object for directly describing gravitational energy, however. (It has too many indices, for example.) Nevertheless, gravity *does* contribute to the total mass–energy of a physical system. The simplest way of seeing this is to consider two masses. When they are far apart, the total energy of the system is somewhat greater than when they are close together, owing to the Newtonian potential energy contribution. Thus, by $E=mc^2$ they must have a slightly greater mass when they are far apart than when they are close together. The difference would have to come from the gravitational field in the space between the masses. But this cannot arise as an integral of the energy density locally defined in T_{ab} because that energy density is zero outside the masses. Also, as mentioned above, (pure) gravitational waves carry energy, yet the energy density throughout the waves is everywhere zero.

These problems are related to the fact that the 'conservation law'

$$\nabla^a T_{ab}=0$$

that is enjoyed by the energy-momentum tensor is a 'covariant' one (∇^a denoting covariant derivative), and does not give rise to the integral conservation law that one would like, namely one asserting that the total energy of a physical system is actually constant. In a well-known attempt to resolve these issues, Einstein introduced a quantity $\mathfrak{T}_{ab}$, referred to as the energy momentum *pseudo*-tensor, which was intended to take the energy of the gravitational field into account. This did give rise to an integral conservation law, but it suffered from the very serious drawback of depending heavily on the particular system of coordinates that happened to have been chosen for the problem at hand. The components of $\mathfrak{T}_{ab}$ therefore had no local physical meaning (a difficulty that was already appreciated by Einstein), and one certainly cannot take this pseudo-tensor description as providing the 'true' measure of the mass–energy distribution in the gravitational field.

One might take the view, nevertheless, that somehow the *curvature* of space–time—as measured by the surviving Weyl components of the curvature tensor—can still represent the 'stuff' of gravitational waves. But, as is indicated by the above arguments, gravitational energy is *non-local*, which is to say that one cannot determine what the measure of this energy is by merely examining the curvature of space–time in limited regions. The energy—and therefore the mass—of a gravitational field is a slippery eel indeed, and refuses to be pinned down in any clear location. Nevertheless, it must be taken seriously. It is certainly *there*, and has to be taken into account in order that the concept of mass can be conserved overall. There is even a good (and positive) measure of mass (due to Bondi 1960, Bondi *et al.* 1962, and Sachs 1962) which applies to gravitational waves, but the non-locality is such that it turns out that this measure can be non-zero in regions where the space–time is actually completely free of curvature. Indeed, the region between two impulsive bursts of gravitational radiation can be completely flat, but the total energy carried by the wave as a whole simply depends on the extent of this flat region, i.e. on the distance between the two bursts. (See Penrose 1966 and Penrose and Rindler 1986; this last reference also describes a 'quasi-local' definition which does take into account the non-local properties of gravitational energy in finite regions.) It seems that, if this mass–energy is to be located at all, it must in such cases be in this flat empty space—a region completely free of matter or fields of any kind. Our 'quantity of matter' is now either there, in the emptiest of empty regions, or it is nowhere at all.

Thus, we see that even before we need consider the mysterious effects of quantum theory, our theories of physics tell us that there is something very odd and counter-intuitive about the nature of matter. We cannot at all draw a clear dividing line between what we call 'matter' or 'substance' and what we call 'empty space'—supposedly, the voids entirely free of matter of any kind. Matter and space are not totally separate types of entity. Actual substance need not be clearly localized in space. These are hints that our treasured intuitive views as to the nature of physical reality are less close to the truth than one would have thought. The nebulous and non-local additional features that the quantum theory injects into our picture of the world greatly strengthen this conclusion, but such conclusions must already be drawn on the basis of classical theory. We must expect, also, that future theory will provide us with yet further shocks to our cherished intuitions.

References

Bondi, H. (1960), 'Gravitational Waves in General Relativity', *Nature* London, no. 186: 535.

Bondi, H., van der Burg, M. G. J., and Metzner, A. W. K. (1962), 'Gravitational Waves in General Relativity, VII: Waves from Axisymmetric Isolated Systems', *Proc. Roy. Soc. London* A269: 21–52.

Einstein, A. (1905), *Annalen der Physik* 17: 639.

Maxwell, J. C. (1865), 'A Dynamical Theory of the Electromagnetic Field', *Phil. Trans. Roy. Soc.* 155: 459–512.

—— (1873), *A Treatise on Electricity and Magnetism*, Oxford University Press.

Penrose, R. (1966), 'General-Relativistic Energy Flux and Elementary Optics', in B. Hoffman (ed.), *Perspectives in Geometry and Relativity*, Indiana University Press, Bloomington, pp. 259–74.

—— and Rindler, W. (1986), *Spinors and Space-Time*, Vol. 2: *Spinor and Twistor Methods in Space-Time Geometry*, Cambridge University Press, p. 427.

Sachs, R. K. (1962), 'Gravitational Waves in General Relativity, VIII: Waves in Asymptotically Flat Space-time', *Proc. Roy. Soc. London* A270: 103–26.

3

Reflections on Ether

SIMON SAUNDERS and HARVEY R. BROWN

1 Introduction

There are special difficulties which obstruct the understanding of both ether and the elimination of ether from physical science. To be sure, there is an *experimentum crucis*—the Michelson–Morley experiment—and there is a rival theory which better accounted for the facts; but we know that in the history of ideas life is not so simple. The ether was a very great idea, and it dominated more than a century of science; it did not vanish overnight.

Nevertheless, the concept of ether as formulated in the Lorentz theory had the ground cut from beneath it. The special theory of relativity obliterated the distinction between field and the ether as formulated by Lorentz. Just how special—or should we say *susceptible*—is the Lorentz theory to the new perspectives of relativity may be gleaned from Einstein's remark a half-century later: 'The physicist of the present generation regards the point of view achieved by Lorentz as the only possible one; at that time, however, it was a surprising and audacious step, without which the later development would not have been possible' (Einstein 1949: 35). The later development is not only special relativity; it is the concept of unitary field theory, in which the energy momentum distribution of the field is a reality just as good as ponderable matter. In fact, Einstein first derived the mass–energy relationship through an application of relativity to electromagnetism; it follows that the field has inertia. As a result, the ether *qua* substance appears wholly redundant; conversely, if the field has inertia, then the field *is* substance. The field is the reality.

The other side of the story is even more devastating for the Lorentz ether–field distinction. Under Lorentz, almost the only other property left to the ether was that it be stationary. The difficulties in reconciling this requirement to the null experiment of Michelson and Morley—using the techniques for handling relative motion of observer and ether provided by the continuum mechanics—is what led (eventually) to the Einsteinian analysis of space and time. In 1905 it appeared incontrovertible that, whether or not there existed a stationary ether, *precisely which* frame of reference was thereby distinguished could not possibly influence the laws of electromagnetism and was to all intents and purposes irrelevant to the description of reality.

The Lorentz ether was pole-axed. This is the situation *circa* 1905. But it is not, of course, the whole of the story.

In March of that same year Einstein conceived of the quantum theory. Over the next two years, and in splendid isolation, he pointed out its significance to the theory of radiation, to the Planck combinational theory of the black-body spectral distribution, and to the molecular theory of heat. His understanding of light quanta at this time was purely phenomenological; but after 1909, he turns with increasing commitment to a theoretical interpretation of light quanta as singularities in a *c*-number field. If he also suggests that each 'singular point be surrounded by a field of force', which when superimposed on one another will 'in their totality produce an undulatory force field little different from that in the sense of the electromagnetic theory of light' (Einstein 1909: 824–5), one sees no hint that these forces are allowed to be non-contiguous, that the forces themselves can be anything other than *c*-number fields. And to this Einstein remains committed for the rest of his life.

In Einstein's new critique of ether (1920 and 1924) we see a revival of the ether terminology; a different sense of ether, and with a new reference. The theory of general relativity had once again transformed the concept of a 'resting frame'. The class of inertial frames, locally transformed into one another by the Lorentz transformations, could be considered a local object, varying from neighbourhood to neighbourhood, and generated from the stress-energy tensor of the total matter, charge, and electromagnetism present in each neighbourhood. In this way the concept of 'resting frame' generalizes to what Einstein now calls the 'ether of mechanics'. It is no longer absolute, and it is effective; it is a part of reality. For precisely the same reasons, the electromagnetic field may be regarded as ether—ether in this general sense, a continuous structure subject to contiguous action, effected by and acting upon its physical environment. The ether of mechanics is a special case of this general kind of field-ether.

Einstein's resurrection of the ether concept brought no converts; these questions were swept aside by the hurricane impact of quantum mechanics. And in quantum mechanics, more properly quantum field theory, Einstein's strategy is completely undermined. He fears as much in the paper of 1924: 'the facts and ideas which together make up the quantum theory threaten to blow up the edifice of field theory altogether'. And from his perspective on ether, which we may suppose equivalent to field-ether, indeed his whole conceptual edifice is in smithereens.

There is, however, another history of ether, which in a number of respects foreshadowed the impact of the quantum theory on the field-ether. It is the representation of a *material* medium as an atomistic structure. This was precisely the 'audacious step' of Lorentz. When we realize that according to Maxwell and Hertz a material medium endowed with electromagnetic properties is ether, we see that Lorentz carried through a classical analogue of the quantization of a massive continuum; he atomized this ether—and in the process created the field-particle dualism inherited (and adapted) by Einstein.

Things are therefore more complicated than they look. Both the theory of space–time transformations and the distinction between field and particle hinged on the concept of ether. The former gave rise to the Minkowski geometry and the pseudo-Riemannian geometry of the ether of mechanics; the latter to the theory of quanta. In this process the Lorentz ether played a transitional role. It was itself born out of the atomization of material media, a sort of 'first quantization'. This ether is the classical field-ether, and it is atomized again in the quantization of the electromagnetic field; it is 'second-quantized'. We no longer have field-ether, we have quantum field-ether.

These are grand generalizations; to fix our ideas we must take a closer look at what Lorentz actually did. We find that his inspiration took root in a conflict no less central, the debate over contiguous action and action-at-a-distance.

2 Lorentz: Creating the Wave-Particle Duality

Recall that an essential mathematical innovation of Maxwell theory was the introduction of the displacement current. This step, and, as it turned out, the concept of electric waves, could be incorporated within the Continental tradition of action-at-a-distance theories. This was Lorentz's first contribution to electrodynamics; he took over Helmholtz's generalized definition of the magnetic potential—a definition which included that of Neumann, Weber, *and* Maxwell—and the theory of dielectric polarization to which it was applied, to deduce Snell's law and the laws of reflection.

This was in 1875; the Helmholtz theory was in 1870; yet Maxwell's theory appeared in 1861 and 1862. On the Continent Maxwell's theory met with resistance, obstructed by the obscure and fanciful ether on which it was based.[1] Helmholtz and Lorentz conceived of a medium, effectively a perfect insulator, which is polarized by the action of the net electric force **E** at each point. The resulting polarization **P** equals $\epsilon\mathbf{E}$, with ϵ the inductive capacity. Include this term in the source equation for the magnetic field *in vacuo*,

$$\nabla \times \mathbf{B} = \mu \mathbf{j} + \epsilon\, \partial \mathbf{E}/\partial t,$$

and one has the Maxwell theory. The essential difference between Maxwell and the Continental school is that the latter conceived of the vector potential **A** defined by the current as a mathematical artifice, not present in the deformation of the medium (except for that part of **A** due to the time variation of **P**), but produced by distance forces.

It was fundamental to the derivations of Helmholtz and Lorentz that there *was* a medium. There was no other reason to consider the action of distance forces at a point in empty space. But they already believed in the existence of *some* medium, because it was required of the wave theory of light. Not because action must be contiguous, but because of the reality of light, the medium must exist. But what sort of medium? For Lorentz the issue was this: is the electromagnetic ether, that is a medium the displacement and deformation of which produces electrical and magnetic effects as described by the Neumann–Weber–Maxwell potentials, a superior model to the elastic–solid ether of Fresnel? The latter was based on Newtonian laws, with central potentials; the former on the velocity-dependent potentials of electrodynamics—more precisely, of Neumann or Weber electrodynamics. The latter might prove effective where the elastic–solid theory had completely stalled.

[1] For references and further details see Whittaker (1910), D'Agostino (1975), McCormach (1970), and especially Hirosiga (1969).

Lorentz was no adherent to the doctrine of contiguous action. In his own words, 'In deriving the equation of motion of electricity, I largely follow Helmholtz. Like this physicist I will start from action at a distance; we shall thus have the advantage of founding the theory on the most direct interpretation of the facts' (1875: 30). Electrodynamical law should not be founded on metaphysical principles of contiguous action; nevertheless, it may be applied to a polarizable medium, and deformations and displacements in that medium will be local structures. The situation is not so different in continuum mechanics, where the Newtonian forces that operate, albeit over short distances, are nevertheless action-at-a-distance forces; as we shall see, Maxwell too had used distance forces in setting up his model of the ether.

Consider now Hertz's great discovery of 1888. Recall that Hertz had generated and detected electrical waves. This experiment appeared to confirm the existence of the electromagnetic ether, but by some it was taken to confirm the doctrine of contiguous action.[2] Lorentz did, of course, embrace this point of view, but not until 1891, and then for reasons that had little to do with the Hertz experiment.[3] In the mean time, he had already taken the first steps towards the separation of the ether from material media, a separation that was far from obvious from the point of view of Maxwell theory, but natural within the action-at-a-distance 'reconstruction' of wave propagation. For Maxwell, deformations within a material medium produce electromagnetic effects; material polarization and current (whatever, precisely, these might be) enter into the field equations. So too does the displacement current in the ether. If both ether and matter ultimately obey Newton's laws, they are one system, instantiating the electromagnetic field. But from the point of view of Neumann–Weber theory, electrical currents act at a distance on volume elements of the ether; if these currents and

[2] Cf. FitzGerald's remarks (1888: 558) in the opening address to the Association of Science: 'The year 1888 will be ever memorable as the year in which this great question has been experimentally decided by Hertz in Germany and, I hope, by others in England. It has been decided in favour of the hypothesis that these actions take place by means of an intervening medium.' Presumably FitzGerald was unaware of or unimpressed with the Helmholtz–Lorentz treatment of electromagnetic waves using distance forces.

[3] It is true that in 1891 Lorentz called the Hertz experiment 'the greatest triumph that Maxwell's theory has attained'. However, it scarcely figured in his detailed critique of the advantages of Maxwell theory over action-at-a-distance theories.

polarizations are associated with molecules, one has a clear separation of the role of the material medium (molecules) and the role of the ether.

It was essential to embrace some version of atomism. Hertz, who did not, also detected inconsistencies in Maxwell's own treatment of the ether, supposedly implementing a theory of contiguous action:

> Maxwell starts with the assumption of direct action-at-a-distance; he investigates the laws according to which hypothetical polarizations of the dielectric ether vary under the influence of such distance-forces; and he ends by asserting that these polarizations do really vary thus, but without being actually caused to do so by distance-forces. This procedure leaves behind it the unsatisfactory feeling that there must be something wrong about either the final result or the way which led to it. Another effect of this procedure is that the formulae retained a number of superfluous, and in a sense rudimentary, ideas which only possessed their proper significance in the older theory of direct action-at-a-distance. (Hertz 1893: 195–6)

The example Hertz cites is the concept of the displacement current. The ether displacement **D** is supposed to result from an applied electric force **E** on a portion of the ether. The two are proportional, as for an elastic medium. But what is **E**, if it is not a distance force? It cannot also be a displacement in the ether, for then one would have one displacement producing another, which is absurd. By this very critique Hertz illustrates the power of the action-at-a-distance perspective; it enables one to distinguish the field arising from a remote system (the applied field **E**) from the field set up by its action on a medium.

Hertz, in his positivist mode, developed a theory that eliminated unobservables—the potentials. Following Maxwell, he denied the distinction between the effects of material media and the effects of the ether. In modern terms, he advocated the use of the fields $\mathbf{D}=\epsilon_0\mathbf{E}+\mathbf{P}$, and $\mathbf{H}=\mathbf{B}/\mu_0-\mathbf{M}$, the net or *phenomenological* fields. In particular, within a body 'the lines of force simply represent a symbol for special conditions of matter'; in just the same way, *in vacuo*, the lines of force are a special condition of ether.

A further comment of Hertz is noteworthy. With reference to Maxwell's equations,

> After these equations are once found, it no longer appears expedient to deduce them (in accordance with the historical course) from conjectures as to the electric and magnetic constitution of the ether and the nature of the acting

> forces—all these things being entirely unknown. Rather is it expedient to start from these equations *in search of such further conjectures* respecting the constitution of the ether. (Hertz 1893: 201; emphasis ours.)[4]

This was essentially Einstein's perspective on Maxwell's theory with one important difference: there could be no ether, for there could be no absolute resting frame. Were there then no further conjectures that one might seek to make concerning the constitution—not of the ether, but of radiation? Was Maxwell theory *complete*?

Return to Lorentz. Already in his Ph.D. thesis—that is, in 1875—he had sharply distinguished the polarization of molecules from the polarization of the ether. This he did in order to explain the fact that the dielectric constants and refractive indices of gases are always close to unity. He made the simple assumption that the dominant electromagnetic contribution comes from the ether, and not from the molecules of a dilute system. Indeed, if the inductive capacity of the molecules is proportional to their number density N, so that $\epsilon = \epsilon_0 + \alpha N$, then from the expression $n^2 = (1 + 4\pi\epsilon)/(1 + 4\pi\epsilon_0)$ (where n is the refractive index), and assuming the density ρ is proportional to N, there follows the law of Arago and Biot, $(n_2 - 1)/\rho = \text{constant}$. In particular, as $\rho \rightarrow 0$ it follows that $n \rightarrow 1$.

Lorentz speculated that, quite generally, 'one should take into account first the ether, and then the molecules lying within it. Then the distance, size, and form of the latter enter into consideration, from which the explanation of dispersion and the plane of polarization will probably result.'[5] The theory of dispersion which followed three years later amply fulfilled his expectations. The basic mechanism for the dispersion was in fact already known; the essential point is to treat each molecule within the dielectric as elastically bound, but driven by the incident wave. Maxwell and Sellmeier independently hit on this mechanism in the context of the elastic solid ether; Lorentz developed a detailed electrodynamic model in which the molecules were treated as charged simple harmonic oscillators.

The model was as follows. The molecule is considered to occupy a small void within the ether, upon which it acts by distance forces. This action, and the effects of a surface charge induced on the boundary of the void, modify the polarization induced in the molecule. In this way

[4] In his much quoted remark, 'Maxwell's theory is Maxwell's equations', Hertz did not mean to *deny* the existence of ether.

[5] Lorentz (1875: 87–8); trans. in Hirosiga (1969: 172).

the polarizability of individual molecules may be defined in terms of the polarization set up by an 'effective' electric field. On the other hand, the external electric field is also related to the polarization field (by the electric susceptibility χ); this quantity is just $\kappa-1$, where κ is the specific inductive capacity or dielectric constant, the square of the refractive index. Eliminating the fields leads to a relationship between the refractive index and the polarizability of the molecule.

This formula depends on the geometry of the cavity and the mass of the molecule, but not on the magnetic moment of the molecule or the intermolecular spacing. Neither does it depend on the frequency of the external field. At this point Lorentz turned to the dependence of the polarizability on the internal structure of the molecule; if this varies with the frequency of the effective field, a dispersive effect results. To this end Lorentz supposed that the molecule can be modelled as an elastically bound charge e of mass m with characteristic frequency ω_0. The polarizability κ is given by $\mathbf{E}=\kappa\mathbf{P}$, and (according to this model) $\mathbf{P}=e\mathbf{x}$, where $\mathbf{x}$ is the displacement of the charge from the centre of oscillation. The equation of motion for the oscillator is $m\ddot{\mathbf{x}}=e\mathbf{E}-k\mathbf{x}$, with $\omega_0^2=k/m$. When $\mathbf{E}$ is an oscillatory field of frequency ω, one obtains the solution $\mathbf{x}=e\mathbf{E}/(k-m\omega^2)$, so that $\kappa=(k-m\omega^2)/e^2$; under these assumptions, one has a theory of dispersion.

Lorentz's inaugural lecture to the University of Leyden was given that same year; its title, 'Molecular Theories in Physics'. In the following decade Lorentz made a number of contributions to the rapidly developing kinetic theory of gases, including a note on Boltzmann's H-theorem. On electromagnetism he published only on the magnetic potential (*à la* Weber and Neumann), on the Hall effect, and on one other topic: the theory of aberration. As we shall see, this could only have pushed him further to embrace an absolute distinction between the electromagnetic ether and material media. But it was not until 1892 that he elaborated the definitive form of what is now called the Lorentz theory; it was, essentially, a more developed form of the dispersion theory of 1878 formulated according to the Maxwell theory, with a new application to the Fizeau experiment.

Lorentz made a public conversion to Maxwell theory in 1891. As arguments in its favour, he considered a technical difficulty of the Helmholtz theory (which need not concern us), and the following qualitative argument. There is the question of energy: where, precisely, does electromagnetic energy reside? It was a claim repeatedly made by Maxwell that the ether itself could be set in motion, and that in this motion resides kinetic energy. Lorentz was impressed by the fact that the potential energy in the action-at-a-distance theory depended on the velocities of charged particles 'and nevertheless is not a kinetic energy in the ordinary sense of the word'. He immediately continues: 'This is a major difficulty: although it is simpler to denote energy dependent on velocity as kinetic energy, here it is not the case. In Maxwell's conception it is considered as such, and in this I find the first reason to give his theory precedence.'[6]

Hereafter Lorentz was to speak of a 'velocity vector' assigned to each point of the ether; these ether velocities, together with the electric displacements, provided the generalized coordinates in Lorentz's Lagrangian treatment on which the 1892 theory was founded. But Lorentz made no attempt to develop a detailed mechanical model; as always, his motives were pragmatic. As we shall see, the Lagrangian theory provided a framework with just enough of a mechanical flavour to obtain the force law of the field on charged particles in motion (the Lorentz force).

It was this sort of analysis that was so conspicuously absent from Hertz's phenomenological approach. It was not just that Hertz had little sympathy for atomism (Lorentz actually derived the force law for a continuous charge distribution); as he complained in the introduction, Hertz 'hardly concerns himself with the relation of electromagnetic actions to the laws of ordinary mechanics'. But there is another connection between Lorentz's great thesis of 1892 and the Hertz theory of 1890; the latter's reluctance to distinguish between that action of the medium defined by the matter content and that owing to ether led to the view that the ether should be considered co-moving with a material body. For consider the consequences of a stationary ether, or an ether incompletely dragged:

> If now we wish to adapt our theory to this view, we have to regard the electromagnetic conditions of the ether and of the tangible matter at every point in space as being in a certain sense independent of each other. Electromagnetic phenomena in bodies in motion would then belong to that class of phenomena which cannot be satisfactorily treated without the introduction of at least two directed magnitudes for the electric and two for the magnetic state. (Hertz 1893: 242)

Now Lorentz's earlier disposition to distant action had led to the

[6] Lorentz (1891: 93–4); translated in Hirosiga (1969: 183–4).

fruitful conception of material media as systems of charged molecules. Adherence to contiguous action might prohibit a particulate view of the non-material ether, but one could still retain a molecular theory of matter. One of Hertz's 'two ethers' might be considered the electric configuration of matter; the remaining ether, the purely electromagnetic one, is most simply considered *at rest.*

3 The Resting Ether and the Possibility of Electrodynamics

The connection is not quite watertight; intuitively, even given an atomistic theory, the ether might be set in motion by the collective influence of a large number of particles. It was here that Lorentz had the particular genius to see that the ether might have no other interaction with matter than that conditioned by its charge distribution—conditioned, in particular, by the Maxwell source equations,

$$\nabla \times \mathbf{H} = 4\pi \mathbf{J}$$
$$\nabla \cdot \mathbf{D} = \rho .$$

(**J** is the sum of the conduction and displacement current densities.) There is no sign here that the ether must be set in bulk motion. Previously, within the action-at-a-distance theory, Lorentz had conceived of the interaction between ether and matter with the molecule at the centre of a cavity within the ether; if the interaction is to be contiguous the cavity must vanish, but if it is purely electromagnetic, the only action on the ether is as given by Maxwell. In this way—mediated, so to speak, by the action-at-a-distance theory—Lorentz conceived of the Maxwell notion of contiguous action without the attendant notions of impenetrability and contact forces. If ether and matter need interact only electromagnetically, there is no reason why they should not permeate each other. In this way Lorentz prepares for a precarious synthesis of field and particle concepts, in which each is an independent reality. The substantive ether—*qua* the microscopic structure of media—splits into two parts. The one may be identified with the atomistic structure of matter; the other with the substratum of pure electromagnetism. Because the two are coupled only through charge, there is no reason why the motion of the one should give rise to a motion of the other, so the substantive electromagnetic ether is also stationary—an absolute resting frame.

The question of the motion of ether had already a long history. Fresnel had proposed that the elastic–solid ether be stationary (unaffected by the motion of the earth) in order to explain the phenomenon of stellar aberration. There is no 'ether drift'. Lorentz's first skirmish with the question of the motion of ether concerned a competing theory due to George Stokes, who had proposed that the velocity of the earth relative to the ether varied continuously from zero (at the earth's surface) to some constant value at sufficiently large distances. On this basis Stokes had obtained agreement with the aberration experiments. Lorentz, in his study of 1886, came to the conclusion that Stokes's conception was inconsistent; the relative velocity could not vanish everywhere at the earth's surface. Dropping this assumption, but retaining a variation in ether velocity, Lorentz obtained a consistent theory midway between that of Fresnel and Stokes. But he considered nothing was gained thereby; Fresnel's theory was more simple, the hypothesis of no ether drift was to be preferred.

But what of the motion of the ether in the interior of a moving medium? Here Fresnel had supposed that the passage of light induces vibrations in that medium; if in motion relative to the ether with velocity v, he further supposed there exists an 'excess' of ether within the medium which partakes of that velocity (while the remaining ether stays at rest)—the hypothesis of 'ether drag'. On this basis he deduced that a wave of velocity u relative to the medium will propagate at the velocity $u+fv$, where f is the *drag coefficient*; in particular, if the ether moves *with* the medium, f will be unity.

On the somewhat *ad hoc* assumption that the density of ether in a body is proportional to the square of the refractive index, Fresnel obtained the value $f=1-1/n^2$. Remarkably, this prediction was confirmed by the null result of Airy's experiment of 1871, an experiment suggested by Fresnel in 1818. In this experiment a star is held fixed at the centre of field of view of a telescope, while the telescope is filled with water. No change was required in the orientation of the telescope. Even more impressive, Fizeau had in 1851 directly measured the Fresnel drag coefficient, by studying the fringe-shift using an interferometer in which coherent beams travelled in opposite directions through flowing water. The observed value of 0.48 compared favourably with the theoretical value of 0.43.

Returning to 1892, it was a weakness of the Hertz theory that it did not account for the Fizeau experiment; but Lorentz chose to challenge the very concept of ether drag:[7]

> In the memoir where Mr. Hertz deals with bodies in motion, he assumes that the ether contained in them moves with the bodies. Now, optical phenomena have shown that this is not always the case. I therefore would like to find out the laws governing the electric motions in bodies which traverse the ether without dragging it, and it seems to me difficult to attain this end without a theoretical idea as a guide. (Lorentz 1892: 168)

Here lies the importance of the concept of a stationary ether to the development of the electrodynamics; if the ether is dragged by matter, then it is dragged by atoms, and the question of the action of the ether on moving charges cannot be posed in any simple way. There is a more simple assumption yet: there is *no* ether drag. On this basis, it was possible to set up the electrodynamics.

The Lorentz force law followed from the Lagrangian theory of the ether together with two hypotheses: the first is the relationship $\nabla \cdot \mathbf{D} = \rho$, which is to hold everywhere within the interior of a charged particle, and the second is the definition of current density within the particle: $\mathbf{J} = \rho\mathbf{v} + \partial\mathbf{D}/\partial t$. Here $\mathbf{v}$ is the velocity at each point within the particle, relative to the stationary ether. Modelling charged particles as extended bodies, however, brought with it the problem of their structure and stability. Here Lorentz simply assumed that the bodies are rigid. On the other hand, an immediate pay-off is that the source expressions are smooth functions of the coordinates just like the fields; variation of the charge distribution in a given volume automatically involves a change in the field energy and current density in that volume, together with the virtual work done by the field on the charge. Including these terms leads to the expression for the total force (the Lorentz force):

$$\mathbf{F} = 4\pi c^2 \int \rho\mathbf{D} \ d\tau + \int \rho(\mathbf{v} \times \mathbf{B}) \, d\tau.$$

The power of this theory was remarkable. With it Lorentz recovered the Maxwell–Hertz theory, results of the old action-at-a-distance theories, and the Lorentz–Lorenz dispersion formula. His new derivation of the Fresnel drag coefficient is, however, of particular interest. In this treatment, the incident radiation induces a

[7] Lorentz may have had in mind the unsatisfactory implication of the Fresnel theory in the context of dispersion; it seems that the 'extent' to which ether is dragged (within one and the same moving body) depends on the frequency of light considered. Is there a different ether for each wavelength of light?

polarization field in the dielectric, which interferes with the incident wave to give rise to a wave of increased phase velocity. Only the polarization current $\partial\mathbf{P}/\partial t$ is changed by the motion of the medium, the displacement current $\epsilon_0\partial\mathbf{E}/\partial t$ is unaffected. Since the polarization current would not exist in the absence of the medium, its dependence on the velocity of the medium is to be expected. To first order in the velocity of the medium, the drag coefficient is proportional to the ratio of the polarization and displacement currents, in accordance with Fresnel's law.[8]

This result is a remarkable achievement, but it should be noted that from a technical point of view the atomistic structure of the material medium plays no essential role; one only needs the idea that the polarization field is associated with the medium, the electric field with a stationary space. That is, we need only to accept the 'two ethers' rejected by Hertz. But it is atomism that makes sense of this idea; the two ethers may coincide in space, but the one is attached to the atoms in motion, the other to the space through which they move, the stationary ether. In this picture the material ether is a dynamical system of charged particles; distinctive properties of ether, the constitutive equations that relate the 'partial fields' **P** and **M** to the external **E** and **B**, are to be determined purely from the properties of this dynamical system. Further, the ether as 'support' to the fields (in this case the polarization and magnetic fields) can be considered as the charge-current distribution itself, therefore the atomic configuration of the medium.

The partial fields admit a dual description. On the one hand, they describe charge and current distributions; on the other, fields on a par with **E** and **B**.[9] But if the charge and current distributions are particulate, they are also the consequence of electromagnetic interactions between these particles; the material ether has therefore a subordinate role to the purely electromagnetic ether. Compare, for example, with Whittaker, writing in 1910:

[8] For a treatment along these lines, see Panofsky and Phillips (1972: 193–5, 279–80). There is also the much simpler derivation on the basis of relativity; for light moving with speed $u=c/n$, through a medium of refractive index n, which is itself moving with speed v, an observer at rest will conclude the light is moving with velocity $V=(c/n+v)/(1+v/nc)$ (by application of the velocity addition law). To first order in v/c, this is just $c/n+(1-1/n^2)v=u+fv$.

[9] Cf. the discussion in Panofsky and Phillips (1972: 462–4) and their closing remark: 'In fact, an understanding of the *dual* physical function of **P** and **M** is the principal requirement for clarity in the classical theory of electric and magnetic media.'

> It may be remarked that Maxwell made no distinction between stress in the material dielectric and stress in the ether: indeed, so long as it was supposed that material bodies when displaced carry the contained aether along with them, no distinction was possible. In the modifications of Maxwell's theory which were developed many years afterwards by his followers, stresses corresponding to those introduced by Maxwell were assigned to the aether, as distinct from ponderable matter; and it was assumed that the only stresses set up in material bodies by the electromagnetic field are produced indirectly; they may be calculated by the methods of the theory of elasticity, from a knowledge of the ponderomotive forces exerted on the electric charges connected with the bodies. (Whittaker 1910: 272)

The material ether appeared derivative to the stationary ether, because electromagnetic forces determine the dynamical behaviour of the atoms; but these arise in the *stationary* ether. The idea that the atomistic structure is itself an aspect to the field—most especially a massive field—did not of course come into consideration. (We shall take up this theme shortly.[10])

But the purely electromagnetic ether, the stationary ether of Lorentz, became increasingly detached from mechanics. In the same year, 1892, Helmholtz showed that Maxwell's expression for the electromagnetic stress evaluated over a closed surface does not vanish; the stationary ether should not, therefore, be considered in equilibrium. Lorentz's (1895) response was simply to deny that electromagnetic stress *acts on ether*—although a manifestation of ether, an action of ether on material bodies, this stress need exert no reaction on ether: Newton's third law is denied. This step is entirely consistent with the basic principle—there is no other action on ether than that arising from charge, as determined by Maxwell. With this, the immobility of the Lorentz ether is assured; at the same time, its detailed description can have no simple relationship to the laws of mechanics. The stage is set for the 'electromagnetic world view' of the turn of the century, the process by which, failing a reduction of the structure of ether to mechanics, mechanics is to be subsumed to the theory of ether.

[10] It is the concept of the massive field that is anachronistic; the idea that material particles (and in particular charge) could be constructed as structures in ether had been around for some time. The theory of vortex filaments, a remarkable achievement due to Helmholtz in 1858, was highly influential and played a prominent role in ether models by the close of the century. However, these unitary theories of ether do not fit very happily with the Lorentz concept of the stationary ether and they are not considered in the text.

In this matter an immediate difficulty concerns the structure of the electron. It was another of the empirical successes of the Lorentz theory that it could account for the so-called normal Zeeman effect. The phenomenon is as follows. Spectral lines are split when there is an external magnetic field, and are circularly polarized. The lines appear as doublets in the direction parallel to the field, and as triplets perpendicular to the field. Lorentz's (1895) explanation led to a spectroscopic determination of the charge to mass ratio of the electron—a value in agreement with Thomson's measurement that same year on the deflection of cathode rays. But as evidence accumulated for the reality of electrons, so too did theoretical difficulties for the simple rigid sphere envisaged by Lorentz. In this problem—the application of electrodynamics to material bodies—the reconstruction of mechanics faced severe test.

The details of these developments need not concern us; it is sufficient that they ended in failure. The other great difficulty confronting the Lorentz theory was the Michelson experiment of 1881, a null result to second order in powers of v/c on the detection of 'ether wind'. The absence of first-order effects—for example the demonstration due to Clausius that electrostatic phenomena could not be effected by the ether wind—was well known; as Potier and Veltmann showed in the early 1870s, *no* experiment sensitive only to effects of first order could detect the ether wind.[11] In a short supplement to his 1892 article, Lorentz explored further the curious cancellations to this order in the context of the electron theory. But the Michelson experiment, which had been repeated in 1887 in collaboration with Morley with a more accurate but null result, resisted such manoeuvres. Something was amiss with the concept of the stationary ether.

We are now approaching a crucial phase in the elimination of ether, but this is a part of our story that has been told many times, and we shall be correspondingly brief. In the same year, 1892, Lorentz proposed the contraction hypothesis; this hypothesis, also suggested by George FitzGerald in 1889, eliminated the second-order effect. All force fields were considered to increase in a medium moving with respect to the ether, in such a way as to contract the dimensions of the medium. Three years later Lorentz presented a more complete

[11] It seems the work of Potier and Veltmann has been largely forgotten; see Newburgh and Costa de Beauregard (1975) for discussion and references.

discussion, including the notion of 'local time', and gave a proof of a 'correspondence', valid to first order in v/c, between the Maxwell equations in a coordinate system stationary with respect to the ether and those for a moving body in a co-moving coordinate system.[12] The coordinate transformations in question were of the form $x'=x-vt$, $t'=t-vx/c^2$; to explain the (second-order) Michelson–Morley result, the contraction hypothesis must be assumed in addition. In 1899 Lorentz wrote down the Lorentz transformations (up to a scale factor); in 1904 he applied them to prove second-order invariance of the field equations. (Use of the non-relativistic velocity addition formula prevented him from obtaining strict invariance.)

Meanwhile, in 1898 Poincaré had questioned the assumptions underlying the traditional concept of time: 'The simultaneity of two events or the order of their succession, as well as the equality of two time intervals, must be defined in such a way that the statements of the natural laws be as simple as possible' (Poincaré 1898). In 1904 he returned to this idea in the context of electromagnetism and the problem of ether: he formulated the relativity postulate and conjectured on an 'entirely new mechanics, which would be characterised above all by the fact that no velocity would be able to exceed that of light . . .' (1904: 316). In June the following year he gave an essentially complete discussion of the invariance of the Maxwell theory under the Lorentz group, and proposed that all of the forces of physics should transform in the same way, and that therefore Newton's theory of gravity must be amended.

Einstein did not know of Poincaré's papers; he knew of Lorentz's work only up to the 1895 paper. Relativity theory, as the mathematical statement of the covariance of physical laws, was proposed simultaneously by Einstein and Poincaré,[13] and it was

[12] Lorentz considered local time subordinate to 'true' (Newtonian) time; by 1899 he was to explain the time dilation as a consequence of the real-length contractions of physical clocks (e.g. the pendulum). This interpretation was not available within the theory of corresponding states of 1895, which led to time dilation alone. As it happens, the transformations of 1895 already contain an invariance principle for the velocity of light: although not isotropic (except to first order in v/c), the speed of a light ray propagating in the direction of relative motion of two frames of reference is strictly invariant. Einstein was familiar with this theory of Lorentz, but not with his later modifications.

[13] The question of precedence must contend with the following circumstance. The first to discover the form invariance of the free Maxwell equations was Woldemar Voigt in 1887. It is a remarkable and most unfortunate fact that this paper of Voigt, which contained the first statement of the Lorentz transformations and the proof of covariance, remained completely unknown throughout this period.

anticipated by Lorentz. However, Poincaré persisted in regarding the length contraction as a real physical effect, and even founded the theory upon this hypothesis; Lorentz did not adapt to the changed situation presented by relativity for many years. It was Einstein alone who followed the implications of relativity to their logical conclusion.

4 Special Relativity and the Electromagnetic Field

Einstein's presentation of the theory was uniquely his own; it rested on two principles:

1 *The relativity principle* The same laws of electrodynamics and optics will be valid for all frames of reference for which the equations of mechanics hold good.
2 *The principle of the constancy of the velocity of light* Light is always propagated in empty space with a definite velocity c which is independent of the state of motion of the emitting body.

Einstein referred to the special theory as a *theory of principle*, founded on hypotheses that are suggested by experience.[14] For all that, the temptation to read into these laws a purely operational significance must be resisted. For example, it would be in error to interpret the relativity principle as the assertion that no electrodynamic or optical experiment can distinguish a privileged inertial frame of reference; this would permit the interpretation of, e.g., length contraction as a real physical effect, experimentally undetectable in a co-moving frame because of the contraction of all measuring rods in that frame. (This was precisely Poincaré's reading of the theory.) In what sense is the relativity postulate founded on experience? A better reading of Einstein's thinking is that he considered the fundamental equations of electromagnetism and optics as themselves phenomenological; in this sense their form invariance may also be considered as a generalization from experience.

The main development of Einstein's ideas is as follows. From the two postulates, applied to an operational analysis of the definition of simultaneity, Einstein established the relativity of lengths and times,

[14] Einstein (1949); other examples cited by Einstein were the impossibility of a *perpetuum mobile*, as a foundation for thermodynamics, and Galileo's law of inertia. Theories of principle are contrasted to *constructive theories*, which rest on hypotheses of a more fundamental but conjectural kind.

and with it the admissibility of transformations that mix space and time coordinates. Assuming homogeneity of space and time, he inferred that the equations must be linear,[15] and from isotropy, and by repeated application of the two postulates, derived the Lorentz transformations.[16] Following a discussion of the interpretation of the contraction and dilation effects, and the derivation of the velocity addition law, he turned to the electrodynamical part of the theory. Here he established the covariance of the free Maxwell–Hertz equations, and compared the old and new interpretations of the Lorentz force law; he then gave relativistic treatments of the Doppler effect and aberration, the transformation law for the amplitudes of light waves, and also for the associated energy and pressure; finally, he proved covariance for the complete Maxwell–Hertz equations in the presence of sources. A final section treated the dynamics of slowly accelerated charges, essentially providing the basis for a relativistic dynamics, independent of electromagnetism. In this section he derived expressions for the transverse and longitudinal mass of a particle. A short note submitted three months later established the mass equivalence of energy.

So much for the bare facts. For our purposes, two considerations are important. First, the stationary ether, *qua* privileged frame of reference, has been eliminated. It is undermined at a stroke, for the relativity postulate militates directly against it. Although the conflict between the Lorentz theory and the Michelson–Morley experiment is removed (to all orders in v/c), there is nothing privileged about the ether frame of reference; one might, so to speak, adopt *any* inertial frame as 'the stationary ether'.[17] But second, there is nothing to be

[15] It seems the first explicit proof that the linearity of the coordinate transformations is a consequence of the homogeneity of space and time is to be found in Berzi and Gorini (1969).

[16] On the basis of the two postulates, Einstein inferred that the light velocity must have the same numerical value in all inertial frames. This inference is non-trivial; see Tzanakis and Kyritsis (1984), Brown and Maia (1990).

[17] Einstein summarized the implications of relativity for the ether postulate in these terms on several occasions. See, for example, his remarks at Salzburg: 'The principle of relativity . . . declares that all laws of nature in relation to a coordinate system K' which is in uniform motion relative to the ether, would be the same as the corresponding laws in relation to a coordinate system K which is at rest relative to the ether. If this is so, however, we have just as much reason to suppose the ether to be at rest relative to K' as to K. It then becomes altogether unnatural to distinguish one of the two coordinate systems K, K' by introducing an ether which is at rest relative to it. From this it follows that one can only reach a satisfactory theory when one can dispense with the ether hypothesis' (Einstein 1909: 819).

gained by always referring the phenomena to such a resting frame. This point deserves some amplification. Witness, for example, Einstein's opening remarks:

> It is known that Maxwell's electrodynamics—as usually understood at the present time—when applied to moving bodies, leads to asymmetries, which do not appear to be inherent in the phenomena. Take, for example, the reciprocal electrodynamic action of a magnet and a conductor. The observable phenomenon here depends only on the relative motion of the conductor and the magnet, whereas the customary view draws a sharp distinction between the two cases in which either the one or the other of these bodies is in motion. For if the magnet is in motion and the conductor at rest, there arises in the neighbourhood of the magnet an electric field with a certain definite energy, producing a current at the places where parts of the conductor are situated. But if the magnet is stationary and the conductor in motion, no electric field arises in the neighbourhood of the magnet. In the conductor, however, we find an electromotive force, to which in itself there is no corresponding energy, but which gives rise—assuming equality of relative motion in the two cases discussed—to electric currents of the same path and intensity as those produced by the electric forces in the former case.
>
> Examples of this sort, together with the unsuccessful attempts to discover any motion of the earth relatively to the 'light medium', suggest that the phenomena of electrodynamics as well as of mechanics possess no properties corresponding to the idea of absolute rest. (Einstein 1905*b*: 891)

Einstein immediately goes on to state the relativity postulate. Now, the thrust of his remarks is as follows. The phenomenon is symmetric between wire and magnet, but the theoretical description is not: in the one case there is an electric field with an associated energy, in the other a pure magnetic field. Yet both act in the same way—inducing a current in the wire. So Einstein is saying that there should be a way of looking at electrodynamic *theory* so as to preserve the symmetry evident in the *phenomenon*—for in the latter there is nothing corresponding to the idea of absolute rest.

The process of induction is, from an atomistic standpoint, nothing but the operation of the Lorentz force on charged particles; Einstein's comparison of the old and new theoretical descriptions bears directly on this question:

> 1. If a unit electrical point charge is in motion in an electromagnetic field, there acts upon it, in addition to the electric force, an 'electromotive force' which, if we neglect the terms multiplied by the second and higher powers of v/c, is equal to the vector-product of the velocity of the charge and the magnetic force, divided by the velocity of light. (Old manner of expression.)

> 2. If a unit electrical point charge is in motion in an electromagnetic field, the force acting upon it is equal to the electric force which is present at the locality of the charge, and which we ascertain by transformation of the field to a system of co-ordinates at rest relatively to the electrical charge. (New manner of expression.) (Einstein 1905*b*: 54)

In this way, those features of electromagnetic theory that seem to pick out a privileged frame of reference are no longer taken as evidence of such a frame; rather, they are made frame-dependent, 'auxillary concepts'—they are not fundamental, not part of the objective reality. These are complications introduced by the choice of frame—and by the concept of ether.

Lorentz did not see these as complications. Indeed, for a man who for forty years had laboured with just these 'auxillary concepts'—displacement fields, electromotive forces, polarization fields, and the magnetic force—they were the very stuff of electromagnetic theory. In 1913 he remarked:

> According to Einstein, it has no meaning to speak of motion relative to the ether. He likewise denies the existence of absolute simultaneity.
>
> It is certainly remarkable that these relativity concepts, also those concerning time, have found such a rapid acceptance.
>
> The acceptance of these concepts belongs mainly to epistemology . . . It is certain, however, that it depends to a large extent on the way one is accustomed to think whether one is most attracted to one or another interpretation. As far as this lecturer is concerned, he finds a certain satisfaction in the older interpretations, according to which the ether possesses at least some substantiality, space and time can be sharply separated, and simultaneity without further specification can be spoken of. In regard to this last point, one may perhaps appeal to our ability of imagining arbitrarily large velocities. In that way, one comes very close to the concept of absolute simultaneity.
>
> Finally, it should be noted that the daring assertion that one can never observe velocities larger than the velocity of light contains a hypothetical restriction of what is accessible to us, which cannot be accepted without some reservation. (Lorentz 1913: 23)

Einstein makes no reference to a further function of ether, what was for Lorentz a principal (and superior) feature of the Maxwell theory: it is the ether that carries the energy of the field. Maxwell and Lorentz considered that this energy could only be a version of kinetic energy (that there must be some substance, in the motion of which inheres kinetic energy). But in relativity, where energy itself constitutes mass, this argument appears simply archaic. In the special theory even the distinction between rest-mass energy and kinetic energy is artificial, and with it the distinction between the ether as substance, the support of the field, and the field itself.

For Lorentz, however, there remained other functions of ether, functions that were not so readily assimilated to the field itself:

> Although, as the foregoing exposition shows, the role of this medium has continually gained in significance, on the other hand the attempts to penetrate further into its nature have fallen increasingly into the background. Since the development of the theory of relativity many physicists have indeed gone so far as to no longer speak of an ether at all, but just of the electromagnetic field propagating itself in space. The question, to what extent this is expedient, must here be left undecided. In part it reduces to a verbal question; if one does not want to say that all forces are transmitted *through the ether*, or will nevertheless have to explain, according to the theory of relativity, that they are all propagated in space with the speed of light. (Lorentz 1914: 333)

Here Lorentz refers to an explanatory role of ether, in accounting for the constancy of the velocity of propagation, relative to ether, and independent of the speed of the source. We look in vain to Einstein's theory for a similar explanation—it is the bald statement of the explanandum which constitutes the second postulate of the theory. For this reason Pauli and Sommerfeld dubbed it the 'ether postulate'—for it demands by fiat that feature of the propagation of light which is most simply and directly explained by the concept of ether.[18] As formulated by Minkowski, however, one has an explanation for light-constancy: space–time has a geometric structure, and this structure entails the existence of an invariant velocity. The fact that light happens to travel at this limiting velocity then legitimates the second postulate.

It is known that Einstein was reluctant to adopt the formal perspective of the Minkowski theory;[19] had he done so, his recognition that the velocity of light could not be constant in the presence of gravitational fields (Einstein 1907) might have led more

[18] Against this, one might argue that it is this feature of electromagnetic law that is explained by the continuum mechanics. But Lorentz had dispensed with a mechanical foundation to the theory of the electromagnetic ether. There may lurk the intuition that the ether *explains* the independence of light speed from the source, but the details of this explanation fall back on the continuum mechanics—or nothing at all.

[19] 'Superfluous learnedness'; remark to V. Bargmann, reported in Pais (1982: 152).

directly to the interpretation of gravity as varying space–time geometry, a point of view that he did not possess prior to 1912.[20] Even then, it was only following an analysis of static fields that Einstein realized that gravity might be associated with curvature in space as well as in time. With these developments, the space–time geometry appears as a physical structure (cf. Einstein's 'ether of mechanics'); Lorentz's stationary ether, its role usurped by the Minkowski geometry, is not so much eliminated as given a new name.

But ether at this level of abstraction played no part in the thinking of Lorentz. What of that other function of ether, the substantive support of radiation? The ether in this capacity is a dynamical structure which underpins the field. Lorentz has his own theory to blame if, by making the ether immobile, by denying the second law, this structure appeared remote and irrelevant to the electrodynamics. Nevertheless, it seems clear enough that Lorentz still hankered after this function of ether.

As well he might. So long as there is an ether, there is a substructure to electromagnetism; there is a definite direction in which the theory must be deepened, even if the tools and techniques of the continuum mechanics can no longer be applied. In this context it matters little that there is, in fact, no privileged reference frame—no more than it matters to the concept of the microscopic structure of a *material* medium.[21] It seems that Lorentz lamented the absence not only of the ether concept, but of any basis by which to deepen the theory, that is to conceive of the field at an instant of time, and to then consider: what kind of thing is this field; what is it *made of*? To respond—it is what it is, as described by the solutions to the Maxwell–Lorentz–Einstein equations of electrodynamics—simply terminates the discussion. Electromagnetism is made a closed book.

[20] For these developments see e.g. Pais (1982: ch. 11). In late 1907 Einstein discovered the equivalence principle and applied it to the time dilation associated with accelerating coordinate systems; for the next five years he was concerned to reformulate the analysis of simultaneity in the presence of gravity, and to understand the incompatibility of the equivalence principle with light constancy. The recognition that the Lorentz transformations cannot hold globally, that they (as also the equivalence principle) must hold only locally, appeared in 1912, where he uses tensor notation for the first time. Minkowski did not live to see the flowering of his geometric perspective; he died in January 1909 at the age of 45.

[21] Good relativists all, we often enough need to think of a system at an instant in time (initial data, field configuration, spacelike hypersurface, etc.).

5 Light Quanta and the *c*-Number Field

In this scenario, the field-ether, a system of coupled *c*-number functions, is the beginning and end of the story. These functions are supposed to describe the field completely. The field-particle duality introduced by Lorentz becomes absolute. In 1905 Einstein could therefore put the matter thus:

> A profound formal distinction exists between the theoretical concepts which physicists have formed regarding gases and other ponderable bodies and the Maxwellian theory. Whilst we consider the state of a body to be completely determined by the positions and velocities of a very large, yet finite, number of atoms and electrons, we make use of continuous spatial functions to describe the electromagnetic state of a given volume, and a finite number of parameters cannot be regarded as sufficient for the complete determination of such a state. (Einstein 1905a: 367)

If Einstein was happy to banish the electromagnetic ether, the atomization of the *material* medium, as prescribed by Lorentz and (in the context of thermodynamics) in accordance with Boltzmann, was for him an empirical issue. To bring the atomic hypothesis to the point of experimental test, he had developed the appropriate tools—the statistical thermodynamics. This work had preoccupied him for more than four years; since he was examining the relationship of atomism to the continuum (or thermodynamic) limit, it is hardly surprising that it should bear most crucially on this new continuum reality, the electromagnetic field. He found that the 'profound formal distinction' leads directly to the conclusion that no thermal equilibrium can exist between radiation and matter. The synthesis of particles and fields conceived by Lorentz is inconsistent with the laws of thermodynamics.

Only if there is no ether, no further microscopic structure to radiation, is this conclusion forced. There is a very good reason why neither Rayleigh, nor Jeans, nor Lorentz saw in the Rayleigh–Jeans spectral distribution cause for concern; they were not concerned because at the microscopic or ultra-microscopic level one may always suppose that the structure of the ether comes into play in new and unforeseen ways. Einstein's development of relativity appears all the more remarkable for its timing;

the relativity paper was in June, the light quantum paper in March.[22]

There is, however, a possible compromise between the two theories: to consider that the electromagnetic field is a *phenomenological* reality. And indeed, the light quantum hypothesis removes the contradiction with thermodynamics just as an ether hypothesis would: by modifying electrodynamic law at high frequencies, equivalently, at short distances. Light quanta take on the role of ether as the microstructure to the field; the Maxwell–Hertz theory appears as the phenomenological limit. As Einstein put it,

> The wave theory of light, which operates with continuous spatial functions, has worked well in the representation of purely optical phenomena and will probably never be replaced by another theory: it should be kept in mind, however, the optical observations refer to time averages rather than instantaneous values. In spite of the complete experimental confirmation of the theory as applied to diffraction, reflection, refraction, dispersion, etc., it is still conceivable that the theory of light which operates with continuous spatial functions may lead to contradictions with experience when it is applied to the phenomena of emission and transformation of light. (Einstein 1905a: 368)

Remarks in a similar vein are scattered throughout his papers; Einstein considered the light quantum hypothesis fully compatible with the free field theory. Further, it is consistent with the special theory. Einstein used the neutral term 'light complex' throughout the relativity paper; on obtaining the transformation equations for energy and frequency, he makes the pointed remark: 'it is remarkable that the energy and the frequency of a light complex vary with the state of motion of the observer in accordance with the same law'. The September paper of 1905, in which the mass-equivalence of energy was derived, ends with the statement, 'If the theory corresponds to the facts, radiation conveys inertia between the emitting and absorbing bodies.' In his 1909 paper he is more specific: although the special thoery of relativity 'does not in any way change our conception of the

[22] B. Hoffmann (1982) has considered the question, Why did not Einstein develop the emission theory of light along the lines of Ritz (1908), in order to explain the null result of the Michelson–Morely experiment? As suggested in the text, Einstein conceded to the Maxwell–Hertz equations phenomenological validity; even if the emission of light is modelled on a ballistic theory of particles (in which case the null result of Michelson and Morely follows immediately), from this point of view it would still be necessary to account for the experiment in terms of the phenomenological wave equations.

structure of radiation, in particular the distribution of energy in the space through which radiation is travelling', nevertheless,

> The electric fields which constitute the light . . . [appear] as independent structures which are emitted from the light sources in accordance with Newton's emission theory of light. As in the latter theory, space which is not permeated by radiation and which is free of ponderable matter would be genuinely empty. (Einstein 1909: 820)

Einstein makes covert appeal to a simple intuition; since relativity requires that the field carry inertia, then the emission of light is at once the emission of inertia. If the ether was required to provide a substantive support for the field, it is now redundant, because the energy distribution of the field is itself substance. In the absence of radiation and ponderable matter there is no substance, there is empty space.

What Einstein does not consider, however, is the relationship of the light quanta to the ether. From the point of view of quantum field theory, the analogy with the atomization by Lorentz of material electromagnetic media is all but exact. Massive fields describe the medium as continuous in the classical limit;[23] the atomistic structure is described by the particle representation of a quantum field. It is exactly the same for the radiation field. In particular, it is *quanta*, whatever their mass, that appear as the substantive substructure of the field.

Einstein's early views on light quanta are consistent with this perspective; for example, in defence of the light quantum hypothesis, he remarks:

> It is by no means certain that we would have to change many of the assumptions regarding interference phenomena as one has at the moment. I would compare [the introduction of light quanta] to the process of the molecularization of the carriers of the electrostatic field. The field produced by atomized electrical particles is not essentially different from earlier assumptions and it is not impossible that in radiation theory something similar will take place. (Einstein 1909: 826)

It was Einstein's work on the quantum theory of gases in 1924 that alerted him to the importance of the wave theory of de Broglie. In late

[23] Conventional media of any complexity will naturally be constructed from bound states of the elementary quanta of the field; it must be admitted that there is no simple or direct route from the field equations to the phenomenological continuum. Nevertheless, the medium is atomistic only because the quantum field is particulate.

1925 Schrödinger, attempting to interpret the Bose–Einstein statistics in the framework of Boltzmann, was led in this way to the wave theory. The Schrödinger field, conceived of as a classical field (this was roughly Schrödinger's point of view), was quantized by Jordan and Klein the following year. But to read into Einstein's writings the suggestion that the field is in some sense *only* a phenomenology of quanta would be a mistake. On the contrary: for Einstein, the field is the support of the quanta. We already detect such a view in the Salzburg address of 1909; it is the basis of what became for Einstein an *idèe fixe*: the light quanta are to be considered singularities in the *c*-number field.

> In the first instance the idea would appear to me the most natural, that the occurrence of the electromagnetic fields of the light is tied to singular (or individual) points, in the same way as the occurrence of electrostatic fields in accordance with the theory of electrons. It is not impossible that in such a theory the entire energy of the electromagnetic field could be considered as localized in these singularities, just as in the old distant action theory. I would imagine every such singular point to be surrounded by a field of force which has the essential character of a plane wave, and the amplitude of which decreases with distance from the singular point. If many such singularities are present in intervals of small distances from one another, distances which are small compared with the dimensions of the field of force of a singular point, then the fields of force will be superposed on one another, and will in their totality produce an undulatory force field which will be little different from an undulatory field in the sense of the present electromagnetic theory of light. (Einstein 1909: 824)[24]

What is implicit is that the 'fields of force' are *c*-number fields. Presumably they may not, however, carry energy, since it is the light quanta (singularities) that comprise the energy distribution within the field. Where the field is continuous presumably it is not field-ether, because it does not transport inertia. We have a new concept of field.

It is a remarkable fact that something similar was to emerge in the theory of gravity. We have already indicated the turn of events that was to follow the extension of relativity to gravity; space–time geometry becomes a dynamical entity in its own right. Einstein's later

[24] So far as we know, Einstein never commented on the implications of the later work of Natanson and Ehrenfest for this neoclassical conception of light. (Recall that Natanson and Ehrenfest in 1911 concluded that light quanta must be non-individuated objects, in effect anticipating the Bose statistics.) If, however, the light quanta are singularities in the field, then their interchange has no meaning, for one obtains the same field configuration.

critique of ether runs something as follows: considering the ether not as a single privileged frame, but as a privileged class of frames (the local inertial frames), this ether enters into the 'causal nexus' of the world; it determines the kinematics of mass–energy distributions and is determined in turn by these distributions.

In field theory, the ether of mechanics is the metric tensor, or perhaps it is the Ricci curvature—in either case it is a tensor field. But what is of interest is that it is not a local field quantity in the sense of electromagnetism (or any other particle fields), for there is no reasonable definition of a local distribution of stress-energy. The tendency to read into general relativity a theory of physical fields is in conflict with the geometric basis of the theory.

Einstein was fully aware of this difficulty. Shortly after his completion of the general relativity, he attempted to define conservation laws for total energy and momentum; in the process (Einstein 1916), he was led to a local stress-energy tensor for the gravitational field (as opposed to other particles or fields that might be present) which was *not* a tensor density. Subsequently it was recognized that this stress-energy was indefinite, and that all its components could be given arbitrary values at a given point (Einstein 1918).

There were many motives for attempting a synthesis of gravity and electromagnetism; the peculiar circumstance, that neither the ether of mechanics nor the *c*-number fields of light in a singularity-free domain admits a reasonable definition of a stress-energy tensor, could only have reinforced Einstein's conviction that the synthesis was necessary. Each appears the key to the other; geometry determines the meaning of space–time relationships, and the quantum theory governs the localization of energy.

We know well enough that Einstein laboured on this task for the rest of his life. In effect, he took the residue from the 'first-quantized' ether—what remains of the ether when the material atomistic structure of the medium is removed—and quantized it again. Within the framework of contiguous action and the *c*-number field, there is again a residue; it may be unified (perhaps) with the ether of mechanics.

Einstein's commitment to the *c*-number field as the exhaustive description of the world scarcely wavered. Let us see how this emerges in his 1924 critique of ether. In this standpoint, the local reality is virtually defined as ether; however, the atomistic structure of physical media is explicitly eliminated from this concept:

> It is rather more generally a question of those kinds of things which are considered as physically real, which play a role in the causal nexus of physics, *apart from the ponderable matter which consists of electrical elementary particles*. Therefore, instead of speaking of an ether, one could equally well speak of physical qualities of space. Now one could take the position that all physical objects fall under this category, because in the final analysis in a theory of fields the ponderable matter, or the elementary particles which constitute this matter, also have to be considered as 'fields' of a particular kind, or as particular 'states' of the space. But one would have to agree that, at the present state of physics, such a point of view would be premature, because up to now all efforts directed to this aim in theoretical physics have led to failure. In the present situation we are *de facto* forced to make a distinction between matter and fields, whilst we hope that later generations will be able to overcome this dualistic concept, and replace it with a unitary one, such as the field theory of today has sought in vain. (Einstein 1924: 85; emphasis ours; trans. on p. 13 above)

This dualistic conception is a precondition of the discussion; so too, it seems, is the commitment to *contiguous action*:

> We can see that for Newton, space was a physical reality, in spite of the peculiarly indirect manner in which this reality enters our understanding. Ernst Mach, who was the first person after Newton to subject Newtonian mechanics to a deep and searching analysis, understood this quite clearly. He sought to escape the hypothesis of the 'ether of mechanics' by explaining inertia in terms of the immediate interaction between the piece of matter under investigation, and all other matter in the universe. *This idea is logically possible but, as a theory involving action-at-a-distance, it does not today merit serious consideration.* We therefore have to consider the mechanical ether which Newton called 'absolute space' as some kind of physical reality. The term 'ether', on the other hand, must not lead us to understand something similar to ponderable matter, as in the physics of the 19th century. (Einstein 1924: 88; emphasis ours; trans. on p. 15 above)

Just as contiguous action binds us to the ether of mechanics as physical reality, it seems also to demand the ultimate (and not merely phenomenological) reality of the fields, whether they be introduced to couple the light quanta, or as descriptions of local geometry. In Einstein's final remark, we have an article of faith:

> But even if these possibilities should mature into genuine theories, we will not be able to do without the ether in theoretical physics, i.e. a continuum which is equipped with physical properties; for the general theory of relativity, whose basic points of view physicists will surely always maintain, excludes

direct distant action. But every contiguous action theory presumes continuous fields, and therefore also the existence of an 'ether'. (Einstein 1924: 93; trans. on p. 20 above)

6 The Ether and Quantum Field Theory

In 1924 Einstein, as also Bohr, considered *c*-number functions to describe exhaustively an ultimate reality. In the electromagnetic case, these functions controlled the interaction of light quanta, which were perhaps singularities in the field, and gave rise to the observed fields. For Bohr these electromagnetic fields were already absolute, for he denied the existence of light quanta. They were both wrong. *C*-number functions do not refer to some ultimate stuff in space–time; they are what Einstein more prudently considered them in 1905—statistical averages, or phenomenological fields, over some other kind of processes, some other kind of entities. Just what these entities are remains unknown—remains the goal of physics to find out. But a first approximation is the concept of light quanta and elementary particles. Consider again the 'two ethers' of Hertz; Lorentz made of the one an ensemble of interacting charged particles, the particle interpretation of massive quantum fields coupled to electromagnetism. Einstein made of the other an ensemble of massless uncharged particles, the photon interpretation of the electromagnetic field coupled to matter. Rather than consider the ether whatever remains when this atomistic structure is removed—the ether of pure vacuum—we may equally consider the ether transmuted into atomism.[25]

What is unsatisfactory about this reading of the ether concept is that the *raison d'être* of ether was the mediation of interactions. So long as only the material ether is atomized, the electromagnetic ether that remains does exactly that. But if we atomize both material and electromagnetic ether, there seems nothing to mediate between the respective quanta.

Quantum theory is atomism—with a difference. It is not so clear

[25] In some sense, quantum field theory is a unitary theory which encompasses wave and particle theories. Here we adopt the perspective that bears most closely on the concept of ether, and the transformation of ether by Lorentz. This perspective was surely anathema to Einstein; however, so far as we know, at no time did he remark on the duality of field and particle as it presents itself in quantum field theory.

that quanta are local objects. This line of interpretation seems definitely to lead to non-local interactions. Consider the following propositions:

1 *Separability* Any observable for a physical system has a definite value, which corresponds to the outcome of any measurement on the physical system designed to measure that observable.
2 *Locality* These values are independent of the occurrence and details of any kind of perturbation suffered by remote systems.

It is well known (Bell 1964) that any theory of physical systems satisfying these conditions is inconsistent with quantum mechanics. Therefore quanta cannot be considered physical systems of this kind; they are non-local systems, or they are local systems subject to non-local law. This is true regardless of whether or not one sets up a mediating field (the field could not mediate the quanta by contiguous action); it seems that Einstein's programme, if successful, must contradict quantum mechanics.

Intuitively, it seems clear that quantum theory does not permit any reasonable expression of the notion of a local 'causal nexus'. If such a thing exists at all in the world, it cannot be exhaustively described by a *c*-number field, certainly not if the quanta are singularities in the field. On a more speculative note, it seems very likely that the notion of a *c*-number function that literally refers—and exhaustively describes—fundamental reality is just incoherent, and not merely false. We shall not know how to make sense of such ideas—or how to establish that they *lack* sense—until we know better just how mathematics can literally refer to the world at all, rather than to the results of measurement (as in the notion of phenomenological fields). We are far from any understanding of this kind—but with quantum theory we have new perspectives.

Consider, for example, elementary quantum mechanics, specifically in the Schrödinger representation on an L^2 function space. In this form quantum mechanics appears as a dualistic theory; the particle aspect appears primarily in the context of measurement theory (which we shall pass over), and in the decomposition of the Hilbert space of the system, for example as an n-fold tensor product, for a system of n particles. (We shall come on to this in a moment.) The notion of a physical substructure to quantum systems no longer gives rise to a phenomenological interpretation of the wave function, because the wave function does not directly refer to a physical entity.

We have learnt of other ways of deepening the theory than by attributing a 'granularity' or suchlike to the L^2 functions themselves; we have learnt to dispense with an ether.

Some examples: first, in a given application we may suppose that additional degrees of freedom exist, and additional potentials and couplings (a methodology that derives from the mechanical foundation of quantum mechanics); second, explicit or hidden symmetries may exist which deepen the description (a new methodology, essentially the creation of Einstein and Dirac); and third, the function spaces used may contain a non-trivial topology (a contemporary focus of research, but also the creation of Dirac with his theory of the monopole).

Quantum field theory stems from the first of these strategies: it was developed as a particular way of effecting a decomposition of the Hilbert space, its representation as *Fock space*. The substructure to radiation—and this applies equally to massive fields—is implicit in the replacement of *c*-number functions by operator-valued fields. The choice of representation of the field provides the physical description of this substructure, and we recover the phenomenological fields in the classical limit. The choice of a Fock representation ensures that the interpretation of this substructure is particulate. Elementary quantum mechanics is, in effect, the explicit theory of this substructure.

In this way, the ether of quantum fields may be considered the theory of the relationship of systems to subsystems, the definition of the irreducible parts of a composite whole. We have come a long way from the concept of a medium; the relationship in question is no longer contiguity.[26] In general relativity contiguity is defined by the usual topology of $\mathbb{R}^4$, but the way the tangent spaces fit together is the business of the space–time metric, Einstein's ether of mechanics. A

[26] In certain approaches to classical and quantum mechanics topology plays a fundamental role. For example, any approach based on measure theory (e.g. the Mackey theory) has a natural topology, as does any approach based on a Borel G-space representation of a space–time symmetry group. (As Weil showed in 1938, the topology of the group is completely determined by the Borel structure.) At best, however, one learns that the position operators, etc., satisfy continuity properties induced from the action of the group. This is satisfactory, and perhaps even necessary, when the system is in some suitable state corresponding to these operators, but equally, it may be an artefact of Hilbert space theory. In the lattice theoretic approach, for example, one does not have a natural definition of topology. (See Saunders 1989: pt. 2 for discussion and references.)

synthesis of gravity and quantum theory must bring into relationship these two ethers.

In this respect the so-called 'Unruh effect' is of particular interest. This concerns the apparent observer-dependence of the particle interpretation of fields on a fixed background metric, and the impossibility of defining a global particle interpretation unless high space–time symmetries are present. On the one hand, the particle interpretation is reminiscent of the 'auxiliary concepts' relegated by Einstein to a subordinate role—which can only be defined by the choice of a frame of reference. Fulling originally formulated the difficulty in Minkowski space; the particle interpretation depends on the decomposition of the field into positive and negative frequency parts, and this decomposition may be effected in inequivalent ways, relative to non-inertial frames of reference.[27] On the other hand, we may suppose that it is precisely the 'ether of mechanics' that should provide an observer-independent decomposition of the field, but that as yet we can formulate this decomposition only in the simplest situations.

These are only some of the avenues before us. What appears most remarkable is the apparent continuity with concepts of ether, albeit in fragmented and more abstract forms. One still remains: the ether as *empty space*.

7 The Ether and Vacuum

We have paid only passing attention to the question of the ether in its ground state, equivalently, radiation-free empty space, equivalently, *vacuum*. It is in consideration of the vacuum that the traditional concept of ether appears most problematic. In the Maxwell sense—what remains when we have removed everything that *can* be removed (therefore radiation but not gravity can be excluded)—the vacuum has structure in classical general relativity. By the same criterion, since radiation *can* be excluded, ether as microstructure to radiation

[27] What is required is a field of timelike isometries, everywhere orthogonal to a family of spacelike hypersurfaces. For Minkowski space, the integral curves of the generator of boosts satisfy this condition; they may be considered the world-lines of constantly accelerated observers, for whom the decomposition in question may be considered in some sense 'natural'. Unruh's analysis of the response of an idealized particle detector in constant acceleration has motivated the view that the differing particle interpretations are equally real, but relative to the acceleration of the observer. See Fulling (1973); Unruh (1974).

would appear to have no existence in the electromagnetic vacuum. Einstein appeared to adopt this point of view on the basis of the special theory of relativity—the zero-valued field has no inertia, hence nothing is present in the pure void.[28]

This conclusion seems even more natural on a naïve interpretation of quanta—the interpretation of radiation as a boson gas. But in quantum field theory, there are two ways in which this conception is deficient. First, there are the zero-point fields; the field excitation cannot be everywhere zero, in the state of particle number zero, so that we cannot after all exclude all electromagnetic activity from a region of space. Second, there is the more subtle feature of quantum theory to which we have already alluded. There is a generalization of the concept of medium; a particle interpretation of a quantum field defines a relationship of system to subsystem which makes no reference to contiguity or locality. This is a rather abstract idea, but it is surely central to the peculiarities of quantum theory; further, this decomposition is non-trivial even when one determines only that state in which no quanta are present. The decomposition must be specified *in order* to define the vacuum. The ether of vacuum has, therefore, some existence—but it is a pale and ghostly shadow of its old self.

On a Lorentzian view, there would have remained the quiescent ether; the distinction between this and that field configuration in which all the fields have the value zero appears vanishingly small, yet in this difference resides the fundamental methodology of nineteenth-century physics. The dominant heuristic pursued by ether theorists—Fresnel, Maxwell, and Lorentz among them—was the application of the methods of continuum mechanics to the ether; it was essential, for the construction of the mathematical equations, to consider the response of the quiescent ether to an applied force or torque. Indeed, the ether was a tool for the understanding of both the mathematics

[28] The same conclusion follows from the traditional concept of ether, the ether of the continuum mechanics. This ether is the 'support' of the field, the excitation of which constitutes the field. In vacuum, the ether is quiescent—but an independent reality. This is true of the metrical field of general relativity, Einstein's 'ether of mechanics'; considered as space–time curvature, this ether often appears in a static role, the 'container'—the medium—of material structures. It is easy to think of gravitational waves as 'ripples' *in* space–time curvature. But we may not consider radiation as 'ripples' in the static electric and magnetic fields, for these fields do not exist in pure vacuum. The metrical field, by contrast, can no more vanish in the vacuum state than can space–time itself, a point stressed by Einstein (1920: 21).

and the phenomena, and in the former case its structure at rest (quiescent) determines its evolution when excited by external forces. There *are* echoes here in modern field theory; specification of the fields as everywhere zero carries with it at least the statement of tensoral type of the fields. In the free case this is enough to fix the field equations completely. (It is in this sense that Minkowski space explains free electromagnetic field theory.) But the ether also provided guidance in the construction of interactions; it is no coincidence that this is just the area in which contemporary theory must fall back on what is little more than a mathematical aesthetic ('minimal coupling'). It is a fundamental limitation of the field–many-particle equivalence established by Dirac that the particle interpretation of a quantum field, most especially for the vacuum, does not assist in the construction of interactions; if this particle interpretation is the substructure to the field, our new ether for old is unhappily impoverished.

But there is the third strategy by which quantum and classical theories alike can be deepened, and that concerns the topology of the *c*-number field solutions. In a nonlinear theory, the zero-valued fields may not satisfy the field equations. In that case the vacuum defined as the field configuration of minimum energy is non-trivial; it has a topological structure, for in general no deformation in the (time-zero) fields will permit one to pass from one vacuum to another without passing through configurations of infinite energy. Considerations of this kind underlie familiar phenomena such as the Bohm–Aharanov effect.

In a suitable regime of energy, the quantum theory of fields of this kind (for the most part Yang–Mills) is supposed to describe quanta. The interactions among these quanta will be determined in part by the vacuum with respect to which they are defined. The vacuum in such a theory determines the phenomenological laws, just as did the ether in the theory of electromagnetism.

Earlier the material ether of ponderable media was peremptorily dismissed from discussion. The vacuum, a space from which everything that can be removed *has* been removed, contains no such matter. But consider the following analogy. A linear (i.e. elastic) mechanical medium will support a certain phenomenology (wave fields, various kinds of stress, linear response coefficients, etc.). This phenomenology may count as 'filled space', its absence as 'empty space'. In particular, empty space may be considered as the no-phonon state (no quanta of sound), or perhaps as the zero-temperature state, the state of minimum energy with respect to the (fixed) atomic structure of the medium. Excite the medium beyond a certain limit and we obtain a nonlinear response; excite it further, and the atomic structure breaks down. If the medium cools, we may see a new atomic structure, a new medium with quite different properties—we will see a new 'vacuum'.

We have described processes that occur in dying stars and the formation of new planets; these 'vacua' are the material ethers of Maxwell and Hertz. Similar processes with no atomic disassociation are a part of everyday life; in fact, nowhere are such phenomena more complex and more intricately related than in biology. Above all, the concept of ether engages a distinction that becomes yet more central, and more profound: the distinction between fundamental and phenomenological law.

References

Bell, J. S. (1964), 'On the Einstein–Podolsky–Rosen Paradox', *Physics*, 1: 195–200.

Berzi, V. and Gorini, V. (1969), 'Reciprocity Principle and the Lorentz Transformations', *J. Math. Phys.* 10: 1518–24.

Brown, H. and Maia, A. (1990), 'Light-speed Constancy *versus* Light-speed Invariance in the Derivation of Relativistic Kinematics' (to appear in *British Journal for the Philosophy of Science*, 1992).

D'Agostino, S, (1975), 'Hertz's Researches on Electromagnetic Waves', *Historical Studies in the Physical Sciences*, 6: 261–323.

Einstein, A. (1905*a*), 'Über einen die Erzeugung und Verwandlung des Lichtes betreffenden heuristischen Gesichtspunkt', *Ann. d. Phys.* 17: 132–48. Trans. A. B. Arons and M. B. Peppard, *Am. J. Phys.*, 33: 367–74 (1965).

—— (1905*b*), 'Zur Elektrodynamik bewegter Körper', *Ann. d. Phys.* 17: 891–921. Trans. W. Perrett and G. B. Jeffery, *The Principle of Relativity: A Collection of Original Memoirs on the Special and General Theories of Relativity*, London, 1923; Dover, 1952.

—— (1907), 'Über das Relativitätsprinzip und die aus demselben gezogenen Folgerungen', *Jahrb. Rad. Elektr.* 4: 411–62; 5: 98–9.

—— (1909), 'Über die Entwicklung unserer Anschauungen über das Wesen und die Konstitution der Strahlung', *Phys. ZS.* 10: 817–25.

—— (1912), 'Lichtgeschwindikeit und Statik des Gravitationsfeldes', *Ann. d. Phys.* 38: 355–69.

Einstein, A. (1916), 'Hamiltonsches Prinzip und allgemeine Relativitätstheorie', *Sitzungsberichte, Preussische Akademie der Wiss.* 1111–16.

—— (1918), 'Der Energiesatz in der allgemeinen Relativitätstheorie', *Sitzungsberichte, Preussische Akademie der Wiss.* 448–59.

—— (1920), *Äether und Relativitätstheorie: Rede gehalten am 5 Mai 1920 an der Reichs-Universität zu Leiden*, Springer, Berlin. Trans. G. B. Jeffery and W. Perrett, *Sidelights on Relativity*, pp. 3–24, New York, 1922; Dover, 1983.

—— (1924), 'Über den Äether', *Schweizerische naturforschende Gesellschaft, Verhandlungen*, 105: 85–93. Trans. S. Saunders, this volume.

—— (1949), 'Autobiographical notes', in P. A. Schilpp (ed.), *Albert Einstein: Philosopher–Scientist*, vol. 1, Open Court, Ill., pp. 2–94.

FitzGerald, G. (1888), 'Address to the Mathematical and Physical Section of the British Association', *Report of the British Association*, p. 558, London.

Fulling, S. (1973), 'Nonuniqueness of Canonical Field Quantization in Riemannian Space–Time', *Phys. Rev.* D7: 2850–62.

Hertz, H. (1893), *Gesammelte Werke von Heinrich Hertz*, 2: *Untersuchungen über die Ausbreitung der Elektrischen Kraft*, Leipzig, 1894. English translation: *Electric Waves, Being Researches on the Propagation of Electric Action with Finite Velocity Through Space*, London, 1893; Dover, 1962.

Hirosiga, T. (1969), 'Origins of Lorentz's Theory of Electrons and the Concept of the Electromagnetic Field', *Historical Studies in the Physical Sciences*, 1: 151–209.

Hoffmann, B. (1982), 'Some Einstein Anomalies', in G. Holton and Y. Elkana (eds.), *Albert Einstein: Historical and Cultural Perspectives, The Centennial Symposium in Jerusalem*, pp. 91–105, Princeton University Press.

Lorentz, H. A. (1875), 'Over de theorie der Terugkaatsing en breking bvan het licht', *Academisch Proefschrift*, Leiden, 1875; *Collected Papers*, vol. 1, pp. 1–192, The Hague, 1935–9.

—— (1891), 'Electriciteit en Ether', *Nederl. Natuuren Geneeskundig Congres*, 4 April 1891; *Collected Papers*, Vol. 9, pp. 89–101, The Hague, 1935–9.

—— (1892), 'La Théorie électromagnétique de Maxwell et son application aux corps mouvants', *Arch. néerl.* 25; *Collected Papers*, vol. 2, pp. 164–343, The Hague, 1935–9.

—— (1886), 'De l'influence du mouvement de la terre sur les phènoménes lumineux', *Arch. néerl.* 21: 103 (1887); *Collected Papers*, vol. 4, pp. 153–214, The Hague, 1935–9.

—— (1895), *Versuch einer Theorie der electrischen und optischen Erscheinungen in bewegten Körpern*, Leiden; *Collected Papers*, vol. 5, pp. 1–138, The Hague, 1935–9.

—— (1913), *Das Relativitätsprinzip*, Leipzig (1920).

—— (1914), *Die Kultur der Gegenwart.*

McCormach, R. (1970), 'H. A. Lorentz and the Electromagnetic View of Nature', *Isis*, 61: 459–97.

Newburgh, R. G. and Costa de Beauregard, O. (1975), 'Experimental Search for Anisotropy in the Speed of Light', *Am. J. of Physics*, 43: 528–30.

Pais, A. (1982), *Subtle is the Lord . . . : The Science and the Life of Albert Einstein*, Clarendon Press, Oxford.

Panofsky, W. K. H. and Phillips, M. (1972), *Classical Electricity and Magnetism*, 2nd edn., Addison-Wesley, Reading, Mass.

Poincaré, H. (1898), 'La Mesure du temps', *Rev. Métaphysique et Morale*, 6: 1–13. Trans. H. Poincaré in *The Value of Science*; trans. G. R. Halsted in *The Foundations of Science*, Science Press, New York, 1913.

—— (1904), 'L'État actuel et l'avenir de la physique mathematique', *Bull. Sci. Math.* 28: 302–24. Trans. in *The Value of Science*, chs. 7–9.

Ritz, W. (1908), 'Recherches critiques sur l'électrodynamique générale', *Ann. Chim. Phys.* 13: 145–275.

Saunders, S. W. (1989), *The Mathematical and Philosophical Foundations of Quantum Field Theory*, Ph.D. thesis, University of London.

Tzanakis, C. and Kyritsis, C. (1984), 'On Special Relativity's Second Postulate', *Annales de la Fondation Louis de Broglie*, 9: 343–52.

Unruh, W. (1974), 'Second Quantization in the Kerr Metric', *Phys. Rev.* D10: 3194–205.

Voigt, W. (1887), 'Über das Doppler'sche Prinzip', *Goett. Nach.*: 41–51.

Whittaker, E. (1910), *A History of the Theories of Aether and Electricity*, vol. 1: *The Classical Theories*, Nelson, London/Harper, New York, 1960.

4

The Negative-Energy Sea

SIMON SAUNDERS

1 Introduction

The Dirac hole theory was developed in response to a growing crisis over the Dirac theory of the electron. It predicts the existence of antiparticles in relativistic quantum theory; the antiparticle came into existence as a 'hole' in a sea of negative-energy particles. In 1972 Heisenberg said 'I think that this discovery of antimatter was perhaps the biggest jump of all the big jumps in physics in our century.' He was speaking of the phenomenology, of pair creation and annihilation processes, the basic mechanisms of relativistic dynamics. But the conceptual basis, the *concept*, of antimatter has a corresponding importance.

If this concept was initially tied to the negative-energy sea, that is not the case any longer. The negative-energy sea remains a widespread heuristic device to introduce antimatter; a review of the hole theory is still given in most elementary textbooks on relativistic quantum theory. But nowadays no one would claim that the negative-energy sea actually exists; it is no longer taken as a literal description of the vacuum. How is it, then, that the hole theory can be dispensed with? What takes its place? We know that in some sense antiparticle states are related to negative-energy states; relativity leads to antimatter because the constraint $E^2-p^2c^2=m^2c^4$ is satisfied by negative energies as well as positive energies. The question is, In what precise sense, if not in the sense of the Dirac hole theory?

One of the most widespread heuristics, due to Feynman and Stueckelberg, identifies antiparticles with negative-energy particles moving backwards in time. Antimatter arises just because it is possible for positive-energy particles to scatter backwards in time with negative energies, emitting energy in the process (pair annihilation). This heuristic finds a natural expression in the path-integral approach to quantum theory. However, this approach marks a decisive break with the canonical theory (in particular, with the exact Hilbert space theory). A good reason to be interested in the canonical formulation of the concept of antimatter is to understand better the relationship between the relativistic and the non-relativistic theory. For these reasons I shall omit discussion of the Feynman–Stueckelberg interpretation.

In what follows I shall first sketch the history of the Dirac theory; afterwards I shall concentrate on a more recent development in the canonical framework which, I believe, throws new light on both the hole theory and the field theory which replaced it. This development has its origins in the Segal theory of quantization; what is characteristic of this approach is that complex numbers are built into the classical solution manifold in a geometric way, and this manifold is then identified with the 1-particle Hilbert space. In this way the negative-energy sea is encoded into the mathematical description of the antiparticle states. Something like this was already achieved in the mid-1930s, but mediated by the fields; the Segal theory makes explicit a Hilbert space analogue. Because my concern is to explore the interpretation of quantum electrodynamics within the canonical framework, I shall consider only the linear quantum electrodynamics. Dirac himself was led to the hole theory purely on the basis of the linear equation; in the linear case we already see all the important features of relativistic quantum theory.

In elementary quantum mechanics the vacuum is very simple; it is the quantum analogue of the Newtonian vacuum. In the vacuum not only are there no particles, but there is no theory. There is no Hilbert space, there is no time evolution, one cannot write down equations for this vacuum. The vacuum concept (as distinct from the concepts of space and time) can be described only informally. We have the same situation in classical particle mechanics. But in quantum field theory (also in continuum mechanics and classical field theory), the vacuum is modelled in the mathematics. One might say that in these theories 'what exists' becomes a dynamical quantity, for which non-existence takes on the value zero. (As one value among others, the vacuum must be modelled in the mathematics.) The idea of 'vacuum' is relativized to the observable content of the theory, be it states of a medium, excitations of a field, or particle number. In quantum electrodynamics, despite the field aspect, the vacuum is defined not as the zero-valued fields (there is no state in which all the fields have eigenvalue zero), still less as a zero-valued wave function (which is not even a state), but rather in terms of the absence of any particles. The canonical vacuum is the state of emptiness.

It might seem that this concept of vacuum is essentially unique, and almost as simple as in the elementary theory. Every particle observable has the value zero with probability one.[1] Nevertheless, there are self-adjoint operators for which this is not the case, for example certain combinations of the quantum fields. From the point of view of these operators, the vacuum is not at all trivial. Properties of the vacuum picked out in this way may still be interpreted in particulate terms (almost entirely, in perturbation theory), and there is a direct connection with the picture of the quantum field as a collection of harmonic oscillators (zero-point energy); but it seems to me that a more immediate problem is to understand why such operators arise in particle mechanics in the first place. In particular, one wants to understand how in the Dirac theory even well defined particle observables are *required* to have vacuum expectation values that are non-zero (and in fact infinite).

There is a more general problem. As I have indicated, the Dirac vacuum brings in its wake the concepts of antimatter and pair creation and annihilation processes. These transform the quantum theory into an edifice of remarkable phenomenological expressiveness and real mathematical complexity. The mathematical framework of non-relativistic quantum field theory was reasonably well understood by the late 1920s;[2] more than sixty years later, the simplest of (non-trivial) relativistic theories still resists any comparable elucidation. My objective is this: to characterize better those features of relativistic theory that are responsible for this pathology.

[1] More precisely, every self-adjoint operator that can be defined as the canonical second quantization of a particle operator has eigenvalue zero in the vacuum state. This is true of the non-relativistic theory; it is also true of the theory developed in Section 7, without recourse to normal-ordering.

[2] I have in mind the proof of equivalence of the interacting Galilean field theory with a many-particle ensemble due to Jordan and Klein (1927); see also Tomonaga (1962). The detailed analysis of Fock space methods came somewhat later (Fock 1932; Cook 1953); these and later developments in the mathematical theory of quantum fields are irrelevant to the present discussion, because one can always restrict the field theory to a finite-particle subspace of the Fock space. (This is not possible in the relativistic case.) For applications of non-Fock representations, see Saunders (1988).

In the historical review of Sections 2–6 we observe shifts in the theoretical perspective in due chronological sequence; however, the framework is throughout tied to fermionic theories. The characterization proposed in Sections 7 and 8 fits naturally into this framework but in fact applies equally to fermionic and bosonic theories.

The latter results are restricted; they apply only to global kinematic observables. This theory is, however, exact. It must be born in mind that the conventional theory can be made rigorous only in the kinematic limit; the Fourier analysis is available only for the free field, and in its absence one has no precise particle interpretation.

2 The Origins of the Hole Theory

In 1928 Dirac wrote down a wave equation, which is Lorentz-covariant and first-order in the time observative:

$$(i\hbar\gamma^{\mu}\partial_{\mu}-mc)\psi(\mathbf{x},t)=0.$$

For an external c-number field with potential $A_{\mu}(\mathbf{x},t)$, one then has, for a particle of charge $-e$,

$$\left[\gamma^{\mu}\left(i\hbar\partial_{\mu}-\frac{e}{c}A_{\mu}(\mathbf{x},t)\right)-mc\right]\psi(\mathbf{x},t)=0.$$

In these equations $\gamma^0, \gamma^1, \gamma^2, \gamma^3$ are 4×4 complex matrices, and ψ is a 4-component complex-valued function on space–time. The γ matrices provide a representation of the Clifford algebra which is unique up to isomorphism. They satisfy $\gamma^{\mu}\gamma^{\nu}+\gamma^{\nu}\gamma^{\mu}=[\gamma^{\mu},\gamma^{\nu}]_{+}=2g^{\mu\nu}$. ψ is usually called a Dirac spinor, or bispinor. Dirac was led to this equation because he was looking for a first-order analogue of the Klein–Gordon equation, namely

$$(\Box+m^2c^2/\hbar^2)\phi(\mathbf{x},t)=0$$

where ϕ is a complex scalar function; for an external electromagnetic potential A_{μ}, this becomes

$$\left[\left(i\hbar\partial_{\mu}-\frac{e}{c}A_{\mu}(\mathbf{x},t)\right)\left(i\hbar\partial^{\mu}-\frac{e}{c}A^{\mu}(\mathbf{x},t)\right)-m^2c^2\right]\phi(\mathbf{x},t)=0.$$

The Dirac equation is a linearized square root of this equation; that is, there are linear combinations P, P' of ∂, A, and m such that $P\psi=0$, and such that $P'P\psi=0$ is the Klein–Gordon equation. For this to be possible, P, P' must contain matrices, and correspondingly ψ must be a many-component object. Dirac found that it is not possible to linearize the square-root equation with two-dimensional matrices; the minimum dimension is four, and then one has the Clifford algebra. This follows from the requirement

$$(-i\hbar\gamma^\mu\partial_\mu - mc)(i\hbar\gamma^\mu\partial_\mu - mc) = \hbar^2\Box + m^2c^2.$$

Dirac wanted a first-order equation because he thought that only then could one find a probability interpretation, and define a transformation theory, as in the non-relativistic quantum mechanics. In retrospect, it is clear that he sought a Schrödinger equation,[3] which must indeed be first-order in time; but Dirac also demanded covariance, which is to ask too much. The result is a wave equation which, used as a Schrödinger equation, leads to a theory that is much more than a sum of its parts.

Initially Dirac had only a fragmented formalism; defining the free Hamiltonian by analogy to Schrödinger theory, he obtained the operator ($i=1, 2, 3$):

$$H_D = -i\hbar\gamma^0\gamma^i\partial_i + \gamma^0mc^2.$$

Likewise, he considered the quantity

$$\int \sum \overline{\psi(\mathbf{x})}\psi(\mathbf{x})\,\mathrm{d}^3x,$$

the probability density. (The summation is over the spinor components.) Unfortunately, although this density is positive definite, the spectrum of the free Hamiltonian is not; formally, there exist functions $w\exp[-i(Et-\mathbf{p}\cdot\mathbf{x})/\hbar)$, $w\in\mathbb{C}^4$, which satisfy the wave equation for both signs of E.

Initially this formalism yielded some striking results: it predicted the correct g-factor for the electron and the Sommerfeld equation for the spectral lines of the hydrogen atom. For these reasons, the negative-energy difficulty did not lead to the abandoning of the theory. It was acknowledged from the beginning; Dirac (1928) suggested that the negative-energy solutions might correspond to positive-charge particles, and that they could be rejected on this basis.

[3] I use the term to mean the infinitesimal form of the unitary time evolution with positive generator on the Hilbert space of states.

A few months later he conceded that they could not simply be excluded from the theory, because in the presence of interactions there might be transitions from positive- to negative-energy states and these could not be eliminated by fiat. The theory was then to be thought of as an approximation. But increasingly, it became clear that the difficulty could not be contained or restricted to any non-trivial dynamical regime. Heisenberg, one of the first to perceive the extent of the departure from the principles of quantum mechanics, went so far as to remark: 'the saddest chapter of modern physics is and remains the Dirac theory'.[4] Much later he was to say: 'up till that time 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 open sea again' (Heisenberg 1963). At this stage one couldn't modify the mathematics too much because there seemed to be too much truth contained in the theory.

There were two further developments that made the difficulty of negative-energy states that much more acute. One was the demonstration, due to Oskar Klein (1929), that even for time-independent potentials there may be no solution of the Dirac equation with only positive-frequency parts (the 'Klein paradox'). The other was the discovery, made independently by Igor Tamm (1930) and Ivor Waller (1930), that the negative-frequency parts played an essential role in the classical limit of the Klein–Nisjima scattering formula; that is, in order to get the Thomson formula, it was necessary in perturbation theory to sum over intermediate negative-frequency states.

These states had to be taken seriously. Dirac saw from the beginning that they had to correspond to particles of opposite charge. In the Klein–Gordon case, which also admits negative-energy solutions, there had been suggestions to the same effect.[5]

But a consistent interpretation proved elusive; classically, a negative-energy negative-charge particle behaves in an external electric field just as a positive-energy positive-charge particle; if the electromagnetic potentials are reversed, it actually behaves in an *identical* way to its positive-energy partner. Its space–time world–line would be identical. But surely, this is a peculiarity of conservative systems; intuitively it seems clear that a negative-energy particle will have to emit energy as it speeds up, and that is unphysical. Dirac

[4] Letter to Pauli, 31 July 1928 (Heisenberg 1928).

[5] See e.g. Fock's (1926) derivation where he used a proper time parameter, and the Klein (1926) derivation on 5-dimensional space–time.

could not just posit that the negative-frequency states are positive-charge positive-energy states.[6]

If one thinks about negative energy, one has to work in terms of the absence of positive energy; and if one also thinks about positive charge, one might be led to think of it as the absence of negative charge. From that, it is a short step to the idea that a particle of positive charge and positive energy might correspond to the *absence* of a particle of negative charge and negative energy.

This notion, that the absence of a particle is a physical thing, and has a dynamical role in the theory, had already been employed in quantum theory, and has precursors in classical physics.[7] The most familiar example is the Bohr theory, where one has electron transitions to orbits that are not closed, which do not have their full complement of electrons. Dirac cited internal conversion, where an inner electron is ejected by absorption of X-rays, and remarked that this absence of an electron is described by a wave function and plays much the same role as a physical particle.

The difference is this: there is nothing analogous to the almost-filled Bohr orbitals; there is nothing with respect to which this absence may be defined. It was here that Dirac made a truly revolutionary hypothesis. The negative-energy particles indeed exist, but they exist everywhere and in such abundance that in general transitions to such states will be forbidden by the Pauli exclusion principle. At the same time, if there were an available negative-energy state, it would appear as an absence of negative energy and negative charge, and hence (relative to the background) as a particle of positive energy and positive charge. If there are very few missing negative-energy electrons, and if these transactions are the only empirical manifestation of the existence of the negative-energy sea, then we would scarcely be aware of its existence.

[6] This reasoning is contained in Dirac (1929). Nevertheless, the identification is made out in Section 7; the problem posed by Dirac is eliminated by use of Segal's methods.

[7] For example, discussing an analysis due to J. J. Thomson of the magnetic field associated with a charged moving conductor, G. F. FitzGerald (1881) showed that the displacement currents set up in the ether by the time-varying electric field could not be circuital. As an example he considered a charged parallel-plate capacitor; if one plate approaches the other, the electric field is 'annihilated' by the plate, the electric displacement is therefore destroyed, and there must exist a corresponding displacement current. (This current evidently has non-zero divergence; FitzGerald then showed that the total current, including that arising from the motion of the charged plate, is circuital.)

The assumption of the negative-energy sea is extravagant, even by the standards of the physics of our day. One assumes that each finite volume of space has infinite charge and infinite energy, to make conceptual sense of the theory. In an illuminating remark, Wightman was later to comment:

> it is difficult for one who, like me, learned quantum electrodynamics in the mid 1940s to assess fairly the impact of Dirac's proposal. I have the impression that many in the profession were thunderstruck at the audacity of his ideas. This impression was received partly from listening to the old-timers talking about quantum-electrodynamics a decade-and-a-half after the creation of hole theory; they still seemed shell-shocked. (Wightman 1972: 99)

Familiarity breeds tolerance; one suspects that for later generations it is not so much that the negative-energy sea is considered a fiction, but that no categorical basis seems to exist by which mathematical artifice may be distinguished from the reality.

To understand the significance of the Dirac vacuum, one has to explore the mathematical background of quantum theory at a deeper level. To understand the immediate context in which Dirac worked, one has to understand the second quantization process, and his theory of the equivalence of the quantized electromagnetic field with a many-boson system. The second quantization will play an important role in all that follows, so I shall start with this theory.

3 Canonical Second Quantization

Starting from a canonical 1-particle theory, with a Hilbert space $\mathfrak{h}$, one defines creation and annihilation operators as maps between n- and $(n+1)$-particle spaces, which are constructed as symmetrized or anti-symmetrized tensor products of the 1-particle Hilbert space. The total Hilbert space must contain all these finite particle subspaces, so it is of the form

$$\mathscr{H}=\mathfrak{F}_S(\mathfrak{h})=\sum_n S_n \bigotimes_{i=1}^{n} \mathfrak{h}_i.$$

Here S indicates the appropriate symmetrization and S_n is a representation of the symmetrization operator for the permutation group of order n. (Later on we suppress the subscript S.) Each $\mathfrak{h}_i$ is a copy of $\mathfrak{h}$.

To make the connection with field theory, it is essential that the particles are identical; that is, the states that are built up by successive applications of the creation operator cannot contain information as to which particle is in which state. This being so, the set of occupation numbers, a string of integers $n_1\ n_2 \ldots n_i \ldots$, is enough to parameterize these states, where each subscript i determines a particular state ϕ_i of the 1-particle theory and each occupation number n_i fixes the number of particles in that state. (We suppose the states ϕ_i form an orthonormal basis for $\mathfrak{h}$.) In terms of this parameterization, the action of the annihilation and creation operators is just

$$a(\phi_i):|n_1 \ldots n_i \ldots\rangle \rightarrow (n)^{1/2}|n_1 \ldots (n_i-1)\ldots\rangle$$

$$a^*(\phi_j):|n_1 \ldots n_j \ldots\rangle \rightarrow (n+1)^{1/2}|n_1 \ldots (n_j+1)\ldots\rangle.$$

(The normalization constants are slightly different in the antisymmetric case.) These operators are adjoints of each other, as our notation suggests; as a result, if one is a linear map on the 1-particle space $\mathfrak{h}$, the other must be *antilinear*; that is, for any complex scalar λ,

$$a^*(\lambda\phi)=\lambda a^*(\phi)$$

$$a(\lambda\phi)=\bar{\lambda}a(\phi) \qquad (1)$$

The antilinearity of the annihilation operator is so important that it is helpful to see why it holds in an intuitive way. For a state $\eta\in\mathfrak{F}_S(\mathfrak{h})$ of the form $f_1\otimes f_2\otimes f_3\otimes\cdots\otimes f_n\oplus$ permutations, and arbitrary $f\in\mathfrak{h}$, we have

$$a(f)\eta=(n)^{1/2}\langle f,f_1\rangle f_2\otimes f_3\otimes\cdots\otimes f_n\oplus \text{ permutations.} \qquad (2)$$

The antilinearity of the annihilation operator is therefore a consequence of the antilinearity of $\langle .\,,.\rangle$ in its first entry, the hermitean inner product on $\mathfrak{h}$.

Using these operators, one can write down the operator on an n-particle Hilbert space which corresponds to a 1-particle operator A on $\mathfrak{h}$, such that this operator makes no reference to particle identity, namely

$$\overset{\longleftarrow\!\!-\!\!-\!\!-\; n\text{-fold tensor sum} \;-\!\!-\!\!-\!\!\longrightarrow}{\underset{\longleftarrow n\text{-fold tensor product}\longrightarrow}{A\otimes\mathbb{1}\otimes\cdots\otimes\mathbb{1}}\oplus\mathbb{1}\otimes A\otimes\cdots\otimes\mathbb{1}\oplus\cdots\oplus\mathbb{1}\otimes\cdots\otimes\mathbb{1}\otimes A}. \qquad (3)$$

For an arbitrary orthonormal basis $\{\phi_i\}$, an equivalent definition is

$$\sum_{ij} a^*(\phi_j)\langle\phi_j, A\phi_i\rangle a(\phi_i).$$

The important point about this operator is that it duplicates the action of (3) on *any* n-particle subspace (i.e. whatever the value of n); therefore this expression, but not (3), can be used in a theory in which the particle number is specified along with the state, that is as part of the specification of the initial conditions. This is a first step towards generalizing the theory to deal with dynamical situations in which the particle number is variable.

The operator $\Sigma_{ij}a^*(\phi_j)\,(\phi_j, A\phi_i)a(\phi_i)$ is called the *second quantization* of the 1-particle operator A; it is usually denoted $d\Gamma(A)$. $d\Gamma$ is a structure-preserving map independent of the orthonormal basis used in its definition. In particular, self-adjointness and positivity are preserved by $d\Gamma$. If A generates the unitary group U, then we *define* the group generated by $d\Gamma(A)$ as the second quantization of U, which we denote $\Gamma(U)$. It follows that

$$\Gamma(U)\,d\Gamma(A)\Gamma(U)^{-1} = d\Gamma(UAU^{-1}), \tag{4}$$

which provides an important class of unitary evolutions on $\mathfrak{F}_S(\mathfrak{h})$ (i.e. those determined by unitary 1-particle evolutions on $\mathfrak{h}$).

The creation and annihilation operators can also be used to construct the total number operator; the quantity $\Sigma_i a^*(\phi_i)a(\phi_i)$ applied to an n-particle state returns that state unaltered, except that it is multiplied by the constant n; likewise, the operator $a^*(\phi_i)a(\phi_i)$ is the number operator for the state ϕ_i. Note that the total number operator is the second quantization of the identity; i.e.,

$$d\Gamma(I) = \sum_{ij} a^*(\phi_j)\,\langle\phi_j, I\phi_i > a(\phi_i). \tag{5}$$

The number operator for the state ϕ_i is the second quantization of the projection operator on to the subspace spanned by the state ϕ_i.

The transformation theory can be applied to these quantities in a rigorous way; formally, one often uses the improper position and momentum eigenfunctions also.

$d\Gamma(A)$ has a simple interpretation. Applied to any many-particle state, it gives the appropriate action of the 1-particle operator on each particle in an ensemble. On states of the form η, the c-number $\langle\phi_i, A\phi_j\rangle$ under the summation is multiplied into each c-number $\langle\phi_j, f_1\rangle$ left as residue of the annihilation of the 1-particle state f_1 (cf. (2)), and the state ϕ_i is returned to the state vector η in its place; since we sum over all i, j, we evaluate the total transition amplitude for each particle under the influence of A. This construction also works for 2-, 3-, or n-particle operators; for example, a second quantized 2-particle operator $\mathrm{d}\Gamma(B)$ is written

$$\sum_{ijkl} a^*(\phi_i)a^*(\phi_j) \ll \phi_i\phi_j, B\phi_k\phi_l \gg a(\phi_k)a(\phi_l)$$

(where I have written $\phi_i\phi_j$ for the symmetrized 2-particle state and $\ll .\,, . \gg$ for the induced inner product on the 2-particle subspace). Figure 1 provides a graphic illustration of the action of this operator.

I shall call the second quantization described above the *canonical* second quantization, to avoid confusion with other formalisms which go by the name of 'second quantized' theories. I have emphasized the intuitive aspects of the canonical second quantization, because if one considers the dynamics of a linear quantum field in an external potential according to its Fock space action, one might expect that the particle aspect of the field should reveal this very simple character, and one should have exactly the same dynamics as for a particle ensemble, each particle of which is subject to this external potential. That is what happens in the non-relativistic case; there, the quantum field theory is very simple and mathematically soluble, subject to the usual limitations of mechanics (for the three-or-more-body problem). There is no difficulty in principle in non-relativistic quantum field theory. In this respect the bilinearity of $\mathrm{d}\Gamma$ in creation and annihilation operators is absolutely crucial; if the evolution is defined in this way (i.e. as the second quantization of an n-particle evolution), the particle number is automatically invariant. This feature is fundamental to quantum mechanics; the dynamics is implemented in terms of the transition of a particle from one state into another (the annihilation of a particle in one state and its creation in another). This bilinearity is not a technicality; it reflects the logical structure of the particle theory.[8]

However, this is not at all the situation in the relativistic case, even for a linear field coupled to an external potential. The problem arises entirely from the existence of negative-energy states; the simplest interactions connect positive- and negative-energy states. Nowadays

[8] There is also a connection with group theory, which leads to the Bargmann mass superselection rule. Essentially, the mass arises not as a Casimir invariant, but in the choice of the central extension of the Galilean group; see e.g. Sudarshan and Mukunda (1974).

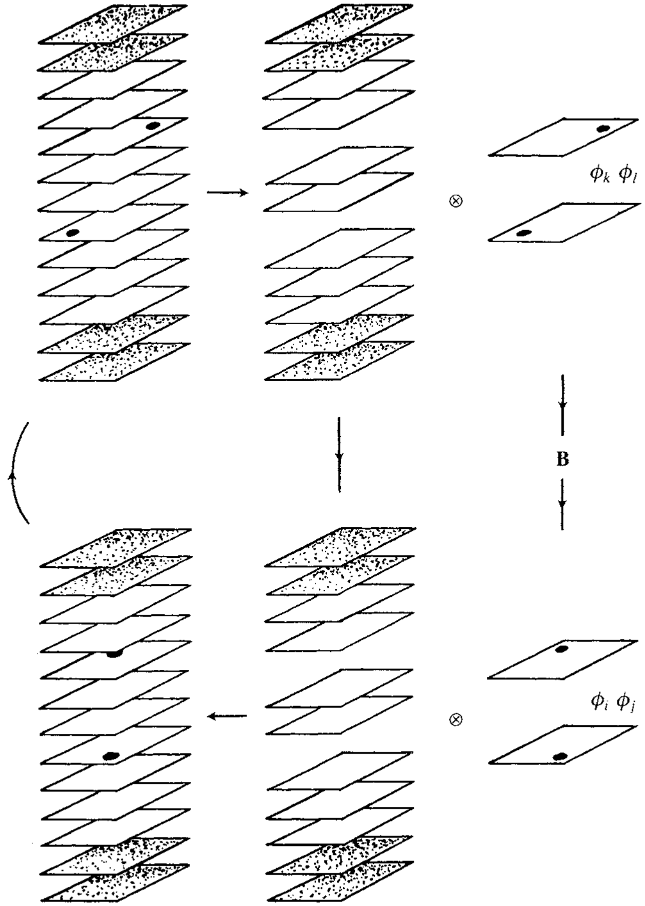


FIG. 1 The action of the operator.

$$\sum_{ijkl} a^*(\phi_i)a^*(\phi_j) \ll \phi_i\phi_{\mathrm{j}}, \mathrm{B}\phi_k\phi_1 \gg a(\phi_k)a(\phi_l)$$

it is usually said that there is no 1-particle relativistic quantum mechanics, so of course the field theory is not the canonical second quantization of any 1-particle theory. Rather, using perturbation theory to read back from the field theory, one learns that what happens on the particle level is pair creation and annihilation, which cannot be understood in terms of particle dynamics (excluding Feynman–Stueckelberg methods).

This point of view arose in the mid-1930s, but Dirac had already considered a similar problem in 1927. He attempted to extend the canonical particle framework so as to describe *individual* absorption and emission processes. This is worth a closer look.

4 The First Dirac Vacuum

One might say that Dirac sought to provide a precise particle interpretation of the emission and absorption of light quanta by atoms.[9] This was Einstein's great 'heuristic hypothesis', which after more than twenty years was at last to find a proper mathematical expression. Dirac had formulated a quantum electrodynamics which led to such processes by applying his theory of action-angle variables, developed in his relativistic q-number (matrix) mechanics of the Compton effect in 1926, transferred to the radiation field in place of the mechanical atom.[10] At the same time, he had discovered that similar techniques applied to the non-relativistic (linear) Schrödinger equation yielded a formalism equivalent to the quantum mechanics of a many-particle boson system (the canonical second quantization which I have just described). Dirac tried to bring the two into correspondence: obviously, the difficulty is the mechanical description of the creation and annihilation of photons.

The problem is simple. Because Dirac wanted to describe the back-reaction on the field, the potential V is no longer an external perturbation, and must be considered a dynamical entity in its own

[9] There are many routes to Dirac's (1927) development of quantum field theory (wave-particle duality, quantum electrodynamics, the correspondence principle, the Einstein theory of A and B coefficients, the Kramers–Heisenberg dispersion theory, the Compton effect); the details need not concern us here.

[10] In his theory of the Compton effect he had used the action-angle operators to describe transitions of the atom, not of the field. The action-angle operators provided the algebraic structure of the creation and annihilation operators; see Dirac (1926*a*, 1926*b*).

right. It becomes, in fact, a quantum field. However, it immediately appears—in support of Einstein's original conceptions—that the Hamiltonian derived from this potential contains an isolated creation (or annihilation) operator.[11] How to relate this theory to the canonical theory? There one can only obtain operators of the form $d\Gamma(A)$; necessarily, these are bilinear in creation and annihilation operators, which is only to say that net particle creation and annihilation processes cannot be described in the mechanical theory.

We have another way of understanding this difficulty if we look at Dirac's proof of the equivalence between a q-number linear Schrödinger equation and a boson ensemble. It is implicit here that the c-number Schrödinger equation for each boson is formally identical to the q-number equation. When one tries to do the same thing for the radiation field coupled to charge, the interaction Lagrangian contains a term linear in the vector potential; hence the field equations are inhomogeneous and no longer linear. Consider for example the electrostatic case; there is an interaction density of the form $V(x, t)\rho(x, t)$, with V the electrostatic potential and ρ the charge density. (In the non-relativistic theory, the latter is bilinear in creation and annihilation operators: no net particle creation or annihilation here.) If one looks at the field equations, the q-number version of what should be the 1-particle Schrödinger equation is obviously nothing of the kind. Because the Lagrangian contains the term ρV, the field equations contain the inhomogeneous term ρ, one has Poisson's equation $\nabla V = -e\rho$, which cannot be understood as a q-number version of a Schrödinger equation. The equivalence between Schrödinger and field equations exploited by Dirac has disappeared.

We are faced with a difficulty which is perhaps more acute than that of the negative-energy states. (Dirac began with the hard problems.) In order to get out of this difficulty and preserve a relationship between the canonical second quantized theory and the quantized field theory, Dirac considered that photon number is also conserved; the apparent annihilation of a photon is in reality the process in which a photon of frequency ν makes a transition to a photon of frequency zero. That is, there are many (infinitely many) zero-frequency photons present all the time, and individual creation and annihilation

[11] According to Dirac, the reality of the potential required its expansion as the sum of creation and annihilation operators for the (positive-frequency) light-quanta. Dirac used a non-relativistic description (for photons!). The negative-frequency difficulty does not arise.

processes become photon transitions to and from this new vacuum. In this way Dirac described the quantized electromagnetic field as a particle theory; this was his resolution of the wave-particle duality as formulated by Einstein.[12] There is no precise mathematical basis to this theory, but we have the first indication that the dynamics can be changed dramatically by modification of the vacuum. This was already clear in 1927.

5 The Negative-Energy Sea

Whereas one might accept that a vacuum filled with zero-frequency photons is still a nothingness, because a zero-frequency photon is just like a vacuum state of a classical field, the situation is different when the vacuum is full of massive charged electrons. What of the gravitational properties of this vacuum? How does this vacuum respond to electromagnetic fields? Even if there are no holes present, there must be other empirical consequences of this idea. What can it mean to have infinite charge and mass in finite volume?

It was a bold and radical and quite outrageous suggestion. Pauli, even after the discovery of the positron, was absolutely against it. As he later put it, 'success seems to have been on the side of Dirac rather than of logic'. We have seen Heisenberg's reaction;[13] Bohr also was sceptical. It is well known that Anderson, the discoverer of the positron, was indifferent to the Dirac theory (Anderson and Anderson 1983). The theory was exotic and speculative; prior to the discovery of the positron, only a handful of experts was concerned with it.

But the hole theory was successful; its heuristic power was

[12] Dirac observed that the 1-particle 'analogue' of photon creation and annihilation (transition to and from the zero-frequency state) could not be written 'as an algebraic function of canonical variables'. In this sense Dirac did not achieve a complete mechanical description of the interaction (contrast with the hole theory). On this point, von Neumann was later to take an anti-realist stand: 'It is difficult to find a direct, clear-cut, interpretation of the interaction energy . . . nevertheless, we can accept this with the interpretation that each model-description is only an approximation, while the exact content of the theory is furnished solely by the expression for the Hamiltonian operator' (von Neumann 1932: 282–3; the 'model-description' is the particle theory, the Hamiltonian is derived from the field).

[13] In early 1934 Heisenberg wrote: 'I regard the Dirac theory . . . as learned trash which no one can take seriously' (1934*a*). Three months later he was able to eliminate it from the formalism (see below).

immense. If there is a hole present, then a positive-energy electron may make a transition to this state with the emission of energy. This would appear as the annihilation of both the electron and the hole. If a negative-energy electron absorbed energy, so that its total energy became positive, it would leave behind it a hole in the negative-energy sea, and this would appear as the creation of a positive-energy electron together with a hole. The inferences follow effortlessly. Phenomena of this kind were soon observed.

The sequence of events went something like this. To begin with, physicists were not then disposed to predict the existence of new kinds of particles. There was only one sub-atomic positively charged particle known, and that was the proton. Weyl had already suggested that the negative-frequency states describe the proton; Dirac rejected their identification with protons, for reasons already summarized, but he suggested that these might appear as the holes in the negative-energy sea. The mass difference, he surmised, may be accounted for by the interaction among all the negative-energy electrons. This was in November 1929; by March of the following year, he (and, independently, Oppenheimer) calculated the cross-section for electron–proton decay. Even given the ambiguity introduced by the electron–proton mass difference, the result was much too large to be consistent with the stability of ordinary matter. The same conclusion was reached by Tamm one month later; by the end of the year Weyl had gone into print retracting his earlier suggestion. In May 1931 Dirac predicted the existence of the positron. It was observed by Anderson that summer.[14]

The way was open to evaluate scattering cross-sections for a number of new phenomena (all to be experimentally observed). Between 1930 and 1935 the following processes were considered (the calculations used c-number potentials for the radiation):

$e^+ + e^- \to \gamma$ (Dirac, Oppenheimer, Tamm)
$\gamma + \gamma \to e^+ + e^-$ (Breit, Wheeler)
$e^+ + e^- \to \gamma \to e^+ + e^-$ (Bhabba)
$\gamma + \gamma \to e^+ + e^- \to \gamma + \gamma$ (Halpern)

The fundamental fact is that, by redefining the ground state of the

[14] This sequence of events is well documented; see e.g. Bromberg (1976) and Pais (1986) for further details and references.

electron theory, the standard ideas of 1-particle quantum theory[15] immediately lead to a phenomenology typical of a many-particle theory. The idea of pair creation and annihilation had actually been around for some time; some progress had been made in studying the equilibrium properties of such processes using semi-classical arguments. But with the hole theory, pair creation and annihilation became 1-particle processes; there is the interpretation, and there is the mathematical formalism. Immediately one could calculate cross-sections and apply the perturbation theory to deduce the existence of more complex processes that proceed by virtual states (such as the Bhabba and Halpern scattering). It is often said that the Dirac hole theory transforms the 1-particle theory into a many-particle theory; we see that it also works the other way: prima facie many-particle processes involving pair creation and annihilation can be treated using the formalism of the 1-particle theory.

In the processes that we have considered the negative-energy sea plays a purely passive role, in restricting the number of negative-energy states available for such transitions. In other processes the sea is more active: its response to an external field should be just like a dielectric. The negative-energy electrons will be polarized and an induced polarization field will be set up. This is the vacuum polarization, first investigated by Dirac in 1934, and here for the first time the full intricacies of the hole theory were encountered. To deal with them Dirac used a variant of the Hartree self-consistent field method. This step leads naturally to the calculation of the effective charge that will produce the 'net' field, that is the external field together with the polarization field of the vacuum; in other words, he was led to the idea of *charge renormalization*. This is the first time that the notion of renormalization entered quantum physics.[16]

[15] Here, by '1-particle system' I mean a particle that may be found in positive- or negative-frequency states. The term 'particle–antiparticle system' is unsatisfactory, because it suggests that one has a particle pair; rather than introduce this or an even more cumbersome terminology, I shall leave it to the context to distinguish the 1-particle system in the present sense from a 1-particle system defined over positive-frequency states alone. No confusion will result.

[16] In other situations, e.g. the interacting non-relativistic field considered by Jordan and Klein (1927: 761–2), the quantum theory effectively *removes* a renormalization problem of the classical theory; the self-energy of the classical field theory disappears through normal-ordering. (This is the only application of normal-ordering to the non-relativistic theory.)

6 The Standard Formalism

With these developments, the Dirac hole theory became part of the basic vocabulary of physics. Every practising high-energy physicist knows the theory well; it has an apparently enduring heuristic role. But the negative-energy sea is no longer considered a literal description of reality. What, then, became of the theory, and how do we do without it today?

There seem to be two answers to this question. The first is that we have learnt to dispense with this heuristic and to rely on the mathematics unaided. The second is that we have a new theory, formally similar, basd on the Feynman–Stueckelberg heuristic and path-integral methods.

This new theory I will have to place on one side. It is the first response that concerns us; it is based on what used to be called the second-quantized hole theory, or the Jordan–Wigner formulation, but now it is called the Dirac field theory. I shall call it simply the *standard formalism*. The fundamental step was to reformulate the hole theory as a second-quantized theory (in a rather imprecise sense). This step was taken by a number of people in 1934—by Fermi, in connection with the theory of β-decay; by Heisenberg, in order to eliminate the negative-energy sea and the asymmetry between positive and negative charge; and by Fury and Oppenheimer, in their systematic reconstruction of the Dirac hole theory. It was reinforced by the new impetus to field theory provided by Pauli and Weiskopf, Yukawa, and the rapid growth of meson physics from the mid-1930s. I shall not discuss these developments; I shall only present the second-quantized version, more or less as did Heisenberg (1934*b*).

To begin with, consider the canonical second-quantization. If for our orthonormal basis we use instead the 'improper' momentum eigenfunctions, we are led to the 'operator' $b_r(p)$ annihilating an electron (of positive or negative energy) of four momentum p. The subscript r picks out one of the two spin eigenstates with respect to a selected component of spin.[17] Taking the Fourier transform,

[17] There is an unfortunate complication here in connecting the standard formalism to the canonical theory. The details will not be relevant to what follows, but in parenthesis let me say this: the plane-wave expansion was first written down for the 1-particle solutions of the Dirac equation. Therefore the bispinor appears explicitly. On 'second-quantizing', the expansion coefficient b_r was made into an annihilation operator. In the canonical framework, it is simpler to work with annihilation operators of the form $a(\phi)$, with ϕ a 1-particle state (and if possible avoid a parametrization of the

extended over both halves of the mass shell,[18] we obtain the point field

$$\psi(x)=\int_{E>0}\sum_r w_r(p)b_r(p)\,\mathrm{e}^{-ip\cdot x/\hbar}\,\mathrm{d}\mu^+$$

$$+\int_{E<0}\sum_r w_r(p)b_r(p)\,\mathrm{e}^{-ip\cdot x/\hbar}\,\mathrm{d}\mu^-.$$

Here the summation is over the two linearly independent spin states, and $\mu^\pm$ is the invariant measure on the mass shells (which includes constants in $\hbar$ and π). It is usual to rewrite this by letting r run over four values, one pair for each sign of the energy: $r=1, 2$ for the positive-mass shell and $r=3, 4$ for the negative-mass shell; i.e., for positive-energy solutions,

$$w_r(\mathbf{p})=w_r((\mathbf{p}^2+m^2c^2)^{1/2}, \mathbf{p}),$$

and for negative-energy solutions,

$$w_{r+2}(\mathbf{p})=w_r[-(\mathbf{p}^2+m^2c^2)^{1/2}, \mathbf{p}]$$

(and similarly for b). One can then carry out the integral over the energy p_0, obtaining the familiar form,

$$\psi(x)=(2\pi\hbar)^{-3/2}\int_{\mathbb{P}^3}\Bigg[\sum_{r=1,2} b_r(\mathbf{p})w_r(\mathbf{p})\,\mathrm{e}^{-ip\cdot x/\hbar}$$

$$+\sum_{r=3,4} b_r(-\mathbf{p})w_r(-\mathbf{p})\,\mathrm{e}^{ip\cdot x/\hbar}\Bigg]\frac{\mathrm{d}^3p}{\sqrt{2p_0}}$$

(here $p_0=+(\mathbf{p}^2+m^2c^2)^{1/2}$; the normalization is $\bar{w}_r w_s=2p_0/c\delta_{rs}$).

We now consider the canonical second quantized operators. By formal application of the transformation theory (using 'improper'

space of states). In that case ϕ includes the bispinor. (This is what we shall do in Section 7.) If one treats instead the quantities $b_r(p)$ as annihilation operators, then they must act on the Fock space over the $s=\frac{1}{2}$ spinor representations constructed by Wigner (1939) (and not over the solution space of the Dirac equation). This feature of the standard formalism is perhaps not a technicality when it comes to perturbation theory; in kinematics, however, it is well understood.

[18] For the time being we can consider the Fourier transform an application of the transformation theory in the canonical theory; viewed in this way, we must expand over a complete set of states, therefore over both positive- and negative-energy states.

position eigenstates[19]), any operator $A(\mathbf{x})$ which is local in the 1-particle theory (a multiplicative function or finite derivative in configuration space coordinates) can be written in the form $\int\Sigma\psi^*(\mathbf{x})A(\mathbf{x})\psi(\mathbf{x})\,\mathrm{d}^3x$. Applied to the 1-particle Hamiltonian H, one obtains, after some manipulation,

$$\mathrm{d}\Gamma(H)=\int\sum_{r=1,2}p_0[N_r^+(\mathbf{p})-N_r^-(-\mathbf{p})]\,\mathrm{d}^3p/p_0,$$

where we define the number operators:

$N_r^+(\mathbf{p})=b_r^*(\mathbf{p})b_r(\mathbf{p})$ (for positive-frequency states)
$N_r^-(\mathbf{p})=b_{r+2}^*(\mathbf{p})b_{r+2}(\mathbf{p})$ (for negative-frequency states).

As in the 1-particle theory, the canonical second quantized energy is indefinite. The total charge is the second quantization of $-e\mathbb{I}$, and one verifies that

$$\mathrm{d}\Gamma(-e\mathbb{I})=-e\int\sum_{r=1,2}[N_r^+(\mathbf{p})+N_r^-(-\mathbf{p})]\,\mathrm{d}^3p/cp_0.$$

This is negative definite, as we expect (since the charge of electrons, whether positive- or negative-frequency, is negative).

So much for the *canonical* second quantization. Despite the formal manipulations, everything can be made rigorous and put into the canonical framework. But if we now consider the action of the negative-energy creation and annihilation operators on the Dirac vacuum, the negative-energy sea, clearly the annihilation operator will create a hole (positron) and the creation operator will annihilate a hole, or will give the value zero if there is no hole present. Accordingly, let us change our notation; we shall replace $b_3(-\mathbf{p})$ by $d_1^*(\mathbf{p})$ and $b_4(-\mathbf{p})$ by $d_2^*(\mathbf{p})$. (The change in sign in $\mathbf{p}$ is for convenience; using the symbol d rather than b eliminates the need for the index values 3 and 4, and the $*$ indicates whether we are dealing with a creation or annihilation operator with respect to the positrons.)

[19] These are doubly improper, since they have little to do with particle position. This part of my treatment is undoubtedly clumsy, and should be replaced by a more fundamental treatment of the Fourier transform. However, I wish to avoid a detour into the representation theory of abelian groups; interested readers are referred to Mackey (1963) for a general perspective.

Similarly, $b^*_{r+2}(-\mathbf{p})$ is replaced by $d_r(\mathbf{p})$; the anticommutation relationships obeyed by these operators are unchanged by these substitutions. (This would not be true if they obeyed commutation relationships).

The effect of these substitutions is that when we evaluate the quantities $d\Gamma(H)$, $d\Gamma(-e\mathbb{1})$, we obtain the quantities as above except that now $N_r^-(-\mathbf{p})=d_r(\mathbf{p})d_r^*(\mathbf{p})$; *the order of the creation and annihilation operators is reversed.* That being so, this term *cannot be interpreted as the positron number operator.*

This can easily be remedied; we use the anticommutation relationships (ACRs) to write these quantities (the original number operators for negative-energy electrons) in terms of number operators for positrons. The latter are given by the quantities $N_r^-(\mathbf{p})=d_r^*(\mathbf{p})d_r(\mathbf{p})$; in this way we obtain $d_r(\mathbf{p})d_r^*(\mathbf{p})=-N_r^-(\mathbf{p})+$positive infinite constant. The expressions for the total energy and charge (E and Q) become:

$$E=\int\sum_{r=1,2} p_0[N_r^+(\mathbf{p})+N_r^-(\mathbf{p})]\,\mathrm{d}^3p/p_0-\text{positive inf. const.}$$

$$Q=-e\int\sum_{r=1,2}[N^+(\mathbf{p})-N^-(\mathbf{p})]\,\mathrm{d}^3p/cp_0-\text{positive inf. const.}$$

The change in sign in these quantities is crucial. The infinite constants correspond, in the hole theory, to the infinite negative energy and negative charge of the negative-energy sea. The remaining contribution to the energy (charge) is now positive definite (indefinite). But as yet the spectrum of these operators is unchanged, nor could it be changed merely by a change in notation and use of the ACRs.

In quantum field theory it is standard practice to subtract such infinite constants (zero-point subtractions) produced by the reordering; the reordering followed by the setting to zero of all c-numbers is called *normal-ordering.* By this 'standard practice', however, *the spectrum of the operator is changed.* The subtraction makes the energy into a positive operator, and the spectrum of the charge operator is no longer negative definite. Mathematically, therefore, the zero-point subtraction is far from trivial. By this same practice, we also find that we have a new basis for the theory: *if one now writes down the momentum expansion for the quantum field $\psi(x)$ with the 'correct' interpretation of the operator components, and normal-orders all expressions for physical observables, one need make no reference* *whatsoever to the Dirac hole theory*. That is, if from the word go we write the quantum field as

$$\psi(x)=\int_{E>0}\sum_{r} w_r(p)b_r(p)\,\mathrm{e}^{-ipx/\hbar}\,\mathrm{d}\mu^+ + \int_{E<0}\sum_{r} w_{r+2}(p)d_r^*(-p)\,\mathrm{e}^{-ipx/\hbar}\,\mathrm{d}\mu^-$$

and normal-order the physical observables, we need never bother with the hole theory. The negative-energy sea has done its work once we have the 'correct' particle interpretation of the field, which is to say the plane-wave expansion above, and once we no longer demand a physical interpretation of the normal-ordering process.[20] In particular, using the Lagrangian theory and Feynman diagrams, we can develop a formal perturbation theory for interactions which lends itself readily enough to intuitive visualization. The relevant heuristic is that particles are created and destroyed in pairs, so as to preserve charge; these are not transition processes of a single particle involving negative- and positive-frequency states. Both intuitively and in the mathematics, we can no longer treat the resulting theory as the canonical second quantization of a 1-particle theory. The technique for making the energy positive (correct momentum space expansion + normal-ordering) does not seem to make any sense at the 1-particle level; a precise correspondence is lost. In this respect we have the same situation as in the Dirac hole theory (with the negative-energy sea as vacuum), but actually the situation is worsened; we have no physical basis for the rift with the canonical theory.

Nowadays no one would regard the use of this Fourier expansion or of the normal-ordering as logically dependent on the Dirac hole theory; they are supposed to stand in their own right. At the same time, the entire theory can be regarded as a quantum field theory, and the link with the 1-particle theory becomes hopelessly tenuous. One looks upon the Dirac equation as a classical field equation, itself derived from a classical Lagrangian. The quantization of this theory is to yield the standard formalism (correctly interpreted), as above.

[20] This is to be considered a purely mathematical technique which does not stand in need of justification. For an account along these lines, see Wightman (1972). In the theory of Section 7, the normal-ordering has a fundamental significance. Path-integral theory also places a more fundamental perspective on normal-ordering. (This is easiest to understand in Euclidean theory; see e.g. Simon 1974, sec. 1.1).

There is, however, a connection between the antimatter fields and the negative-energy solutions. The latter contribute negative energy to the total field energy. (The use of anticommutators on quantization allows us to change the sign of this contribution, depending on whether we consider it a creation field or an annihilation field.) This part of the field ψ is the antiparticle (creation) field. But the negative-energy solutions disappear from the Fock space description (there are no negative-energy states); there is a doubling-up of states, and their distinction is made at the q-number level. What were before calculations of transition amplitudes at the level of the states become analysis at the level of the fields; particularly, it is analysis on the c-number bispinors that occur in the plane-wave expansion.

It does not appear possible to understand antimatter at the level of the states. As a result, certain questions, such as the definition of states when one does not have a scattering situation, or the meaning of the Wigner negative-energy representations of the Lorentz group (which now appear to be excluded by fiat), cannot even be formulated. It is the canonical theory that imparts precision to these questions.

Let me pursue the question of the independence of the standard formalism from the Dirac hole theory. The problem is to justify the plane-wave expansion of the fields. It turns out that the necessary assumptions have a natural interpretation in field theory; the field must be a linear combination of creation and annihilation operators. This is implicit in some earlier discussions, but (so far as I know) there is no very clear statement prior to Weinberg (1964). One reason for this neglect is that the precise definition of the creation and annihilation operators—namely an explicit action on a concrete Fock space—was not and could not have been available prior to the late 1950s because of the difficulty in relating the Dirac bispinors to the Wigner spinor representations. This problem (cf. fn. 17) requires the distinction between representations on Hilbert space and those on Hilbert space bundles;[21] the explicit bispinor c-numbers that occur in the plane-wave expansion should be understood as transformation matrices between the $\mathbb{C}^4$ fibre sitting over the base space and the Wigner $\mathbb{C}^2$-valued Hilbert space of spinors.

But neither was Weinberg concerned with the logical status of the plane-wave expansion. He was trying to demonstrate the independence of the S-matrix theory from Lagrangian methods. To this end

[21] A crucial link that was first investigated in a physical application by Joos (1962).

he was driven to work from first principles. He found[22] that in kinematics, if one is to get covariant operators that commute (or anticommute) at spacelike separation, it is necessary to take linear combinations of creation and annihilation operators. Even then, it is not necessary that one have an annihilation operator for a particle state and a creation operator for an antiparticle state. (They could be creation and annihilation operators for a single species of particle.) However, in that case the field will (in general) be complex, but it will not transform in any simple way under (global) gauge transformations; it will not transform as $\psi \rightarrow \exp(i\theta)\psi$, with θ a *c*-number. The reason is that, if the field ψ is a linear combination of creation and annihilation fields on the *same* Fock space, then if the annihilation field (say) transforms as $b \rightarrow \exp(i\theta)b$, the creation field must transform as $b^* \rightarrow \exp(-i\theta)b^*$ because it is the adjoint field. So the final upshot is that, if we want to have a covariant, causal, field that is complex, but transforms simply under gauge transformations, then we must introduce a new Fock space (the antiparticle space), and the linear combination of annihilation operators on the particle space and creation operators on the antiparticle space (together with its adjoint) is the only possible operator expansion.

Weinberg was happy to have isolated simple assumptions concerning the field, sufficient for the derivation of the plane-wave expansions; with these, the Feynman rules could be defined, and on the *S*-matrix philosophy no further appeal to a dynamical theory (such as the Lagrangian theory) was necessary. For our purposes, what is important is that these are just the properties of the field as required in Lagrangian theory; the field is covariant, causal, and gauge-covariant.

In this way the standard formalism can be considered logically independent of the Dirac hole theory. However, one must still motivate the normal-ordering process, and one finds that the antiparticle states have no relationship to the negative-energy 1-particle states. On the contrary, they are identical to the positive-energy states. This needs careful consideration.

The antiparticle states (the elements of the antiparticle Fock space) behave identically under Lorentz transformations, including space and time inversions, as the particle states. (In particular, they have

[22] The summary that follows is a slight modification of Weinberg (1964), along the lines of Novozhilov (1975).

positive energy.) One merely supposes that these states are *distinct*, so that $b_r(\mathbf{p}) \neq d_r(\mathbf{p})$. What makes them distinct? First and last, it is the action of the fields. It is the way the fields couple the two kinds of states that leads to the characteristic dynamics whereby pairs of particles (one in each class of states) are destroyed and created. What prevents single creation and annihilation processes, or pair processes of other kinds (each from the same class of states), is the requirement that the Hamiltonian be gauge-invariant.

This is something new to the principles of elementary quantum mechanics, although its impact—in particular, its implications for measurement theory and the transformation theory—is somewhat reduced under the rubric of charge superselection. It is often said that operators that connect states of different total charge do not exist. But the meaning of this statement is unclear. There is nothing comparable in the non-relativistic theory. The mass superselection rule has a different origin; the gauge invariance of the Hamiltonian is there a consequence of the fact that it is self-adjoint.

I have omitted mention of charge conjugation. One might think that the explicit definition of this operator will clarify the relationships between matter and antimatter on the one hand, and positive- and negative-frequency states on the other. It is even said that the introduction of charge conjugation restored the symmetry between positive and negative charge, and cleared the way to the elimination of the negative-energy sea. In fact, the charge conjugation adds little to our understanding.

In the Weinberg construction, this operator (denote $\mathfrak{C}$) is defined by the interchange of the b operators with the d operators, or simply by the interchange of the two Fock spaces (for particle and antiparticle). Since these are identical as function spaces, it is trivial that $\mathfrak{C}$ is unitary.

However, at the level of the fields it can be written in a way that is formally identical to the charge operator $\mathscr{C}$ in the 1-particle theory, which is antiunitary. The details are as follows. If one takes the adjoint (complex conjugation plus matrix transposition) of the Dirac equation in the presence of an external field,

$$\left[\gamma^{\mu}\left(i\hbar\partial_{\mu}-\frac{e}{c}A_{\mu}(x)\right)-mc\right]\psi(x)=0,$$

one obtains the equation (superscript t is matrix transpose)

$$\bar{\psi}^t(x)\left[\bar{\gamma}^{\mu t}\left(-i\hbar\partial_\mu - \frac{e}{c}A_\mu(x,t)\right) - mc\right] = 0.$$

From the defining properties of the γ matrices, it follows that $\gamma^0\bar{\gamma}^{it}\gamma^0 = \gamma^i$ $(i=1, 2, 3)$; so inserting a factor $\gamma^0\gamma^0$ between $\bar{\psi}^t$ and $\gamma^{\mu t}$, and operating from the right by γ^0, one obtains, on taking the matrix transpose,

$$\left[\gamma^{\mu t}\left(-i\hbar\partial_\mu - \frac{e}{c}A_\mu(x,t)\right) - mc\right]\gamma^{0t}\bar{\psi}(x) = 0.$$

This equation is not quite in the right form, because of the transposition of the γ matrices. However, if there exists a matrix C such that $C\gamma^{\mu t}C^{-1} = -\gamma^\mu$, we may insert a factor $C^{-1}C$ and operate from the left by C to obtain

$$\left[\gamma^\mu\left(i\hbar\partial_\mu + \frac{e}{c}A_\mu(x,t)\right) - mc\right]\psi^c(x,t) = 0$$

where we have written $\psi^c = C\gamma^{0t}\bar{\psi}$ (the positron state). This is the Dirac equation for a particle of *positive* charge.

Actually, if ψ is a positive-frequency state, then ψ^c is a negative-frequency state, so that to obtain positive-frequency solutions of the positive-charge Dirac equation we must take the charge conjugate of negative-frequency negative-charge solutions. The map $\mathscr{C}$: $\psi \rightarrow \mathbf{C}\gamma^{0t}\bar{\psi} = \psi^c$ is called the 1-particle charge conjugation operator; a matrix $\mathbf{C}$ with the defining property above exists and can be chosen unitary; because of the complex conjugation, however, $\mathscr{C}$ is antilinear (hence antiunitary).

However, it seems that one cannot consistently define charge conjugation within the 1-particle theory. The reason is that the total charge (and likewise the charge current density) do not change sign under the 1-particle charge conjugation. Since the charge operator is just $-e\mathbb{I}$, it is obvious that this operator is invariant under charge conjugation, contrary to physical requirements.

The relationship with the charge conjugation $\mathfrak{C}$ for the quantum field is as follows. The formal application of the operator $\mathscr{C}$ to the field takes one from the field to its adjoint, hence is equivalent to the interchange of b and d, even though it defines transformations of the form $b \rightarrow b^*$, $d \rightarrow d^*$. Since these transformations are antilinear, it seems we must have an antiunitary transformation (and an anti-automorphism with respect to the fields); but if $\mathscr{C}$ is automorphic and one normal-orders after its application, then one obtains the same transformation as $\mathfrak{C}$. The total charge now changes sign under charge conjugation, because (normal-ordered) it is no longer a multiple of the identity.

This is a curious situation; the charge conjugation appears at once antilinear at the level of the fields and unitary at the level of the Fock space. On the other hand, if the gauge transformation of the fields is induced by that of the states, the particle and antiparticle states must stand in antilinear correspondence,[23] in contradiction to the unitarity of $\mathfrak{C}$.

The standard formalism presents puzzling features, and the relationship between the 1-particle and field-charge conjugation operators is one more example. The particle interpretation of the fields, from which follows all of the mathematical pathology of relativistic quantum theory, is secured if the fields are covariant, gauge-covariant, and cusal, but these requirements make no sense at the 1-particle level. The distinction between matter and antimatter cannot be made out at the level of the states, and the negative-energy representations of the Lorentz group play no role in the theory. We are a long way from the clear-cut heuristics of the Dirac hole theory.[24]

Confronted with this situation, one feels a certain exasperation. Surely the antiparticle *field* is the negative-frequency *field* solution of the *field* equation, just as the negative-energy *state* is the solution of the *wave* equation. The connection between antimatter and negative energy should be direct and simple. In order to make it so, we must find a formulation of the canonical theory which directly relates q-number and c-number versions of the same equations. Fortunately, that has been worked out for us.

7 Quantization and Complex Numbers

I refer to the so-called *geometric quantization*, due to several workers, but above all to Irving Segal.[25] He was concerned specifically with the

[23] i.e., if the transformation $a(f) \rightarrow \exp(i\theta)a(f)$ arises from the transformation $f \rightarrow \exp(i\theta)f, f \in h$, and f and g are respectively particle and antiparticle states, then we must have $g \rightarrow \exp(-i\theta)g$.

[24] These obscurities are eliminated in the theory that follows; I omit, however, a discussion of covariance and microcausality, which hinge on the analysis of locality (see Saunders 1989, sec. 3.4).

[25] See e.g. Segal (1964, 1967). For a review of the more general theory, see e.g. Woodhouse (1980).

quantization of linear classical fields; in its more developed form (following Souriau 1966, and Kostant 1970) it provides a quantization process—and with it a representation theory—which generalizes the Dirac correspondence between commutation relationships and the Poisson bracket. This theory leads to a rigorous quantum theory of much more general systems, e.g. constrained systems on manifolds; however, we will make use only of the basic construction provided by Segal.

This construction is applicable in all cases where a rigorous Fock space representation has been established in quantum field theory. It can be considered a generalization of the canonical theory; its novel features disappear in the non-relativistic limit.

Suppose that we have a linear dynamical system and an associated phase space, that is a pair V, ω where V is a real vector space[26] and ω a bilinear form, antisymmetric for bosons (the symplectic form) and symmetric for fermions; and let us introduce a canonical transformation J such that $J^2 = -1$. With the aid of this, we can construct a *complex* vector space and a sesquilinear inner product, and can complete the vector space to obtain a Hilbert space[27] (denote V_J). The point of this construction is that *symplectic (bosons) or orthogonal (fermions) transformations on* V *which preserve* J *automatically become unitary transformations on* $\mathrm{V_J}$. In particular, the Hamiltonian flow, the group of transformations on V corresponding to the (classical) time evolution, becomes a weakly continuous group of unitary transformations on V_J so long as it preserves J. The quantum mechanical Hilbert space $\mathscr{H}$ is then given as the space of *analytic*[28] functions on this space. In quantum field theory, V is already a function space, the space of classical solutions to the classical field equations,[29] and $\mathscr{H}$ is naturally represented as the Fock space over V_J. The complexified

[26] This does not mean that V might not also be a complex vector space (i.e. that it is also complex-linear); the point is that we use only the real-linear structure.

[27] This procedure was foreshadowed in the work of Stueckelberg and his co-workers in the early 1960s; see e.g. Stueckelberg and Guenin (1962) and references therein.

[28] Unitarity and analyticity are both defined with respect to J. This Hilbert space of analytic functions was first introduced by Fock (1928); see Bargmann (1961) for a systematic treatment. An analytic function has a power series expansion; when it is defined on a function space (as in classical field theory), the term in this expansion linear in vectors in V is the 1-particle component of this state.

[29] More precisely, the space of Cauchy data for the field. For the Dirac field, this can be identified with square-integrable $\mathbb{C}^4$-valued functions on $\mathbb{R}^3$. The theory can be formulated in a covariant way, but we do not need this here.

phase space has a natural correspondence with the 1-particle subspace of $\mathscr{H}$. In the simplest case the dynamics is simply lifted from the classical Hamiltonian flow on V, so that we can regard the induced evolution as the canonical second quantization of a 1-particle evolution. In this way we can preserve a very close correspondence with the 1-particle theory (or, equivalently, with the c-number solutions to the field equations). Indeed, although we start from a field theory, the relationship of the field to the particle interpretation is the same as in the canonical second quantized theory.[30]

For interacting theories of physical interest (even linear theories) one cannot put this construction on any simple basis; in particular, there does not exist a canonical complex structure J which is preserved under the time evolution. We see that the complex structure J, which tells us what we mean by complex numbers in the Hilbert space theory, also tells us what we mean by particle number (or more generally a particle interpretation) for a quantum field. The favourable case roughly coincides with the situation where the field is kinematic (it actually includes time-independent external couplings); otherwise we shall suppose that interactions lead to a change in J and the particle interpretation is shifted; quanta have been created or destroyed.

In perturbation theory, too, one defines the asymptotic states (and, by the assumption of completeness, even the interacting states) in terms of the kinematic description of the quantum field. So this is a familiar limitation. What is the kinematic description? It is provided by the decomposition of the field into positive- and negative-frequency parts. Here the complex numbers that enter into the classical fields play a crucial role; for a given spacelike hypersurface $\mathscr{S}$ and solution ϕ, one finds functions ϕ^+ and ϕ^- such that $\phi = \phi^+ + \phi^-$, and when $\mathscr{S}$ is translated in time the complex phase of these functions on $\mathscr{S}$ rotates in opposite directions.[31]

This complex structure, the i that (may) appear in the field equations, we shall call the *natural complex structure*. These are the

[30] On this basis I shall at times speak of the field quantization as a canonical second quantization of a 1-particle theory. By this I mean no more than that from the field quantization one can read off the 1-particle theory, to which it is related by the functor Γ.

[31] A sufficient condition for the decomposition to be possible is that the space–time admits a timelike Killing vector field everywhere orthogonal to a family of spacelike hypersurfaces; see e.g. Ashtekar and Magnon (1975).

complex numbers that are usually used in Hilbert space theory. They are also the complex numbers that are used to define the gauge transformation properties of the fields (and to define the gauge-invariant objects in the theory, i.e. the observables). We recall that, if we switch this gauge transformation to the Hilbert space level, the particle and antiparticle states are rotated in opposite directions (cf. fn. 23). In some sense, multiplication by i at the level of the fields is mirrored in the Hilbert space by multiplication by the imaginary unit (of the Hilbert space) on the particle states, and by minus the imaginary unit on the antiparticle states. We may conjecture that it is J that determines what we mean by complex numbers in the quantum theory of a classical system, in particular by the particle interpretation of a quantum field, and that J is related to the natural complex structure through the decomposition into positive- and negative-frequency parts. That is just what happens; J is given by multiplication by i for a positive-frequency solution, and multiplication by $-i$ for a negative-frequency solution.

To see the implications for what we mean by positive- and negative-energy 1-particle states, let us use the canonical second quantization and suppose that the free evolution of the quantum field is generated by the free evolution on the 1-particle subspace as in equation (5), which by the foregoing can be identified with the space V_J. In that case, with respect to J, the field evolves as the canonical second quantization Γ of the unitary evolution:

$$f \rightarrow \exp(-JHt/\hbar)f$$

where $f \in V_J$. The requirement that the energy be positive means that H must be a positive operator. (It is self-adjoint because of Stone's theorem.) Equivalently, H can have only positive (generalized) eigenfunctions. Since we know that the solution manifold V contains positive- and negative-frequency solutions, this appears inconsistent with the infinitesimal form of the evolution (the Schrödinger equation). However, it is the complex structure J which must now be used; that is, we now have

$$Hf = J\hbar \frac{\partial f}{\partial t}. \tag{6}$$

When J is given by multiplication by $-i$ on the negative frequency solutions, of the form $f^- = w \exp(iEt/h)$, we obtain $Hf^- = Ef^-$. If J has the natural action (multiplication by i) on the positive frequency states, then H will be a positive operator on V_J; since (6) must yield the same evolution as the field equation, it is obvious that $H=-iJH_N$, where H_N is the usual Dirac Hamiltonian. (The meaning of this notation will shortly become clear.)

To make further progress, we need a little more of the theory of complex structures on orthogonal and symplectic spaces. For the bosonic field, we assume that the solution manifold V is a complex linear vector space equipped with a symplectic form ω, with the usual complex structure given by multiplication by i (the natural complex structure). In favourable cases one can find a canonical mapping J on V (i.e. with $\omega(Jf, Jg)=\omega(f, g)$) such that $J^2=-1$. It then automatically follows that

$$(f, g)_J=\omega(f, Jg)$$

is a symmetric form on V, and that

$$\langle f, g\rangle_J=(f, g)_J+i\omega(f, g)$$

is sesquilinear on V_J; that is, it is sesquilinear with respect to multiplication of elements of V by 'the complex numbers' $a+Jb$, $a, b\in\mathbb{R}$. When $(.\,,.)_J$ is non-degenerate, it follows that $\langle .\,,.\rangle_J$ provides a Hilbertian norm and we may complete to obtain a Hilbert space.

This is how we obtain the bosonic field theory; to obtain the fermionic theory we have instead of a symplectic form a *symmetric non-degenerate bilinear form.* For the Dirac field it is

$$S(\psi, \phi)=\frac{1}{2}\left(\int \bar{\psi}\phi\, \mathrm{d}^3x+\int \psi\bar{\phi}\, \mathrm{d}^3x\right) \tag{7}$$

(here and in the following the spinor summation is suppressed); now $S(J\psi, \phi)$ is automatically antisymmetric and we may write

$$\langle f, g\rangle_J=S(\psi, \phi)+iS(J\psi, \phi), \tag{8}$$

so that in both cases what fixes the Hilbert space and the properties of the operators is the complex structure J. In particular, we must choose J such that the Hamiltonian H is a positive operator.

This is at the 1-particle level; at the level of the Fock space, the creation and annihilation operators also depend critically on the complex structure J. To see this we recall equation (1):

$$a(if)=-ia(f)$$
$$a^*(if)=ia^*(f).$$

Therefore in terms of the real-linear self-adjoint field,

$$\Phi(f)=(\hbar/2)^{1/2}[a(f)+a^*(f)],$$

we have[32]

$$a(f)=(2\hbar)^{-1/2}[\Phi(f)+i\Phi(if)]$$

$$a^*(f)=(2\hbar)^{-1/2}[\Phi(f)-i\Phi(if)], \tag{9}$$

and, as follows from the action of a, a^*, we see that

$$[a(f), a^*(g)]_{\pm}=\langle f, g\rangle,$$

and that for fermions

$$[\Phi(f), \Phi(g)]_{+}=\hbar S(f, g),$$

while for bosons

$$[\Phi(f), \Phi(g)]_{-}=i\hbar\omega(f, g).$$

In the present approach it is the field Φ that is considered fundamental, defined by an algebra independent of the complex structure J. If, in the foregoing, we consider the Hilbert space complex numbers given by J, we obtain from (9) new creation and annihilation operators:

$$a_J(f)=(2\hbar)^{-1/2}[\Phi(f)+i\Phi(Jf)]$$

$$a_J^*(f)=(2\hbar)^{-1/2}[\Phi(f)-i\Phi(Jf)], \tag{10}$$

The field Φ is called the *Segal field*. It is real-linear, causal, and self-adjoint.[33] Assuming that it is given, the freedom in the particle interpretation corresponds to the freedom in the choice of J. Once J is chosen, then (10) tells us what are to count as creation and annihilation operators, and it is J that tells us the (anti)commutation

[32] Usually the factor $\sqrt{\hbar}$ appears only in the relationship between the Segal field and the creation and annihilation operators in the bosonic case. The issue here is a little subtle and I shall not pursue it. For a conservative critique, see Rosenfeld (1963: 355).

[33] In non-relativistic quantum mechanics, the Segal field is of the form $P+Q$, where P and Q are the momentum and position operators. Its physical interpretation is obscure; mathematically, it is the generator of the Weyl algebra of the fields (boson case), and algebraically, it generates the Clifford algebra of the fields (fermion case).

relationships for a_J, because it defines the inner product $\langle . , . \rangle_J$ (via (8)). In this way, we find[34]

$$[a_J(f), a_J^*(g)]_\pm = \langle f, g \rangle_J$$
$$[a_N(f), a_N^*(g)]_\pm = \langle f, g \rangle_N . \qquad (11)$$

As we have seen, if we choose the complex structure

$$J = iP^+ - iP^-$$

(where $P^\pm$ are projection operators on to the positive- and negative-frequency subspaces of V_J), the negative-frequency solutions no longer have negative energy in the sense of Schrödinger, i.e. according to (6). We shall call this choice of J the *particle* complex structure. *We make the natural assumption that the negative-frequency states are the positive-energy antiparticle states.*

Now consider the relationship between the creation and annihilation operators defined by the natural and particle complex structures (what we shall call the natural and particle creation and annihilation operators). The two are linked through the Segal field Φ, which is independent of the complex structure. We see that:

$$\begin{aligned}(2\hbar)^{1/2}a_N(f) &= \Phi(f) + i\Phi(if) = \Phi(f) + i\Phi[J(P^+ - P^-)f]\\ &= \Phi(f^+) + \Phi(f^-) + i\Phi(Jf^+) - i\Phi(Jf^-)\\ &= (2\hbar)^{1/2}[a_J(f^+) + a_J^*(f^-)],\end{aligned}$$

and similarly for $a_N^*(f)$; that is,[35]

$$a_N(f) = a_J(f^+) + a_J^*(f^-)$$
$$a_N^*(f) = a_J(f^-) + a_J^*(f^+).$$

We see that the natural annihilation and creation operators are combinations of annihilation and creation operators for particles and antiparticles; this is just the particle interpretation of the (usual) physical fields, for example the Dirac field ψ and its adjoint ψ^*. We

[34] Quantities defined with respect to the natural (or particle) complex structure have subscript N (respectively, J).

[35] So far as I know, these equations first appeared in Bongaarts (1972). Segal treated only the real scalar field; for the treatment of the complex scalar field, see Saunders (1989, sect. 3.4).

can therefore identify a_N with ψ, etc., and $a_J(f^+)$ with b, $a_J(f^-)$ with d, and $a_J^*(f^-)$, $a_J^*(f^+)$ with b^* and the d^*, respectively.[36]

Now that we have the particle creation and annihilation operators, we can canonically second-quantize any 1-particle operator. Since for the identity $\mathbb{1}$,

$$\langle f^+, \mathbb{1}g^-\rangle_J = \langle f^+, g^-\rangle_J = 0,$$

we have from (5) that

$$\mathrm{d}\Gamma_J(\mathbb{1}) = \sum_k a_J^*(f_k)a_J(f_k) = \sum_k a_J^*(f_k^+)a_J(f_k^+) + \sum_k a_J^*(f_k^-)a_J(f_k^-).$$

Therefore with the obvious identifications,

$$\mathrm{d}\Gamma_J(\mathbb{1}) = N^+ + N^-.$$

Similarly, one sees that the total energy is positive: if H_N is the Hamiltonian defined by the natural complex structure, then $-iJH_N$ is the particle Hamiltonian which is clearly positive (what we denoted H in (6)), and its second quantization is the sum of the energy of all the particles and the energy of all the antiparticles. The 1-particle *charge operator*, on the other hand, is given by iJe; obviously, the positive-frequency states have eigenvalue $-e$, and the negative frequency states have eigenvalue $+e$; the total charge operator is its canonical second quantization:

$$\mathrm{d}\Gamma_J(iJe) = -eN^+ + \lambda eN^-.$$

In fact, all the global kinematic observables can now be obtained by a canonical second quantization with respect to the particle complex structure. At the same time, since we now see that the physical fields are the natural annihilation and creation operators (with their particle interpretation fixed by the particle complex structure), the conventional theory can now be understood as a *canonical second quantization with respect to the natural complex structure followed by normal-ordering*. Both the Fock space action of the natural creation and annihilation operators and the normal-ordering process are defined by the particle complex structure.

It is helpful to prove this equivalence in detail. For this we need the explicit relationship between $\langle .\,,.\rangle_J$ and $\langle .\,,.\rangle_N$. From (7), (8) it follows that

[36] The identification is figurative (see fn. 17).

$$\langle f, g\rangle_N = \int \bar{f} g \, \mathrm{d}^3x,$$

whereas

$$\langle f, g\rangle_J = \langle f^+, g^+\rangle_N + \langle g^-, f^-\rangle_N. \tag{12}$$

(Note carefully the order of f and g.) We also need the relationship between 1-particle operators defined by the two complex structures. To determine this we consider only those observables X which can be obtained from the classical theory and which preserve the two complex structures; X then generates a one-parameter group of orthogonal transformations on V, which is the same as the group of transformations generated by X_N on V_N and by X_J on V_J (each as generators of unitary transformations with respect to the relevant complex structure). It then follows that

$$X_N = -iJX_J. \tag{13}$$

(We have already used this relationship above.)

The equivalence in question is therefore between

$$:\mathrm{d}\Gamma_N(X_N): = :\sum_{kj} a_N^*(f_k)\langle f_k, X_N f_j\rangle_N a_N(f_j):$$

(where we must normal-order with respect to the action of a_N on $\mathscr{F}(V_J)$, that is in terms of the particle complex structure J), and

$$\mathrm{d}\Gamma_J(X_J) = \sum_{kl} a_J^*(f_k)\langle f_k, X_J f_l\rangle_J a_J(f_l).$$

In evaluating the first expression, the requirement that the Hamiltonian flow generated by X preserves the complex structures (in particular J) means that X_N, X_J do *not* connect negative- and positive-frequency states (i.e. $\langle f^+, X_N f^-\rangle_N = \langle f^+, X_J f^-\rangle_J = 0$). We then obtain

$$:\mathrm{d}\Gamma_N(X_N): = \sum_{k,l} [a_J^*(f_k^+)a_J(f_l^+)\langle f_k^+, X_N f_l^+\rangle_N$$
$$-a_J^*(f_l^-)a_J(f_k^-)\langle f_k^-, X_N f_l^-\rangle_N].$$

(Note carefully the order of the indices k, 1; the minus sign is due to normal ordering.) From (12) and (13), it now follows that

$$\langle f_k^+, X_N f_l^+\rangle_N = \langle f_k^+, X_J f_l^+\rangle_J$$
$$\langle f_k^-, X_N f_l^-\rangle_N = -\langle f_l^-, X_J f_k^-\rangle_J,$$

and we obtain

$$:\mathrm{d}\Gamma_N(X_N):=\sum_{k,l}[a_J^*(f_k^+)a_J(f_l^+)\langle f_k^+, X_J f_l^+\rangle_J + a_J^*(f_l^-)a_J(f_k^-)\langle f_l^-, X_J f_k^-\rangle_J];$$

that is,

$$:\mathrm{d}\Gamma_N(X_N):=\mathrm{d}\Gamma_J(X_J), \tag{14}$$

as claimed.

8 Interpretation

What is the upshot of all this? We cannot say that the conventional theory is equivalent in all respects to the canonical second quantized theory with respect to the particle complex structure; this is true only for a limited class of global operators (which preserve particle number). In particular, the equivalence does not hold for local multiplicative operators, for these connect positive- and negative-frequency states. (They are 'odd' operators, in the sense of Schrödinger; equivalently, they do not commute with J.[37]) For these the RHS of (14), if considered a perturbation, would induce transitions from particle to antiparticle states, which would be a complete disaster. In this situation we must read back from the standard formalism; it seems that the complex structure must also change under the evolution. One must abandon or extend the canonical theory, because one does not have a fixed Hilbert space (of the form $\mathscr{F}(V_J)$) to host a unitary evolution.

In the standard formalism, the LHS of (14) still makes sense as an operator on $\mathscr{F}(V_J)$: that is why it is possible to develop a formal perturbation theory. From the canonical point of view, the standard formalism is a quantum mechanics over *two* complex structures; the natural complex structure is used to determine the creation and annihilation operators (which are the physical fields) and the

[37] There are intimate connections with the problem of defining a (position space) Born interpretation in relativistic theory; in the Foldy–Wouthuysen representation, equivalently the Newton–Wigner representation, J is a local operator. To keep the discussion within reasonable bounds I shall not pursue the matter here; relevant material may be found in Segal (1964), Goodman and Segal (1965), Streater (1988), and Saunders (1989).

canonically second quantized operators, but their action on Fock space is expressed in terms of the particle complex structure. (The physical Fock space is $\mathscr{F}(V_J)$, and not $\mathscr{F}(V_N)=\mathscr{F}(V)$.) Just because the two complex structures do not coincide, the operators $:\mathrm{d}\Gamma_N(X_N):$ (or $\mathrm{d}\Gamma_N(X_N)$, for that matter) contain products of linear combinations of particle creation and annihilation operators. Therefore, they describe pair creation and annihilation processes.

In the non-relativistic field theory, the particle complex structure is identical to the natural complex structure. In this case the physical fields are simply the canonical particle creation and annihilation operators; in the 1-particle theory, every real-linear operator extends to a complex-linear (or antilinear) operator. The implication of the foregoing is that these are special features of the non-relativistic limit which do not survive in the relativistic theory, not even in kinematics.

There is also the implication that in some sense the relativistic theory is the canonical theory where the concept of *charge* replaces the concept of *particle*. What the natural creation and annihilation operators (the physical fields) create and destroy are *units of charge*. The natural complex structure attaches to the charge, and the particle complex structure to the particle number. This is borne out by the role of the total charge and number operators as generators of phase transformations (gauge transformations of the first kind). The charge generates phase transformations in the physical fields; i.e.,

$$\exp(i\mathrm{d}\Gamma_N(\mathbb{1})\theta): a_N(f)\to a_N(\exp(i\theta)f)=\exp(-i\theta)a_N(f)$$

(using the antilinearity of a_N on $\mathscr{F}(V_N)$), but the number operator $\mathrm{d}\Gamma_J(\mathbb{1})$ generates phase transformations in the complex numbers that are used in the Hilbert space:

$$\exp(J\,\mathrm{d}\Gamma_J(\mathbb{1})\theta): a_J(f)\to a_J(\exp(J\theta)f)=\exp(-i\theta)a_J(f)$$

(using the antilinearity of a_J on $\mathscr{F}(V_J)$). For example, if we express the action of $\exp(i\,\mathrm{d}\Gamma_N(\mathbb{1})\theta)$ on a_N, in terms of the physical Hilbert space $\mathscr{F}(V_J)$, we find the curious properties that we noted were required in the Weinberg construction:

$$\begin{aligned} a_N(f)\to a_N[\exp(i\theta)f] &= a_J[\exp(J\theta f^{+})]+a_J^{*}[\exp(-J\theta f^{-})] \\ &= \exp(-i\theta)a_J(f^{+})+\exp(-i\theta)a_J^{*}(f^{-}). \end{aligned}$$

Therefore, despite the fact that according to the canonical theory the annihilation and creation operators transform oppositely, the particle annihilation operator and antiparticle creation operator transform in the same sense under the gauge transformations of the physical fields. We see explicitly that the gauge transformations of the Dirac fields *are* induced by gauge transformations on the Hilbert space, but by the natural complex structure and not by the particle one. The conserved quantity is the charge;[38] the conserved quantity under rotations in the particle complex structure is the particle number. Because the interaction Hamiltonian is required to be gauge-invariant with respect to the natural complex structure (i.e. bilinear in the natural annihilation and creation operators, the Dirac field, and its adjoint), it is the former and not the latter that is conserved in the dynamics.

It is now easy to understand why the charge conjugation operator has such different properties in the 1-particle theory and in the standard formalism. Recall that this operator is unitary in the field theory, but antilinear in 1-particle theory; further, it does not (as it should) change the sign of the 1-particle charge. We can formulate these differences as follows. Whereas

$$\mathfrak{C}::\mathrm{d}\Gamma_N(-e\mathbb{I}):\rightarrow -:\mathrm{d}\Gamma_N(-e\mathbb{I}):,$$

in the 1-particle theory the operator $-e\mathbb{I}$ is invariant under any unitary (or antiunitary) mapping defined on V_N. But there is a corresponding map on V_J, because $:\mathrm{d}\Gamma_N(-e\mathbb{I}):=\mathrm{d}\Gamma_J(ieJ)$ and the latter is a canonical second quantization (with no normal ordering); hence in particular $\Gamma(U)\,\mathrm{d}\Gamma_J(X)\Gamma(U)^{-1}=\mathrm{d}\Gamma(UXU^{-1})$ (where U is unitary if and only if $\Gamma(U)$ is unitary). On V_J the charge conjugation must be unitary.

This is just what we find; a simple calculation shows that $(Jf)^c=Jf^c$.[39] With respect to our new notion of complex numbers, the charge conjugation is linear, despite the complex conjugation contained in its action (intuitively, $(J(f_1^+ + f_2^-))^c = -if_1^{+c} + if_2^{-c} \approx -if_1^- + if_2^+ = J(f_1^- + f_2^+) = Jf^c$). At the same time, the 1-particle charge now changes sign, because it is no longer the identity on V_J, but rather the operator eiJ, so that under $\mathscr{C}$ (using its linearity with respect to J, and antilinearity with respect to N),

$$\mathscr{C}eiJ\mathscr{C}^{-1}=-eiJ.$$

[38] Strictly speaking, we should be second-quantizing $-e\mathbb{I}$ rather than $\mathbb{I}$ to obtain the physical charge.

[39] In detail: in a representation with γ^0 real, one has $\gamma^0=\gamma^{0t}$, $\gamma^0\gamma^{\mu\dagger}\gamma^0=\gamma^\mu$ († is complex conjugation followed by matrix transposition). Recalling that $\mathscr{C}\gamma^\mu\mathscr{C}^{-1}=-\gamma^{\mu t}$, $P^\pm=\pm i\gamma^\mu p_\mu/m$, then $(Jf)^c=\mathscr{C}\gamma^{0t}\overline{Jf}=-i\mathscr{C}\gamma^{0t}(\gamma^\mu p_\mu)^{\dagger t}\gamma^{0t}\mathscr{C}^{-1}f^c=i\gamma^\mu p_\mu f^c=Jf^c$.

In summary, we are led to a point of view in which the (linear) standard formalism appears as a quantum mechanics of charge ('charge dynamics'), represented (through the normal ordering) in the kinematic limit as a quantum mechanics of particles. In both cases the quantum mechanics is in canonical second quantized form. On this view there is no non-trivial particle dynamics on a fixed complex-linear space (not even in Fock space[40]); that is, there no particle Hilbert space description of the dynamics. In the kinematic case, only observables that commute with J can be defined.

To pursue these implications would take us too far afield and into difficult terrain. For what it is worth, C^*-algebra theory also leads to a similar conclusion, in so far as one finds that the evolution cannot be unitarily implemented within any one representation, and in semi-classical quantum gravity the metric dependence of the particle interpretation of quantum fields has a natural expression in terms of inequivalent complex structures on Hilbert space.[41] The almost complete failure of the constructive programme in quantum field theory—in which the existence of a particle Fock space is an axiom—is in itself remarkable.[42] The present theory offers a new perspective which is both simple and radical. It is simple because (for linear fields) the standard formalism can be understood as a canonical second quantization; contact with the 1-particle theory is preserved. Relativistic quantum theory is cast into a form almost identical to the non-relativistic theory. It is radical because the modification of the complex structure ramifies throughout the interpretive and mathematical framework of the theory.[43] It may be that one can formulate a

[40] In the present framework Fock space is defined as $\mathscr{F}(V_J)$; J cannot be preserved by the evolution.

[41] See Ashtekar and Magnon (1975), Woodhouse (1980: 284–7). For a self-contained introduction along more conventional lines, see Birrel and Davies (1982).

[42] This programme has culminated in the result that $\lambda\phi^4$ theory in $3+1$ dimensions, a super-renormalizable quantum field theory, has in fact only the trivial solution $S=1$. A similar conclusion seems to hold for QED itself. For a review see e.g. Huang (1989).

[43] It seems to me that the absence of complex linearity during interactions, or with respect to certain operators, *may* have a decisive bearing on measurement theory. As I have argued elsewhere (Saunders 1988), the measurement problem requires exact mathematics; we know that the world is not Galilean, and that relativistic effects, however small, are always present in the measurement process. See also the conjecture of Maxwell (1988), according to which scattering processes differing in particle number for the outgoing states should *not* be considered coherent, and that of Penrose (1989), where the '1-graviton' criterion (for longitudinal gravitons) likewise signals the breaking of coherence. The present framework may provide a theoretical basis for these conjectures.

theory in which the complex structure is changing with time; however, we must first formulate an interpretation of a Hilbert space theory in which the usual local operators are no longer complex-linear. If these things can be done, we will have an interpretation of relativistic theory of some beauty; the complex numbers of the Hilbert space theory will determine the particle interpretation and will themselves change with time, subject to the dynamics. In this way the existence of particles is built into the dynamics. For linear interactions this could even be studied at the 1-particle level.

This is a radical view; a more conservative approach is that there is no exact theory of interactions, not because complex linearity fails but because the standard formalism is incomplete.[44] Actually, the best one can do is claim that, *so long as one has a scattering situation*, there is a relativistic quantum theory on a particle Hilbert space. There is, after all, a clear enough intuition of a particle dynamics, in which an incoming particle state evolves into a coherent superposition of outgoing particle states, all within a fixed Hilbert space. This seems to conform to the basic framework of elementary quantum mechanics.

But charge superselection, the fact that only gauge-invariant operators count as physical observables, is not an incidental feature to this picture; on the contrary, charge superselection, and with it the construction of a field theory over the tensor product of the particle and antiparticle Fock spaces, can be looked on as a way of rewriting quantum mechanics over a 'number' field with imaginary unit of the form J, in the conventional format of the non-relativistic theory. The unrestricted complex linearity of the non-relativistic theory is preserved through elaborate constraints on what is to count as an observable. The distinction between the two Fock spaces still depends on the decomposition of the field into positive- and negative-frequency parts (this is why we need a scattering situation); each Fock space has the natural complex structure, but the usual self-adjoint operators of elementary theory cannot act on these spaces. (They map states out of each Fock space altogether.)

Is this *unrestricted* complex linearity? This may better be considered a fiction; if it is so qualified at the relativistic level, the

[44] This point of view is particularly natural in path-integral quantization. Although the present theory indicates that the standard formalism actually describes the dynamics in the natural Fock space ('charge dynamics'), the path-history space appears to preserve complex linearity at the expense of unidirectional evolution in time.

unrestricted version that appears when we descend to the non-relativistic limit may be an idealization of no fundamental physical significance.

I shall conclude with a last look at the negative-energy sea. The complex structure J leads to a reinterpretation of the negative-energy solutions, much as did the hole theory; consider now their relationship. Here we must bear in mind the fact that the Dirac negative-energy sea enforces a dynamical interpretation as well as a kinematic one. The properties of antimatter were deduced both from the assumption of the Dirac vacuum and from dynamical considerations (transitions to vacant negative-energy states with the emission of energy). By means of the exclusion principle, these properties could be deduced from the physical picture of the vacuum.

Consider now the following theorem, which holds when V is a *finite* (say s)-dimensional space. (I owe this observation to Professor G. Segal of the Mathematical Institute, University of Oxford.) In the fermion case the natural Fock space is the exterior algebra over V, i.e. $V_0 \oplus_{n=1}^{s} \Lambda_{i=1}^{n} V_i$ (where each V_i is a copy of V and V_0 is the one-dimensional vector space $\mathbb{C}$). Consider each n-particle subspace $\Lambda_{i=1}^{n} V_i$; there is a unique *antilinear* isomorphism with the $(s-n)$-particle subspace, or a linear isomorphism $\approx$ with the complex conjugate $(s-n)$-particle subspace:

$$\overset{n}{\underset{i=1}{\Lambda}} \bar{V}_i \approx \overset{s-n}{\underset{i=1}{\Lambda}} V_i. \tag{15}$$

(This is the generalization of the Hodge $*$-operator to the complex case.[45]) For the negative-frequency subspace of V (the antiparticle states), this is equivalent to replacing the action of multiplication by i by multiplication by $-i$, so it also implements the change from the natural complex structure on these states to the particle complex structure. It does so, however, by replacing a vector in the n-antiparticle state by a vector in the $(s-n)$-particle state. But this is the prescription of the Dirac hole theory; the vacuum state, with $n=0$ and s infinite, is replaced by the $(s-n=\infty)$-particle state[46]—the negative-energy sea. We can guess what is going on; even when $n \neq 0$,

[45] The Hodge $*$ transformation is an isomorphism resting on the canonical duality between differential forms and tangent vectors. For real spaces this duality is bilinear; in the complex case it is sesquilinear. (There is no positive-definite non-singular bilinear form on $V \times V$, with V complex.)

[46] Note that, for s finite, both states are rays, i.e. one-dimensional.

(15) continues to ensure that the 'Dirac dual' $(s-n)$-particle state,[47] using the natural complex structure, is equivalent to the physical n-antiparticle state using the particle complex structure. The holes behave as do the antiparticles.[48]

We see that the Dirac vacuum enforces a different notion of complex numbers at the Hilbert space level from that suggested by the non-relativistic theory. We may conclude that the negative-energy sea is what the particle vacuum looks like using the wrong notion of complex numbers (the natural complex structure). If the particle vacuum is to appear really empty, then we must use the particle complex structure at the Hilbert space level.

References

Anderson, C. D. and Anderson, H. L. (1983), 'Unraveling the Particle Content of Cosmic Rays', in L. M. Brown and L. Hoddeson (eds.), *The Birth of Particle Physics*, Cambridge University Press, pp. 131–54.

Ashtekar, A. and Magnon, A. (1975), 'Quantum Fields in Curved Space-times', *Proc. Roy. Soc. London* A346: 375–94.

Bargmann, V. (1961), 'On a Hilbert Space of Analytic Functions and an Associated Integral Transform, Part 1', *Comm. Pure App. Math.* 14: 187–214.

Birrel, N. and Davies, P. (1982), *Quantum Fields in Curved Space*, Cambridge University Press.

Bongaarts, P. (1972), 'Linear Fields according to I. E. Segal', in R. Streater (ed.), *The Mathematics of Contemporary Physics*, Academic Press, New York, pp. 187–208.

Bromberg, J. (1976), 'The Concept of Particle Creation before and after Quantum Mechanics', *Historical Studies in the Physical Sciences*, 7: 161–91.

Brown, H. and Harré, R. (1988), *The Philosophy of Quantum Field Theory*, Clarendon Press, Oxford.

Cook, J. (1953), 'The Mathematics of Second Quantization', *Trans. Am. Math. Soc.* 74: 222–45.

Dirac, P. A. M. (1926*a*), 'Relativity Quantum Mechanics with an Application to Compton Scattering', *Proc. Roy. Soc.* A111: 405–21.

[47] i.e. the state with $(s-n)$ negative-frequency electrons present, where s is the total number of states available; alternatively, the Dirac vacuum with n holes.

[48] This conclusion appears valid even when J is changing with time, that is for arbitrary evolutions. At each instant there is a correspondence between the holes and the antiparticles.

—— (1926*b*), 'The Compton Effect in Wave Mechanics', *Cam. Phil. Soc.* 23: 500–7.

—— (1927), 'The Quantum Theory of the Emission and Absorption of Radiation', *Proc. Roy. Soc.* A 114: 243–65.

—— (1928), 'The Quantum Theory of the Electron', *Proc. Roy. Soc.* A117: 610–24.

—— (1929), 'A Theory of Electrons and Protons', *Proc. Roy. Soc.* A126: 360–5.

—— (1931), 'Quantized Singularities in the Electromagnetic Field', *Proc. Roy. Soc.* A133: 60–72.

—— (1934), 'Discussion of the Infinite Distribution of Electrons in the Theory of the Positron', *Proc. Camb. Phil. Soc.* 30: 150–63.

FitzGerald, G. (1881), 'Note on Mr J. J. Thomson's investigation of the electromagnetic action of a moving electrified sphere', *Sci Proc. Roy. Dublin Soc.*, New series 3: 250–54.

Fock, V. (1926), 'Zur Schrödingerschen Wellenmechanik', *Zeit.f. Phys.* 38: 242–50.

—— (1928), 'Verallgemeinerung und Losung der Diracschen Theorie des Elektrons', *Zeit.f. Phys.* 55: 127–40.

—— (1932), 'Konfigurationsraum und zweite Quantelung', *Zeit.f. Phys.* 21: 78–89.

Goodman, R. and Segal, I. (1965), 'Anti-locality of Certain Lorentz-invariant Operators', *J. Math. and Mech.* 14: 629–38.

Heisenburg, W. (1928), Letter to Pauli, 31 July, in Hermann *et al.* (1979, 1: 251).

—— 1934*a*), Letter to Pauli, 8 February in Hermann *et al.* (1979, 2: 279).

—— (1934*b*), 'Bermerkungen zur Diracschen Theorie des Positrons', *Zeit.f. Phys.* 90: 209–31; 92: 692.

—— (1963), Interview with T. Kuhn, 12 July 1963; transcript in the Niels Bohr Library, American Institute of Physics, New York.

—— (1972), 'Development of Concepts in the History of Quantum Theory', in J. Mehra (ed.), *The Physicist's Conception of Nature*, D. Reidel, Dordrecht, pp. 264–75.

Hermann, A., von Meyenn, K., and Weiskopf, V. (eds.) (1979) *Wolfgang Pauli, Scientific Correspondence*, Springer, New York.

Huang, K. (1989), 'Triviality of the Higg's Field', *Int. J. Mod. Phys.* A4: 1037–53.

Joos, H. (1962), 'Zum Darstellungstheorie der inhomogenen Lorentzgruppe als Grundlage quantenmechanischer Kinematik', *Fort. d. Phys.* 10: 65–146.

Jordan, P. and Klein, O. (1927), 'Zum Mehrkorproblem der Quantentheorie', *Zeit f. Phys.* 45: 751–65.

Klein, O. (1926), 'Quantentheorie und fünfdimensionale Relativitätstheorie', *Zeit f. Phys.* 37: 895–906.

Klein, O. (1929), 'Der Reflexion von Elektronen an einem Pontialsprung nach der relativistischen Dynamik von Dirac', *Zeit. f. Phys.* 53: 157–65.

Kostant, B. (1970), 'Quantization and Unitary Representations', in C. T. Taam (ed.), *Lectures in Modern Analysis*, iii: *Lecture Notes in Mathematics*, vol. 170, Springer, Berlin, pp. 87–208.

Mackey, G. (1963), 'Group Representations in Hilbert Space', in I. Segal, *Mathematical Problems of Relativistic Physics*, American Mathematical Society, Providence, RI, pp. 113–30.

Maxwell, N. (1988), 'Quantum Propensiton Theory: A Testable Resolution of the Wave/Particle Dilemma', *Brit. J. Phil. Sci.* 39: 1–50.

Neumann, J. von (1932), *Grundlagen der Quantenmechanik*, Springer, Berlin. English trans. R. T. Beyer, Princeton University Press, 1955.

Novozhilov, Y. (1975), *Introduction to Elementary Particle Theory*, Pergamon Press, Oxford.

Pais, A. (1986), *Inward Bound*, Clarendon Press, Oxford.

Penrose, R. (1989), *The Emperor's New Clothes*, Clarendon Press, Oxford.

Rosenfeld, L. (1963), 'On Quantization of Fields', *Nuc. Phys.* 40: 201–20.

Saunders, S. (1988), 'The Algebraic Approach to Quantum Field Theory', in Brown and Harré (1988: 149–86).

—— (1989), *The Mathematical and Philosophical Foundations of Quantum-Field Theory*, Ph.D. thesis, University of London.

Segal, I. (1964), 'Quantum Field and Analysis in the Solution Manifolds of Differential Equations', in W. Martin and I. Segal (eds.), *Analysis in Function Space*, MIT Press, Cambridge, Mass., pp. 129–53.

—— (1967), 'Representations of the Canonical Commutation Relationships', in F. Lurcat (ed.), *Cargese Lectures on Theoretical Physics*, Gordon and Breach, New York.

Simon, B. (1974), *The $P(\phi)_2$ Euclidean (Quantum) Field Theory*, Princeton University Press.

Souriau, J. (1966), 'Quantification Géométrique', *Comm. Math. Phys.* 1: 374–98.

Streater, R. (1988), 'Why Should Anyone Want to Axiomatize Quantum Field Theory?' in Brown and Harré (1988: 135–48).

Stueckelberg, E. and Guenin, M. (1962), 'Quantum Theory in Real Hilbert Space', *Helv. Phys. Acta* 34: 675–98.

Sudarshan, E. C. G, and Mukunda, N. (1974), *Classical Dynamics: A Modern Perspective*, John Wiley, New York.

Tamm, I. (1930), 'Über die Wechselwirkung der freien Elektronen mit der Strahlung nach der Diracschen Theorie des Elektrons und nach der Quantenelektrodynamik', *Zeit. f. Phys.* 62: 545–68.

Tomonaga, S. (1962), *Quantum Mechanics*, vol. II: *New Quantum Theory*, North-Holland, Amsterdam.

Waller, I. (1930), 'Die Strenung von Strahlung durch gebundene und frie

Elektronen nach der Diracschen relatistischen Mechanik', *Zeit.f. Phys.* 61: 837–51.

Weinberg, S. (1964), 'Feynman Rules for Any Spin', *Phys. Rev.* D133: 1318–22.

Wightman, A. (1972), 'The Dirac Equation', in A. Salam and E. Wigner (eds.), *Aspects of Quantum Theory*, Cambridge, University Press, pp. 95–115.

Wigner, E. (1939), 'On Unitary Representations of the Inhomogeneous Lorentz Group', *Ann. of Math.* 40: 149–204.

Woodhouse, N. (1980), *Geometric Quantization*, Clarendon Press, Oxford.

5

The Vacuum on Null Planes

GORDON N. FLEMING

1 Historical Introduction

In 1949 Dirac published a comparison of three distinct approaches that one could take to describing the initial conditions and dynamical evolution of Lorentz covariant quantum theory (Dirac 1949, 1950). One motivation for presenting the comparison was Dirac's recognition that, in the two approaches that were not used at the time, the description of dynamical evolution had simpler properties, in some respects, than those in the conventional approach. The conventional approach was called by Dirac the Instant Form of quantum dynamics, and the two alternatives were called the Front Form and the Point Form.

The Instant Form employs the specification of initial data and fundamental commutation relations among the basic dynamical variables on spacelike hyperplanes, which can always be associated with a definite *instant* in some inertial frame (Fig. 1). The dynamical evolution of the system is determined by the structure, expressed as a functional of the basic dynamical variables, of those Poincaré group generators which, interpreted actively, are associated with transformations that change the initial data hyperplane.[1] In that frame of reference in which the initial data hyperplane is instantaneous, those generators are: (1) the time component of the total 4-momentum (Hamiltonian), for infinitesimal time translations, and (2) the boost generators, for infinitesimal reorientations of the hyperplanes. The remaining six generators of the Poincaré group—those for spatial translations and rotations—correspond to transformations that

Presented to the 1987 Oxford University Symposium on the Vacuum in Quantum Field Theory, June 1987.

[1] I take 'Poincaré group' to be synonymous with Inhomogeneous Lorentz group. See Wigner (1939); Wightman (1960: 159).

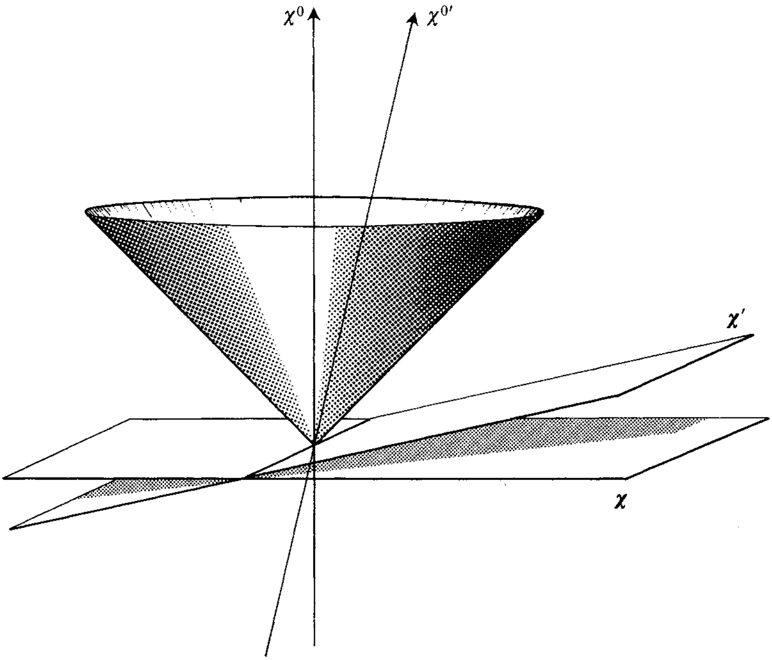


FIG. 1 Spacelike hyperplanes and the light cone for Instant Form quantization (in 2 + 1 dimensions).

leave the instantaneous hyperplanes unchanged but transform sets of initial data on them into other possible sets. These transformations comprise the so-called stability group of the instantaneous hyperplanes; and, whether instantaneous or not (a frame-dependent matter), the stability group of any spacelike hyperplane is isomorphic to the six-parameter Euclidean group of spatial translations and rotations. As stated before, this is the conventional approach to relativistic quantum theory, with which we are all familiar, and even my arguments on behalf of hyperplane dependence[2] are just intended to foster the explicit exploitation of the hyperplane dependence that is inherent in this scheme and/or can be consistently added to it.

In the Front Form, as envisaged by Dirac, the initial data and the fundamental commutation relations are presented on null hyper-

[2] Fleming (1966, 1985, 1988); Fleming and Bennett (1989).

planes, or null planes for short (Fig. 2). These null planes are flat three-dimensional sections of Minkowski space–time which are tangent to light cones and thus, in any inertial frame, have the 'appearance' of an advancing wave-*front* of a plane electromagnetic wave in vacuum.The dynamical evolution of a physical system in this scheme is monitored from null plane to null plane either by translation, with the null plane remaining parallel to itself, or by reorientation (sliding around the light cone), but always preserving the null-plane character. It turns out, as Dirac discovered, that null planes have a seven-parameter stability group, so that the dynamical evolution is determined by the structure of only three independent generators of the Poincaré group. To put it another way, when interactions are turned on in the Front Form, only three Poincaré generators change their structure. In Dirac's view, this recommended the further study of the Front Form to the physics community.

The three Poincaré generators that carry the dynamical structure in the Front Form are not any of the ten generators with which we are familiar from our experience with the Instant Form. Instead, one analyses the Poincaré group with respect to those subgroups and quotients associated with the transformation and propagation of null-plane data. In the context of that analysis, the three dynamical generators appear, the remaining having purely kinematical roles to play. From this analysis one learns that the stability group of a null plane is isomorphic to $(\mathrm{E}_2 \underset{\sim}{\times} \mathrm{D}) \underset{\sim}{\times} \mathrm{T}_3$, where *D* is the dilation group.

In the Point Form the initial data and fundamental commutation relations were given on Lorentz-invariant hyper-hyperboloids of revolution lying inside future light cones. As indicated by their definition, the stability group for these hyper-hyperboloids is isomorphic to the homogeneous Lorentz group, leaving the four translation generators to carry all the dynamical structure. The advantage in this form is that the dynamical evolution is generated by a subgroup of the Poincaré group and the subgroup is Abelian. Nevertheless, with the exception of a few studies exploring the general formalism,[3] this form has not enjoyed any subsequent development, and I will not refer to it further.

The Front Form, however, after lying neglected for two decades, was rediscovered in the late 1960s in the context of so-called current algebra calculations in high-energy particle physics. These calcula-

[3] Fubini *et al.* (1973); Sommerfield (1974) di Sessa (1974); Gromes *et al.* (1974).

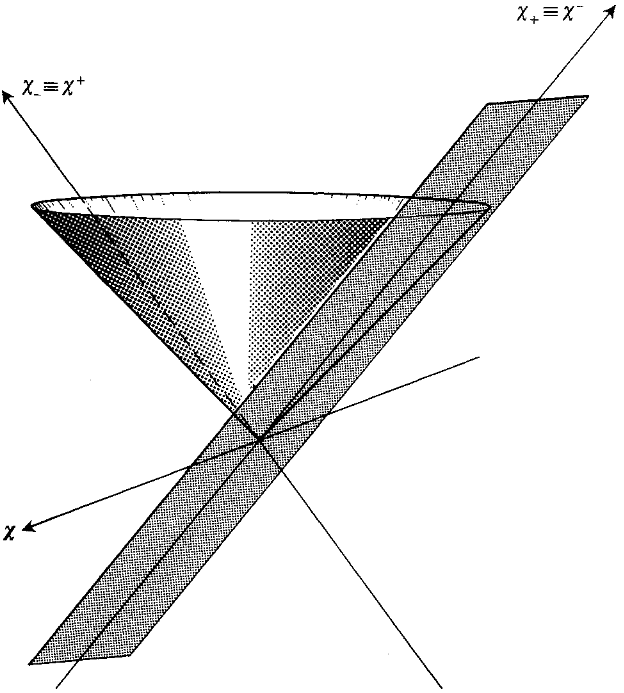


FIG. 2 A null hyperplane with associated coordinates and the light cone for Front Form quantization (in 2+1 dimensions).

tions depended critically upon the use of clever approximations for performing a sum over a complete set of intermediate states in a non-perturbative calculation of matrix elements of commutators of generalized charge and current operators. Fubini and Furlan (1965) showed that these clever approximations were much more easily come by and justified in an inertial frame in which the 3-momentum of some physical system of particles was very, very large, approaching infinity in the ideal limit. In 1966 Weinberg provided an 'explanation' of the utility of these infinite momentum frame calculations by showing that in perturbation theory, treated in the old-fashioned time-ordered way (Heitler graphs rather than Feynman graphs), all vacuum fluctuation graphs and the troublesome Z-graphs disappeared (Weinberg 1966). Susskind (1968) and Bardacki and Halpern (1968) then showed that the infinite momentum limit methods were equivalent to calculations with new time coordinates which labelled members of a set of parallel null planes. With this, Dirac's idea had been rediscovered. Then in the 1970s, systematic expositions of standard quantum field theories in null-plane form,[4] as well as rigorous studies of the foundations of null-plane quantization of fields,[5] appeared. The essential upshot of these studies, which by no means exhausted all the important issues that could be addressed, were as follows.

1 In the absence of interactions, null-plane quantization of fields is physically equivalent to conventional Instant Form quantization of fields. They are two forms of the same (trivial) theory.

2 Notwithstanding this equivalence, the number of independent basic dynamical variables is essentially cut in half when one makes the transition from Instant Form to Front Form or null-plane quantization. Whereas in the Instant Form the commutator (anti-commutator) algebra of the basic fields is, modulo constraints, Abelian (anti-Abelian) and must be augmented by the canonical conjugate fields to form an irreducible algebra, in the Front Form the basic fields themselves already form an irreducible algebra and the canonical conjugate fields are explicit functionals of the basic fields.

3 *The vacuum appears to be trivial in null-plane quantized field theory!* In other words, if one defines the free-field vacuum in the usual way, as the state devoid of field quanta, then turning on interactions does not modify, or undermine, the defined vacuum. The quanta-free state remains the stationary ground state of the system. This means that an interaction-independent Fock space structure can be identified (Fock 1932, 1934) and Haag's Theorem (Haag 1955), prohibiting the interaction picture, does not apply. Perturbatively, this triviality of the vacuum can be attributed to the positive semi-definite spectrum of one component of the null-plane 3-momentum. Such a spectrum prohibits the virtual creation or annihilation of quanta out of or into the vacuum.

4 *The invariance of the vacuum is not the invariance of the world!* That is, Coleman's Theorem (Coleman 1966) does not apply to null-plane quantization. Again, this is a consequence of the spectrum condition mentioned above, coupled with the fact that for massive

[4] Kogut and Soper (1970); Rohrlich (1971); Chang *et al.* (1973); Chang and Yan (1973); Yan (1973); Casher (1876); Bart and Fenster (1977).

[5] Leutwyler *et al.* (1970); Driessler (1975, 1977); Gal-Ezer and Horwitz (1976); Streit (1977).

states the lower end of the semi-definite spectrum corresponds to infinite energy.

5 In the absence of interactions, the field algebras for a given spin but different masses, restricted to a null plane, are unilaterally equivalent (Leutwyler *et al.* 1970). This reversal of the situation found in Instant Form quantization is ultimately traceable to the mass parameter not appearing in the momentum-space-invariant volume element when null-plane variables are employed.

Spurred by these developments, a number of workers adopted null-plane quantization ideas to the introduction of relativistic bound state wave functions for dealing with quark models of hadrons.[6] At the same time, however (late 1970s), strong cautionary notes were sounded concerning the singular nature of certain calculations in the null-plane formalism,[7] and the apparent absence of some desirable features for building a rigorous scattering theory such as the cluster decomposition property (Suzuki *et al.* 1976). Unfortunately, these questions and doubts concerning the internal consistency of null-plane quantization were not finally resolved before the interests of the theoretical physics community were redirected by the move to centre-stage of the topics of instantons and non-perturbative vacuum structure,[8] supersymmetry,[9] and, lastly, string theories.[10]

Today the null-plane quantization of fields is no longer systematically studied. We do not really know whether it is equivalent, in the presence of interactions, to the Instant Form quantization of fields. The apparent triviality of the vacuum is prima facia evidence that it is not so equivalent, at least not for modern field theories with complex vacuum structure (Aitchison 1985). But where and why does the equivalence break down, and could it be reinstated with some fixing? Can one even address the Casimir interaction problem in the context of null-plane quantization?[11] We don't know, and nobody seems to be studying the problem. On the other hand, several workers still use null-plane perturbation theory to do high-energy collision calcula-

[6] Ida and Yabuki (1976a; 1976b); Berestetsky and Terentev (1976); Terentev (1977); Thorn (1979*a*, 1976).

[7] Hagen and Yee (1976); Singh and Hagen (1977); Steinhardt (1979).

[8] Coleman (1977); Callan and Coleman (1977).

[9] Wess and Zumino (1974*a*, 1974*b*, 1974*c*); Wess (1976).

[10] Scherk and Schwartz (1974); Gliozzi *et al.* (1977); Green and Schwartz (1984, 1985).

[11] See Casimir (1948); Ambjørn and Wolfram (1983).

tions, and use null-plane quantization concepts to analyse hadrons as quark-bound states.[12] All of this is carried out with total neglect of the underlying vacuum structure problem. Is this a justifiable or even a consistent practice? We don't know!

In the following sections, I will present a survey of some of the concepts and technical ideas involved in null-plane quantization and its relation to Instant Form quantization. No pretence is made to either exhaustive coverage or rigour of treatment. The discussion should be adequate to give readers unfamiliar with null-plane quantization a feel for the issues involved in the following questions: Is null-plane quantized field theory physically equivalent to Instant Form quantized field theory in those cases where interactions are present (especially if non-perturbative methods must be employed to extract the consequences of the theory), and, if they are not equivalent, coud they be made equivalent by 'modest' adjustments? To illuminate these issues, I discuss the kinematics of null-plane coordinates in Section 2, free-field quantization for a scalar field in Section 3, and vacuum structure in Section 4. In Section 5 I briefly entertain two conjectures on how the apparent incompatibility of null-plane and Instant Form quantization might be removed by 'modest' adjustments.

2 Null-Plane Coordinates and Poincaré Transformations

A null-plane coordinate system is one in which the expression for the invariant interval, ΔS^2, between two space-time points, takes the off-diagonal form

$$\Delta S^2 = 2\Delta x^+ \Delta x^- - \Delta \bar{x}^2 \tag{1}$$

where $\Delta\bar{x}$ is a Euclidean two-vector. The coordinates x^+, x^-, and $\bar{x} \equiv (x^1, x^2)$ can always be related to some Minkowski coordinate system (inertial frame) by

$$x^\pm \equiv \frac{1}{\sqrt{2}}(x^0 \pm x^3), \qquad \bar{x} \equiv (x^1, x^2). \tag{2}$$

[12] Hornbostel, Brodsky and Pauli (1990).

The invariant form (1) corresponds to a metric tensor of the form

$$g^{\mu\nu} = \begin{array}{c|cccc} & + & - & 1 & 2 \\ \hline + & 0 & 1 & 0 & 0 \\ - & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & -1 & 0 \\ 2 & 0 & 0 & 0 & -1 \end{array} \tag{3}$$

which, in turn, yields the inner product

$$p^\mu x_\mu = p^0 x^0 - \mathbf{p}\cdot\mathbf{x} = p^+ x^- + p^- p^+ - \bar{\mathbf{p}}\cdot\bar{\mathbf{x}} \tag{4}$$

where $p^\pm$ are defined in the manner of (2).

Making the conventional choice to regard x^+ as the 'time' variable, we see that different values of x^+ designate different members of a family of parallel null planes.To apprehend the action of the Poincaré group transformations on these null-plane coordinates and the parallel null planes singled out by them, it is convenient to break down the Poincaré transformations into subsets. The general Poincaré transformation can be analysed into a composition of transformations from these subsets. We begin with members of the stability subgroup.

(i) Translations *in* the null plane:

$$x^{-\prime} = x^- + a^-, \qquad \bar{x}' = \bar{x} + \bar{a}, \qquad x^{+\prime} = x^+. \tag{5}$$

(ii) Rotation *in* the null plane:

$$x^{\pm\prime} = x^\pm, \qquad \bar{x}' = \cos\theta\bar{x} + \sin\theta\bar{x}_\perp; \tag{6a}$$

$$\bar{x}_\perp \equiv (x^2, -x^1). \tag{6b}$$

(iii) 'Galilean boost' *in* the null plane:

$$x^{-\prime} = x^- - \bar{U}^-\bar{x} + \frac{(\bar{U}^-)^2}{2} x^+, \qquad \bar{x}' = \bar{x} - \bar{U}^- x^+; \tag{7a}$$

$$x^{+\prime} = x^+. \tag{7b}$$

(iv) Dilatation:

$$ax^{\pm\prime} = (\lambda^\pm) x^\pm, \qquad \bar{x}' = \bar{x}. \tag{8}$$

All the transformations that can be formed from these sets by composition leave the $x^+=0$ null plane unchanged. Such transformations form the stability group of the $x^+=0$ null plane. This is a seven-parameter group. The restriction to $x^+=0$ is required only by the dilatations that would clearly alter a null plane with $x^+\neq 0$. As we will see in a moment, however, for *each* of the $x^+\neq 0$ null planes there is a transformation, different from but analogous to dilatation, which leaves *that* null plane unchanged. Continuing with the listing, we have the following.

(v) Translation *of* the null plane:

$$x^{-\prime}=x^-, \qquad \bar{x}'=\bar{x}, \qquad x^{+\prime}=x^+ + a^+. \tag{9}$$

(vi) 'Galilean boost' *of* the null plane:

$$x^{-\prime}=x^-, \qquad \bar{x}'=\bar{x}'=\bar{x}-\bar{U}^+x^-; \tag{10a}$$

$$x^{+\prime}=x^+ - \bar{U}^+\bar{x}+\frac{(\bar{U}^+)^2}{2}x^-. \tag{10b}$$

These last sets contain transfer motions that change all the x^+ null planes. We do note, as promised, that the combination of type (iv) and a type (v) transformation,

$$x^{-\prime}=\frac{1}{\lambda}x^-, \qquad \bar{x}'=\bar{x}, \qquad x^{+\prime}=\lambda x^+ + a^+, \tag{11}$$

leaves invariant the particular null plane designated by $x^+ = a^+/(1-\lambda)$. This transformation plays the role of dilatation in the stability subgroup for this null plane.

Several of the transformations have been referred to as 'Galilean'. This points to the similarity in their formal structure to Galilean-like boosts in a two-dimensional Euclidean space. This 'Galilean' structure was initially much prized by practitioners of null-plane dynamics, both for its intuitive appeal (one could use one's non-relativistic intuition in setting up calculations) and for the simplifications in dynamical structure it seemed to impose (see later sections). Eventually, however, it came to be seen as a two-edged sword, bringing with it, as it did, unpleasant singular behaviour in many perturbation theory calculations.

3 Null-Plane Quantization of Free Fields

The equivalence of null-plane and conventional, Instant Form, quantization for free fields will be illustrated by the simple example of a single self-adjoint scalar field. In the case of Instant Form quantization, the space–time field operator can be expressed in the form of a Fourier integral as

$$\hat{\phi}(\mathbf{x}, x^0)=(2\pi)^{-3/2}\int\frac{\mathrm{d}^3k}{2k^0}\left\{\hat{A}(\mathbf{k})\exp[i(\mathbf{k}\cdot\mathbf{x}-k^0x^0)] + \hat{A}^+(\mathbf{k})\exp[-i(\mathbf{k}\cdot\mathbf{x}-k^0x^0)]\right\} \qquad (12)$$

where the $\hat{A}^+(\mathbf{k})$ create free quanta with 4-momenta $k^\mu=(k^0, \mathbf{k})=[\sqrt{(\mathbf{k}^2+\kappa^2)}, \mathbf{k}]$ and satisfy

$$[\hat{A}(\mathbf{k}'), \hat{A}(\mathbf{k})]=0; \qquad [\hat{A}(\mathbf{k}'), \hat{A}^+(\mathbf{k})]=2k^0\delta^3(\mathbf{k}-\mathbf{k}'). \qquad (13)$$

These commutation rules in turn yield the canonical equal-time commutation rules for the space–time field,

$$[\hat{\phi}(\mathbf{x}', x^0), \hat{\phi}(\mathbf{x}, x^0)]=0; \qquad (14a)$$

$$[\hat{\phi}(\mathbf{x}', x^0), \partial_0\hat{\phi}(\mathbf{x}, x^0)]-\delta^3(\mathbf{x}=\mathbf{x}'). \qquad (14b)$$

We now attempt to simply rewrite these relations in terms of the coordinates and variables appropriate to null-plane quantization. Introducing

$$k^\pm=\frac{1}{\sqrt{2}}(k^0\pm k^3) \qquad (15a)$$

and

$$\bar{k}\equiv(k^1, k^2), \qquad (15b)$$

we note that

$$\frac{\mathrm{d}^3k}{2k^0}=\frac{\mathrm{d}^2\bar{k}\,\mathrm{d}k^+}{2k^+}, \qquad (16a)$$

$$2k^0\,\delta^3(\mathbf{k}-\mathbf{k}')=2k^+\,\delta(k^+-k^{+\prime})\,\delta^2(\bar{k}-\bar{k}'), \qquad (16b)$$

and

$$k^0x^0-\mathbf{k}\cdot\mathbf{x}=k^-x^++k^+x^--\bar{k}\cdot\bar{x}. \qquad (17)$$

If we add to this the recognition that creation or annihilation of quanta with four momenta, $\mathbf{k}, k^0 = \sqrt{(\mathbf{k}^2+\kappa^2)}$, is equivalent to creation or annihilation of quanta with null-plane momenta $\bar{k}$, k^+, $k^- = (\bar{k}^2+\kappa^2)/2k^+$, we can put

$$\hat{A}^+(\mathbf{k}) \equiv \hat{A}^+(\bar{k}, k^+) \tag{18}$$

and write

$$\begin{aligned}\hat{\phi}(\mathbf{x}, x^0) &\equiv \hat{\phi}(\mathbf{x}, x^-, x^+)\\ &= (2\pi)^{-3/2}\int \frac{\mathrm{d}^2\bar{k}\,\mathrm{d}k^+}{2k^+}\{\hat{A}(\bar{k}, k^+)\\ &\qquad \exp[i(\bar{k}\cdot\bar{x}-k^+x^- - k^-x^+)]\\ &\qquad +\hat{A}^+(\bar{k}, k^+)\exp[-i(\bar{k}\cdot\bar{x}-k^+x^- - k^-x^+)]\}\end{aligned} \tag{19}$$

where

$$[\hat{A}(\bar{k}', k^{+\prime}), \hat{A}(\bar{k}, k^+)] = 0 \tag{20a}$$

$$[\hat{A}(\bar{k}', k^{+\prime}), \hat{A}^+(\bar{k}, k^+)] = 2k^+\,\delta(k^+ - k^{+\prime})\,\delta^2(\bar{k}-\bar{k}') \tag{20b}$$

and

$$k^- = \frac{k^2+\kappa^2}{2k^+}. \tag{20c}$$

The basic space–time commutation relations are now to be evaluated at equal x^+ rather than at equal x^0, i.e. on members of the appropriate family of null planes rather than on instantaneous hyperplanes. Using (19) and (20), we find

$$[\hat{\phi}(\bar{x}', x^{-\prime}; x^+), \hat{\phi}(\bar{x}, x^-; x^+)] = \tfrac{1}{4}\varepsilon(x^- - x^{-\prime})\,\delta^2(\bar{x}-\bar{x}'), \tag{21a}$$

$$[\hat{\phi}(\bar{x}', x^{-\prime}; x^+), \partial_-\hat{\phi}(\bar{x}, x^-; x^+)] = \tfrac{1}{2}\,\delta(x^- - x^{-\prime})\,\delta^2(\bar{x}-\bar{x}'). \tag{21b}$$

As these last are the basic commutation relations postulated for null-plane quantized scalar fields, we see that in the case of free fields null-plane quantization is just a simple rewriting in null-plane coordinates of conventional, Instant Form quantization.

We are immediately moved to ask what it is that blocks this rewriting when interactions are present. In simplest terms, the answer is that in the presence of the interactions the quantities

$$\hat{A}(\mathbf{k})\exp(-ik^0x^0) \qquad \text{and} \qquad \hat{A}^+(\mathbf{k})\exp(ik^0x^0) \tag{22}$$

occurring in (12) become

$$\hat{A}(\mathbf{k}, x^0) \qquad \text{and} \qquad \hat{A}^+(\mathbf{k}, x^0), \tag{23}$$

while the quantities

$$\hat{A}(\bar{k}, k^+)\exp(i\bar{k}^- x^+) \qquad \text{and} \qquad \hat{A}^+(\bar{k}, k^+)\exp(ik^- x^+) \tag{24}$$

occurring in (19) become

$$\hat{A}(\bar{k}, k^{+\prime}, x^+) \qquad \text{and} \qquad \hat{A}^+(\bar{k}, k^{+\prime}, x^+) \tag{25}$$

It is now impossible to establish the connection between (23) and (25) prior to solving the non-trivial, nonlinear equations of motion for the interacting fields. If we *assume* equivalence of the two forms of quantization, then a temporally highly non-local relation between the two types of creation and annihilation operators can be written down. But in the absence of the assumption, we cannot test for the satisfaction of this relation without solving the equations of motion.

Returning now to (21) we note that the equal null-plane commutator for the field with itself is not zero (21a), while the commutator of the field with its canonical conjugate (21b), is obtained by explicit differentiation, *on* the null plane, of the field–field commutator. To confirm that the canonical conjugate to the field *is* just the x^- derivative of the field, and thus is a null-plane generalized functional of the field, we need only rewrite the free field Lagrangian density,

$$L(x) = \tfrac{1}{2}\{\partial^\mu \hat{\phi}(x)\, \partial_\mu \hat{\phi}(x) - \kappa^2 \hat{\phi}^2(x)\}, \tag{26a}$$

in null-plane coordinates,

$$L(x) = \tfrac{1}{2}\{2\partial_+ \hat{\phi}(\bar{x}, x^-, x^-; x^+)\, \partial_- \hat{\phi}(\bar{x}, x^-; x^+) \\ - \bar{\partial}\hat{\phi}(\bar{x}, x^{-\prime}; x^+)\, \bar{\partial}\hat{\phi}(\bar{x}, x^-; x^+) - \kappa^2 \hat{\phi}^2(\bar{x}, \bar{x}'; x^+)\}. \tag{26b}$$

This yields

$$\hat{\Pi}(\bar{x}, x^-; x^+) \equiv \frac{\partial L(\bar{x}, x^-; x^+)}{\partial[\partial_+ \hat{\phi}(\bar{x}, x^-; x^+)]} \\ = \partial_- \hat{\phi}(\bar{x}, x^-; x^+). \tag{27}$$

The factor of $\frac{1}{2}$ occurring in (21b) as well as the non-vanishing of

(21a), can be traced to the functional dependence of $\hat{\phi}(\bar{x}, x^{-}; x^{+})$ and $\hat{\Pi}(\bar{x}, x^{-}; x^{+})$, which renders the standard canonical commutation relations internally inconsistent on null planes.

Since the canonical conjugate to the field on a null plane is determined by the field itself on the same null plane, it appears that the field, over all the points of the null plane, comprises an irreducible set of operators. In other words, every operator in the state space of the field-theoretic system can be written as a null-plane generalized functional of the field itself. In particular, the null-time derivative of the field, $\partial_{+}\hat{\phi}(\bar{x}, x^{-}; x^{+})$, is not required in the functionals. The dynamical circumstance that permits this is the first-order character of the equations of motion for the field in null-plane coordinates. Thus, the wave equation

$$(\Box + \kappa^{2})\hat{\phi}(x) = 0 \tag{28a}$$

becomes

$$(2\partial_{+}\partial_{-} - \bar{\partial}^{2} + \kappa^{2})\hat{\phi}(\bar{x}, x^{-\prime}\, x^{+}) = 0. \tag{28b}$$

This equation can be integrated over x^{-} to yield

$$\partial_{+}\hat{\phi}(\bar{x}, x^{-}; x^{+}) = \frac{1}{4}\int dx^{-\prime}\varepsilon(x^{-} - x^{-\prime})\,(\bar{\partial}^{2} - \kappa^{2})\hat{\phi}(\bar{x}, x^{-\prime}; x^{+})$$

$$+ \text{arbitrary function of } \bar{x} \text{ and } x^{+}. \tag{29}$$

The arbitrary function is set equal to zero on the ground that a non-zero value would contribute infinite energy to the states of the system. With this elimination we see that $\partial_{+}\hat{\phi}(\bar{x}, x^{-}; x^{+})$ is itself a functional of the field over the x^{+} null plane, thus confirming the irreducibility of the null-plane field algebra.

4 Vacuum Structure on Null Planes

There are two intuitively natural ways to define a vacuum state for a quantum-field-theoretic system. One is that the vacuum is the state of lowest energy, which energy can conventionally be taken to be zero. The existence of such a state, necessarily an energy eigenstate, presumes the energy spectrum of the field system to have a lower bound. Such a lower bound seems required to prevent runaway collapse of the world modelled by the system. The second intuitive definition of the vacuum state is that it is the state devoid of field quanta, i.e. the state that is transformed to the null vector, 0, by application of *any* annihilation operator. Both of these definitions appeal to our conception of the vacuum as the state with nothing in it, the empty state. In the first definition it has no energy, while in the second it has no quanta, no particles. Unfortunately, in quantum theory, as we know, the concepts of energy and particles have become sufficiently subtle as to play havoc with our intuitive expectations. A third approach to the vacuum concept, closely related to the previous ones but not quite as immediately appealing, is through symmetry considerations. The vacuum state is that state which is invariant under all transformations of the space-time symmetry group of the theory. We expect the former definitions to yield this feature since, if the vacuum is empty, how can it appear differently to different inertial observers?

In practice, these concepts of the vacuum mesh well together in the absence of interactions, whether one employs Instant Form quantization or null-plane quantization. Upon turning on interactions, however, the null-plane quantization scheme *seems* to yield a distinct benefit in that the three conceptions of the vacuum state remain compatible. In the Front Form quantization scheme the presence of interactions renders the three conceptions incompatible. The state without field quanta is not even stationary, let alone free of energy and space–time-invariant, while the zero-energy, space–time-invariant state carries a complex array of field quanta. Let's see how this works.

In the absence of interactions, the normal-ordered Hamiltonian, expressed in terms of momentum eigenstate creation and annihilation operators, is

$$\begin{aligned}\hat{P}^0 &= \int \frac{\mathrm{d}^3p}{2p^0}\, p^0 \hat{A}^+(\mathbf{p})\hat{A}(\mathbf{p}) \\ &= \frac{1}{2}\int \mathrm{d}^3p \hat{A}^+(\mathbf{p})\hat{A}(\mathbf{p}) \qquad (30a)\end{aligned}$$

for the instant form, and

$$\hat{P} = \int \frac{\mathrm{d}p^+}{2p^+}\, \mathrm{d}^2\bar{p} \left(\frac{\bar{p}^2+\kappa^2}{2p^+}\right) \hat{A}^+(\bar{p}, p^+)\hat{A}(\bar{p}, p^+) \qquad (30b)$$

for the null-plane form. Similar expressions hold for all the generators of the Poincaré group of space–time symmetry transformations. For all these operators, when applied to a state vector, the first thing that hits the state vector is the annihilation operator, $\hat{A}(\mathbf{p})=\hat{A}(\bar{p}, p^{+})$. If the state in question contains no field quanta, it is transformed into the null vector and we have an eigenvector with zero eigenvalue of the energy, the null-plane energy, or whatever quantity is represented by the space–time generator we apply.

Now suppose we turn on interactions. For simplicity we will assume the interaction to be a normal-ordered $\lambda\hat{\phi}(x)^4$ term added to the Lagrangian density. This will add to the Instant Form energy operator the term

$$\lambda \int \mathrm{d}^3x\!:\!\hat{\phi}(\mathbf{x}, x^0)^4\!: \tag{31a}$$

and to the null-plane energy operator the term

$$\lambda \int \mathrm{d}^2\bar{x}\, \mathrm{d}x^{-}\!:\!\hat{\phi}(\bar{x}, x^{-}; x^{+})^4\!: \tag{31b}$$

where the colons signify normal-ordering of the fourth power of the field. When rewritten in terms of momentum eigenvector creation and annihilation operators, both of these expressions yield a term that does not contain any annihilation operators. For the Instant Form energy, we get the term

$$\begin{aligned}(2\pi)^{-6}\lambda \int &\frac{\mathrm{d}^3p_1}{2p_1^0}\frac{\mathrm{d}^3p_2}{2p_2^0}\frac{\mathrm{d}^3p_3}{2p_3^0}\frac{\mathrm{d}^3p_4}{2p_4^0}\,\delta^3(\mathbf{p}_1+\mathbf{p}_2+\mathbf{p}_3+\mathbf{p}_4)\\ &\hat{A}^{+}(\mathbf{p}_1, x^0)\hat{A}^{+}(\mathbf{p}_2, x^0)\hat{A}^{+}(\mathbf{p}_3, x^0)\hat{A}^{+}(\mathbf{p}_4, x^0),\end{aligned} \tag{32a}$$

while for the null-plane energy, we get the term

$$\begin{aligned}(2\pi)^{-6}\lambda \int &\frac{\tilde{\mathrm{d}}p_1}{2p_1^{+}}\frac{\tilde{\mathrm{d}}p_2}{2p_2^{+}}\frac{\tilde{\mathrm{d}}p_3}{2p_3^{+}}\frac{\tilde{\mathrm{d}}p_4}{2p_4^{+}}\,\delta(p_1^{+}+p_2^{+}+p_3^{+}+p_4^{+})\\ &\delta^2(\bar{p}_1+\bar{p}_2+\bar{p}_3+\bar{p}_4)\hat{A}^{+}(\bar{p}_1, p_1^{+}; x^{+})\hat{A}^{+}(\bar{p}_2, p_2^{+}; x^{+})\\ &\hat{A}^{+}(\bar{p}_3, p_3^{+}; x^{+})\hat{A}^{+}(\bar{p}_4, p_4^{+}; x^{+})\end{aligned} \tag{32b}$$

where $\tilde{\mathrm{d}}p\equiv \mathrm{d}p^{+}\,\mathrm{d}^2\bar{p}$.

Now, on the face of it, when these terms in the energy operators hit the quanta free state, they will transform it to a linear superposition of 4-quanta states. The quanta free state is no longer the zero-energy eigenstate, it would seem. This is true for the Instant Form, but there is a saving catch for the null-plane form. Unlike the other momentum variables, the variables $p_1^+, \ldots, p_4^+$ have a positive semi-definite spectrum, and the delta function, $\delta(p_1^+ + p_2^+ + p_3^+ + p_4^+)$, can yield a contribution only if $p_1^+ = p_2^+ = p_3^+ = p_4^+ = 0$. But this, remembering (20c), is the case in which each quanta contributes infinite energy to the system. It seems desirable on physical grounds, and is consistent with the remarks following (29) systematically to eliminate such states from the formalism. More rigorously, one can show that if $|\Omega; x^+)$ is the no-quanta state at null-time x^+ and if $|\psi)$ is any normalizable state such that all its n-quanta state functions are of arbitrary fast decrease at infinity in *all* components of the individual quanta 4-momenta, then

$$(\Psi|\hat{P}^-|\Omega; x^+) = 0. \tag{33a}$$

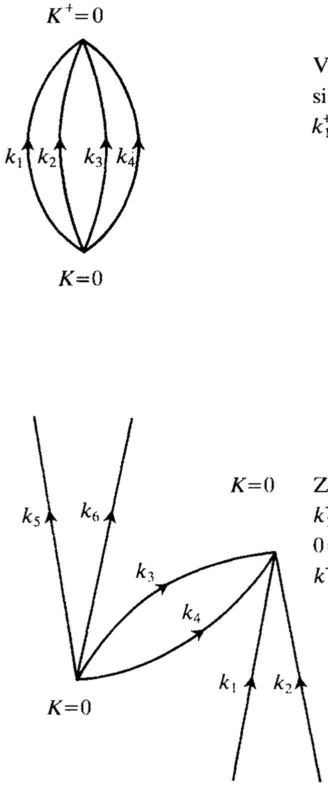
Since the $|\Psi)$ in question are dense in the state space, the demand that (33a) be a continuous antilinear functional of $|\Psi)$ yields

$$\hat{P}^-|\Omega; x^+) = 0, \tag{33b}$$

and the no-quanta state is the stationary vacuum.

This conclusion holds up when one examines the dynamical evolution of the system from the standpoint of null-plane perturbation theory. Employing what is sometimes called 'old-fashioned' or 'time-ordered' perturbation theory, but using null-time ordering rather than ordinary time ordering, we remember that, at each interaction (vertex) in the perturbative calculation, all the components of the total 3-momenta must be *rigorously* conserved. But in the null-plane formalism the $\hat{P}^+$ component has only a non-negative spectrum. Thus, no quanta can emerge from the vacuum ($p^+ = 0$) with non-vanishing p^+, not even temporarily as virtual states, and no set of quanta, with non-vanishing p^+, can disappear into the vacuum without violating p^+ conservation. In the language of perturbation theory diagrams, this means that all vacuum fluctuation diagrams and all so-called Z-graph diagrams are eliminated since they violate p^+ conservation at the vertices (see Fig. 3). The one loophole of quanta with all $p^+ = 0$ emerging from or annihilating into the vacuum is nicely dispatched by the fact that the corresponding $p^- = \infty$ yields infinite energy denominators in the calculation, thus killing the probability for the transition.

We mention in passing that our presumed choice to work with null-



Vacuum fluctuation diagram forbidden since $k_1^+ + k_2^+ + k_3^+ + k_4^+ = 0$ implies $k_1^+ = k_2^+ = k_3^+ = k_4^+ = 0$.

Z-graph diagram forbidden since $k_1^+ + k_2^+ + k_3^+ + k_4^+ = 0$, and $0 = k_3^+ + k_4^+ + k_5^+ + k_6^+$ implies $k_1^+ = k_2^+ = k_3^+ = k_4^+ = k_5^+ = k_6^+ = 0$.

FIG. 3 Null-time-ordered perturbation theory diagrams forbidden owing to conservation of non-negative total K^+ at each interaction vertex.

time-ordered perturbation theory is not much of a choice at all. In Instant Form perturbation theory the relationship between the time-ordered version and the manifestly covariant Feynman version is just that in the latter one calculates at once the contribution from all time-ordered diagrams related to one another by rearranging the time ordering of their vertices (accompanied by appropriate particle–antiparticle transformations). In most instances, those rearrangements yield *Z*-graph structure, which is forbidden in null-plane perturbation theory. Thus, in the null-plane formalism there is little difference between time-ordered perturbation theory and Feynman perturbation theory.

We now turn to a different aspect of vacuum structure, the relationship between global and local symmetries. In 1966, Coleman published a paper with the provocative title. 'The Invariance of the Vacuum is the Invariance of the World'. In that paper he showed that a generalized charge, which can be expressed as the volume integral of the time component of a local generalized 4-current density, can have the vacuum as a null eigenvector if and only if the generalized 4-current density is locally conserved. Having the vacuum as a null eigenvector means that the vacuum state would be invariant under the unitary group resulting from exponentiating the generalized charge, while having the generalized 4-current density conserved is the standard manifestation of an underlying continuous symmetry via Noether's theorem, at least in a Lagrangian formulation. This was a surprising result, since one intuitively expected the vacuum state to be empty of generalized charges whether a corresponding local symmetry was present in the theory or not. The consensus that seemed to emerge from this theorem was that, in the absence of the corresponding local symmetry, the generalized charges did not really exist as well defined operators (Orzalesi 1970). In this way one could avoid the awkward question of the value of such charges in the vacuum state. With the study of null-plane quantization, however, a second, welcome, surprise came in the recognition that Coleman's theorem did not hold in the null-plane case. Let's see how that works at the heuristic level.

First, in the Instant Form quantization scheme, let the generalized charge be

$$\hat{Q}_{x^0} \equiv \int \mathrm{d}^3 x \hat{y}^0(\mathbf{x}, x). \tag{34}$$

Then, if $|\Omega)$ is the vacuum state and we assume it is a null eigenvector, i.e. if

$$\hat{Q}_{x^0}|\Omega) = 0, \tag{35}$$

we have

$$\begin{aligned} 0 = \frac{\mathrm{d}}{\mathrm{d}x^0}(\hat{Q}_{x^0}|\Omega)) &\equiv \int \mathrm{d}^3 x \frac{\partial}{x^0} \hat{y}^0(\mathbf{x}, x^0)|\Omega) \\ &= \int \mathrm{d}^3 x \, \partial_\mu \hat{y}^\mu(\mathbf{x}, x^0)|\Omega), \end{aligned} \tag{36}$$

since

$$\int \mathrm{d}^3x \boldsymbol{\nabla}\cdot\hat{\mathbf{y}}(\mathbf{x}, x^0)|\Omega) = \oint_{\substack{\text{closed surface at}\\ \text{spacelike infinity}}} \mathrm{d}\boldsymbol{\sigma}\cdot\hat{\mathbf{y}}(\mathbf{x}, x^0)|\Omega) = 0. \tag{37}$$

For the 4-divergence of the 4-current, however, we have

$$\partial_\mu \hat{y}^\mu(x) \equiv \hat{\phi}(x) = \exp\left(\frac{i}{\hbar}\hat{P}^\mu x_\mu\right)\hat{\phi}(0)\exp\left(-\frac{i}{\hbar}\hat{P}^\mu x_\mu\right), \tag{38}$$

and since

$$\hat{P}^\mu|\Omega) = 0, \tag{39}$$

we find

$$\begin{aligned} 0 &= \int \mathrm{d}^3x \partial_\mu \hat{y}^\mu(\mathbf{x}, x^0)|\Omega) \\ &= \exp\left(\frac{i}{\hbar}\hat{P}^0 x^0\right)(2\pi\hbar)^3\delta^3(\mathbf{p})\hat{\phi}(0)|\Omega) \end{aligned} \tag{40a}$$

or

$$\delta^3(\mathbf{p})\hat{\phi}(0)|\Omega) = 0. \tag{40b}$$

Now let $\langle \mathbf{p}=0,\ p^0=Mc, \ldots |$ denote *any* null eigenbra of total 3-momentum with definite total energy $p^0c = Mc^2$. From (40b), it follows that

$$(\bar{\mathbf{p}}=0, p^0=Mc, \ldots |\hat{\phi}(0)|\Omega) = 0. \tag{41a}$$

Since both $|\Omega)$ and the scalar field $\hat{\phi}(0)$ are invariant under homogeneous Lorentz transformations, this last equation is equivalent to

$$(p^\mu, \ldots |\hat{\phi}(0)|\Omega) = 0 \tag{41b}$$

for a complete set of p^μ eigenbras. Therefore

$$\hat{\phi}(0)|\Omega) = 0. \tag{42}$$

But it is well known in Instant Form local field theory that a *local field* that has the vacuum state as a null eigenvector is itself zero. Thus,

$$\hat{\phi}(0) = 0 \tag{43}$$

and from (38),

$$\hat{\phi}(0)=\partial_\mu \hat{y}^\mu(x)=0.$$

In the null-plane quantization scheme, we start from

$$\hat{Q}_{x^+}\equiv\int dx^-\, d^2\bar{x}\hat{y}^+(\bar{x},x^-;x^+) \tag{44}$$

and

$$\hat{Q}_{x^+}|\Omega)=0, \tag{45}$$

which yields

$$0=\frac{d}{dx^+}(\hat{Q}_{x^+}|\Omega))=\int dx^-\, d^2\bar{x}\,\partial_+\hat{y}^+(\bar{x},x^-;x^+)|\Omega)$$

$$=\int dx^-\, d^2\bar{x}\partial_\mu\hat{y}^\mu(\bar{x},x^-;x^+)|\Omega). \tag{46}$$

Once again using (38) and (39), we find, this time,

$$\delta(p^+)\,\delta^2(\bar{p})\hat{\phi}(0)|\Omega)=0 \tag{47a}$$

or

$$(p^+=0,\bar{P}=0,p^-=\infty,\ldots|\hat{\phi}(0)|\Omega)=0 \tag{47b}$$

for any null eigenbras of $\hat{P}^+$ and $\hat{\bar{P}}$. But in this case Lorentz transformations cannot change the $p^+=0$, $p^-=\infty$ values. Thus we cannot generate a complete set of total 4-momentum eigenbras by Lorentz transformations, and thus we cannot obtain the vector equation (42). Coleman's theorem does not hold in null-plane quantization. The null-plane vacuum can be *empty* of generalized charge in the *absence* of a corresponding local conservation law and associated symmetry.

5 Conjectures on Equivalence

Let us first consider the Casimir effect in quantum electrodynamics or, more generally, Casimir-like effects in general quantum field theories (see Casimir 1948). This shall be identified here as the problem posed by the introduction, into an otherwise vacuum state of a quantized field, of pieces of macroscopic matter carrying mobile sources of the quantized field (charged particles) which are constrained to remain in the pieces of matter. One calculates the geometry-dependent change of energy of the system relative to the undisturbed vacuum state, and from this, the forces exerted on the macroscopic matter. The standard approach, in Instant Form quantized field theory, is to treat the macroscopic matter as simply changing the boundary conditions on the vacuum fluctuations of the quantized field and thereby changing the energy of the system.

This will not do in null-plane quantized field theory, as there are no vacuum fluctuations to modify. It would seem that there, a more detailed account of the coupling between the matter-bound sources and the quantized field is required to reproduce the Casimir-like effects. *If* such a treatment were to succeed, it would suggest that the standard boundary condition approach in the Instant Form is misleadingly glib and simplistic. It would be interesting to resolve this question.

Finally, is it at all possible for null-plane quantized field theory to be equivalent to Instant Form quantized field theory for non-Abelian gauge fields with non-trivial vacuum structure of the degenerate or multiple-phase sort found in Quantum-Chromodynamics, Electroweak, and Grand Unification theories? On the face of it, our account of the vacuum in null-plane field theory would suggest that the equivalence was not possible. The null-plane vacuum seems inherently trivial. We should remember, however, that in order to determine a unique solution to the field equation for a free scalar field (and these considerations would continue to apply to all other cases in null-plane field quantization), we discarded an infinite family of solutions on the ground that they would contribute infinite energy to the system. If one could consistently associate these infinite energies with something like the zero-point energies of distinct vacuum phases, one might be able to employ the discarded solutions to model in null-plane quantization the vacuum complexity we see in Instant Form quantization. In any case, one should not yet regard the issue of the relationship between the distinct forms of field quantization, discovered long ago by Dirac, to be closed.

References

Aitchison, I. J. R. (1985), 'Nothing's plenty; The vacuum in modern quantum field theory', *Contemp. Phys.* 26: 333–91.

Ambjorn, J. and Wolfram, S. (1983*a*), 'Properties of the Vacuum. 1. Mechanical and Thermodynamic', *Ann. Phys.* 147: 1–32.

—— (1983), 'Properties of the Vacuum. 2. Electrodynamic', *Ann. Phys.* 147: 33–56.

Bardacki, K. and Halpern, M. B. (1968), 'Theories at Infinite Momentum', *Phys. Rev.* 176: 1686–99.

Bart, G. R. and Fenster, S. (1977), 'Free field theories of spin-mass trajectories and quantum electrodynamics in the null plane', *Phys. Rev.* D16: 3354–64.

Berestetsky, V. B. and Terentev, M. V. (1976), 'Light Front Dynamics and Nucleons Consisting of Relativistic Quarks', *Sov. J. Nucl. R.* 24(5): 547–54.

—— (1977), 'Nucleon Form Factors and Light Front Dynamics', *Sov. J. Nucl. R.* 25(3): 347–54.

Callan, C. G. jun. and Coleman, S. (1977), 'Fate of the False Vacuum. II. First Quantum Corrections', *Phys. Rev.* D16: 1762–8.

Casher, A. (1976), 'Gauge Fields on the Null Plane', *Phys. Rev.* D14: 452–64.

Casimir, H. B. G. (1948), 'On the attraction between two perfectly conducting plates', *Koninkl. Ned. Akad. Wetenschap. Proc.* 51: 793–6.

Chang, S. J. and Yan, T. M. (1973), 'Quantum Field Theories in the Infinite Momentum Frame. II. Scattering Matrices of Scalar and Dirac Fields', *Phys. Rev.* D7: 1147–61.

——, Root, R. G., and Yan, T. M. (1973), 'Quantum Field Theories in the Infinite Momentum Frame. I. Quantization of Scalar and Dirac Fields', *Phys. Rev.* D7: 1133–46.

Coleman, S. (1966), 'The Invariance of the Vacuum is the Invariance of the World', *J. Math. Phys.* 7: 787.

—— (1977), 'Fate of the False Vacuum: Semiclassical Theory', *Phys. Rev.* D15: 2929–36.

Dirac, P. A. M. (1949), 'Forms of Relativistic Dynamics', *Rev. Mod. Phys.* 21: 392–9.

—— (1950), 'Generalized Hamiltonian Dynamics', *Can. J. Math.* 2: 129–48.

di Sessa, A. (1974), 'Quantization on hyperboloids and full space-time field expansion', *J. Math. Phys.* 15: 1892–1900.

Driessler, W. (1975), 'Comments on Lightlike Translations and Applications in Relativistic Quantum Field Theory', *Comm. Math. Phys.* 44: 133–41.

—— (1977*a*), 'On the Structure of Fields and Algebras on Null Planes. I. Local Algebras', *Acta. Phys. Austriaca* 46: 63–96.

—— (1977*b*), 'On the Structure of Fields and Algebras on Null Planes. II. Field Structure', *Acta. Phys. Austriaca* 46: 163–96.

Fleming, G. N. (1966), 'A Manifestly Covariant Description of Arbitrary Dynamical Variables in Relativistic Quantum Mechanics', *J. Math. Phys.* 7: 1959–81.

—— (1985), 'Toward a Lorentz Invariant Quantum Theory of Measurement', *Minicourse in Fundamental Physics*, Univ. of Puerto Rico Press.

—— (1988), 'Hyperplane Dependent Quantized Fields and Lorentz Invariance', in R. Harré and H. Brown (eds.), *Philosophical Foundations of Quantum Field Theory*, Oxford University Press: 93–115.

—— with Bennett, H. (1989), 'Hyperplane Dependence in Relativistic Quantum Mechanics', *Found. Phys.* 19: 231–67.

Fock, V. A. (1932), 'Konfigurationsraum und zweite Quantelung', *Z. Physik* 75: 622–47.

—— (1934), 'Electrodynamics of Quantum Theory', *Phys. Zeit. Sov.* 6: 425–69.

Fubini, S. and Furlan, G. (1965), 'Renormalization Effects for Partially Conserved Currents', *Phys.* 1: 229–47.

——, Hansen, A. J., and Jackiw, R. (1973), 'New Approach to Field Theory', *Phys. Rev.* D7: 1732–60.

Gal-Ezer, E. and Horwitz, L. P. (1976), 'Charges as Null Plane Integrals Over Tensor Densities', *Lett. Math. Phys.* 1: 147–54.

Gliozzi, F., Scherk, J., and Olive, D. (1977), 'Supersymmetry, supergravity theories and the dual spinor model', *Nucl. Phys.* B122: 253–90.

Green, M. B. and Schwartz, J. H. (1984), 'Anomaly cancellations in supersymmetric $D = 10$ gauge theory and superstring theory', *Phys. Lett.* 149B: 117–22.

—— (1985), 'Infinite cancellations in SO(32) superstring theory', *Phys. Lett.* 151B: 21–5.

Gromes, D., Rothe, H. J., and Stech, B. (1974), 'Field quantization on the surface $x^2 = \text{constant}$', *Nucl. Phys.* B75: 313–32.

Haag, R. (1955), 'On Quantum Field Theories', *Dan. Mat. Fus. Med.* 29(12): 1–37.

Hagen, C. R. and Yee, J. H. (1976), 'Light Cone Quantization: Study of a Soluble Model with q-number Anticommutator', *Phys. Rev.* D13: 2789–804.

Hornbostel, K., Brodsky, S. J., and Pauli, H. C. (1990), 'Light Cone Quantized QCD in $1+1$ Dimensions[1], *Phys. Rev.* D41: 3814–21.

Ida, M. and Yabuki, H. (1976*a*), 'The Kinematical Structure of Lightlike Wave Functions', *Prog. Theor. Phys.* 55: 1606–18.

Ida, M. and Yabuki, H. (1976*b*), 'The Kinematical Structure of Lightlike Wave Functions. II', *Prog. Theor. Phys.* 56: 297–310.

Kogut, J. and Soper, D. (1970), 'Quantum Electrodynamics in the Infinite Momentum Frame', *Phys. Rev.* D1: 2901–14.

Leutwyler, H., Klauder, J. R., and Streit, L. (1970), 'Quantum Field Theory on Lightlike Slabs', *Nuovo Cim.* 66A: 536–54.

Orzalesi, C. A. (1970), 'Charges and Generators of Symmetry Transformations in Quantum Field Theory', *Rev. Mod. Phys.* 42: 381–408.

Scherk, J. and Schwartz, J. H. (1974), 'Dual models for non-hadrons', *Nucl. Phys.* B81: 118–44.

Singh, L. P. S. and Hagen, C. R. (1977), 'Interacting Rarita–Schwinger Field on the light front', *Phys. Rev.* D16: 347–53.

Sommerfield, C. M. (1974), 'Quantization on Spacetime Hyperboloids', *Ann. Phys.* 84: 285–302.

Steinhardt, P. J. (1979), 'Problems of Quantization in the Infinite Momentum Frame', Harvard Univ. preprint, HUTP-79/A033.

Susskind, L. (1968), 'Model of Self Induced Strong Interactions', *Phys. Rev.* 165: 1535–46.

Suzuki, T., Tameike, S., and Yamada, E. (1976), 'Some Remarks on Null Plane Quantization', *Prog. Theor. Phys.* 55: 922–8.

Terentev, M. V. (1976), 'On the structure of the wave functions of mesons considered as bound states of relativistic quarks', *Sov. J. Nucl. Phys.* 24: 106–9.

Thorn, C. B. (1979*a*), 'Quark confinement in the infinite momentum frame', *Phys. Rev.* D19: 639–51.

—— (1979*b*), 'Asymptotic freedom in the infinite momentum frame', *Phys. Rev.*, D20: 1934–40.

Weinberg, S. (1966), 'Dynamics at Infinite Momentum', *Phys. Rev.* 150: 1313–18.

Wess, J. (1976), 'Supersymmetry', *Acta. Phys. Austriaca* Suppl., 15: 475–98.

—— and Zumino, B. (1974*a*), 'Supergauge transformations in four dimensions', *Nucl. Phys.* B70: 39–50.

—— (1974*b*), 'A Lagrangian world invariant under supergauge transformations', *Phys. Lett.* 49B: 52–4.

—— (1974*c*), 'Supergauge invariant extension of quantum electrodynamics', *Nucl. Phys.* B78: 1–13.

Wightman, A. S. (1960), 'L'Invariance dans la Mecanique Quantique Relativiste', in C. de Witt and R. Omnes (eds.), *Dispersion Relations and Elementary Particles* (John Wiley): 159–226.

Wigner, E. P. (1939), 'On Unitary Representations of the Inhomogeneous Lorentz Group', *Ann. Math.* 40: 149–204.

Yan, T. M. (1973*a*), 'Quantum Field Theories in the Infinite Momentum

Frame. III. Quantization of Coupled Spin-One Fields', *Phys. Rev.* D7: 1760–80.

—— (1973*b*), 'Quantum Field Theories in the Infinite Momentum Frame. IV. Scattering Matrix of Vector and Dirac Fields and Perturbation Theory', *Phys. Rev.* D7: 1780–1800.

6

The Physical Significance of the Vacuum State of a Quantum Field

D. W. SCIAMA

1 Introduction

Even in its ground state, a quantum system possesses fluctuations and an associated zero-point energy, since otherwise the uncertainty principle would be violated. In particular, the vacuum state of a quantum field has these properties. For example, the electric and magnetic fields in the electromagnetic vacuum are fluctuating quantities. This leads to a kind of reintroduction of the ether, since some physical systems interacting with the vacuum can detect the existence of its fluctuations. However, this ether is Lorentz-invariant, so there is no contradiction with special relativity. The aim of this paper is to discuss the physical significance of these zero-point fluctuations and to comment on the zero-point energy.

This discussion is not straightforward because we meet at the outset two fundamental unsolved problems. Despite this handicap, we can make considerable progress in understanding the subject, because provisional calculations which have been constructed to sidestep these problems lead to numerical results for various effects which are in remarkable agreement with experiment. Presumably, then, some significance attaches to these calculational procedures, although a full understanding is still lacking.

The problems are as follows.

1 The zero-point energy of a quantum field in Minkowski space–time is infinite.

This article is a revised version of material to be found in 'Black Holes and Fluctuations of Quantum Particles: An Einstein Synthesis', in Francesco de Finis and Giunti Barbera (eds), *Relativity, Quanta and Cosmology*, 1979; reprinted London, 1979.

2 This zero-point energy can produce no gravitational field if the Minkowski space–time is to be the consistent background geometry. Moreover, any actual field so produced (perhaps associated with a cosmological term in Einstein's field equations of general relativity) is known from experiment to be extremely small on any scale except possibly a cosmological one.

Problem 1 is often solved by arguing that, gravitation apart, the zero of energy can be changed by a fixed amount without altering any physically significant quantities. One may therefore subtract out the zero-point energy of a quantum field in Minkowski space–time, thereby reducing it to the more comfortable value of zero. Technically this is often done by normal-ordering the field operators. This process does not remove the field fluctuations, which, as I have said, are physically measurable.

Problem 2 remains unsolved. The most attractive proposal made so far uses the fact that the zero-point energy of a fermion field is negative, while that of a boson field is positive. Perhaps the fermion and boson fields that actually exist in nature are such that their zero-point energies exactly cancel out. This would be true, for example, if supersymmetry were an exact symmetry of nature, since the supersymmetry relates fermion fields to boson fields in just the required way. Unfortunately, we know from observation that supersymmetry can be at most a broken symmetry. (For example, we know that there does not exist a spin zero partner of the electron with the same mass as the electron itself. Any such electron must have a mass exceeding 80 GeV.) No one has been able to show that the breaking of supersymmetry could leave intact the cancellation of the zero-point energies.

A further complication arises from the fact that the physical effects produced by zero-point fluctuations of the field can equally well be attributed to other components of the physical system, owing to the various couplings that exist between the components. This is a very familiar situation in physics which has caused some confusion in the literature on our topic, although it is now well understood. To give an elementary example of what is involved, we can consider the classical electrostatic interaction between two charges. This interaction can be attributed either to a Coulomb-type long-range effect directly from one charge to the other, or as arising from a Lorentz force exerted by the field of one charge and evaluated at the position of the other charge. These pictures are related by elementary algebra involving the equations for the coupling between the various degrees of freedom, and one cannot say that only one of them leads to a 'physically real' representation. This may seem to be an obvious point, but we will see later that one must be careful about it in connection with zero-point fluctuations.

In attempting to understand this subject, I have found it helpful to follow an historical approach, particularly because the idea of zero-point energy in quantum theory was originally proposed fourteen years before it was shown to be a property of a harmonic oscillator in the quantum mechanics of 1925. Numerous attempts to detect zero-point energy differences in different systems were also made in those early years, and, as we shall see, the first reliable positive measurement was made only a few months after the discovery of quantum mechanics, and constituted the first experimental test of that revolutionary theory.

2 The Zero-Point Energy of a Harmonic Oscillator

Zero-point energy was introduced into physics by Max Planck in 1911. In a renewed attempt to understand the interaction between matter and radiation and its relation to the black body radiation spectrum, Planck put forward the hypothesis that the absorption of radiation occurs continuously while its emission is discrete and in energy units of $h\nu$. On this hypothesis, the average energy $\bar{\epsilon}$ of a harmonic oscillator at temperature T would be given by

$$\bar{\epsilon} = \frac{1}{2} h\nu + \frac{h\nu}{e^{h\nu/kT} - 1}.$$

Thus the oscillator would have an energy $\frac{1}{2}h\nu$ even at the absolute zero of temperature.

Planck abandoned his hypothesis in 1914 when Fokker showed that an assembly of rotating dipoles interacting classically with electromagnetic radiation would possess statistical properties (such as specific heat) in conflict with observation. Planck then became convinced that no classical discussion could lead in a satisfactory manner to a derivation of his distribution law for black body radiation. Nevertheless the idea of zero-point motion for a quantum harmonic oscillator intrigued physicists, and the possibility was much discussed before its existence was definitely shown in 1925 to be required by quantum mechanics, as a direct consequence of Heisenberg's uncertainty principle.

The second argument is more relevant to the subject matter of this paper. It involves the influence of zero-point motion on the Einstein–Hopf (1910) derivation of the black body radiation spectrum. That derivation involved a study of the interchange of energy and momentum between a harmonic oscillator and the radiation field. It led to the Rayleigh–Jeans distribution rather than to Planck's because of the classical assumptions which were originally used for the harmonic oscillators and the radiation field. In the Einstein–Hopf calculation the mean square momentum of an oscillator was found to be proportional to the mean energy of the oscillator and to the mean energy density of the radiation field. Einstein and Stern now included the zero-point energy of the oscillator in its mean energy, in the hope of deriving the Planck rather than the Rayleigh–Jeans distribution, but found that they could do so only if they took $h\nu$ rather than $\frac{1}{2}h\nu$ for the zero-point energy.

The reason for this difficulty is their neglect of the zero-point energy of the radiation field itself. It is not a consistent procedure to link together two physical systems only one of which possesses zero-point fluctuations in a steady state. These fluctuations would simply drive zero-point fluctuations in the other system or be damped out, depending on the relative number of degrees of freedom in the two systems. This point is fundamental to our later considerations. I merely note it here, and return to it in detail later on. Parenthetically, I may add that it was Nernst who, in 1916, first proposed that 'empty' space was everywhere filled with zero-point electromagnetic radiation.

3 The Discussion of Einstein and Stern (1913)

We now know that the main considerations of Einstein and Stern are wrong, and we mention them here purely for their historical interest. However, one correct remark that they made will be of importance for us later on. When one takes the classical limit $kT \gg h\nu$ for the Planck distribution, one finds

$$\frac{h\nu}{e^{h\nu/kT}-1} \rightarrow kT - \frac{1}{2}h\nu + h\nu \times O\left(\frac{h\nu}{kT}\right),$$

whereas

$$\frac{1}{2}h\nu + \frac{h\nu}{e^{h\nu/kT}-1} \rightarrow kT + h\nu \times O\left(\frac{h\nu}{kT}\right).$$

In a sense, then, the correct classical limit is obtained only if the zero-point energy is included.

Einstein and Stern gave two independent arguments in favour of retaining the zero-point energy, one involving rotational specific heats of molecular gases and the other, the derivation of the Planck distribution itself. In the first argument it was assumed that a freely rotating molecule would possess a zero-point energy. However, we now know that this is not the case according to quantum mechanics.

4 The Experimental Verification of Zero-Point Energy

For a single harmonic oscillator, the existence of a zero-point energy would not change the spacing between the various oscillator levels and so would not show up in the energy spectrum. It might, of course, alter the gravitational field produced by the oscillator. This is a difficult question which we shall touch on later. For the experimental physicists influenced by Planck and by Einstein and Stern, demonstrating the existence of zero-point energy amounted to finding a system in which the *difference* in this energy for different parts of the system could be measured. It was soon realized that a convenient way to do this was to look for isotope effects in the vibrational spectra of molecules. The small change in mass associated with an isotopic replacement would lead to a small change in vibrational frequency and so to a small change in the zero-point energy, and in the energies of all the vibrational levels. These changes might then show up in the vibrational spectrum of the system.

Many attempts were made to find this effect, but the first conclusive experiment was made by Mulliken in 1925, in his studies of the band spectrum of boron monoxide. This demonstration (made only a few months before Heisenberg (1925) first derived the zero-point energy for a harmonic oscillator from his new matrix mechanics) provided the first experimental verification of the new quantum theory.

Since then, zero-point effects have become commonplace in quantum physics, for example in spectroscopy, in chemical reactions, and in solid-state physics. Perhaps the most dramatic example is their role in maintaining helium in the liquid state under its own vapour pressure at absolute zero. The zero-point motion of the atoms keep them sufficiently far apart on average so that the attractive forces between them are too weak to cause solidification. We can express this in a rough way by defining an effective temperature T_{eff} such that kT_{eff} is equal to the zero-point energy per atom. For helium T_{eff} exceeds a classical estimate of the melting point under its own vapour pressure. Thus, even close to absolute zero helium is 'hot' enough to be liquid.

5 Dynamic Effects of Zero-Point Energy

Usually the boundary conditions associated with a physical system limit the range of normal modes that contribute to the ground state of the system and so to the zero-point energy. A trivial example is a harmonic oscillator, which corresponds to a single normal mode of frequency v and so to a zero point energy $\frac{1}{2}hv$. In more complicated cases the range of normal modes may depend on the configuration of the system. This would lead to a dependence of the ground-state energy on the variables defining the configuration and so, by the principle of virtual work, to the presence of an associated set of forces. One important example of such a force is the homopolar binding between two hydrogen atoms when their electron spins are antiparallel (Hellmann 1927). When the protons are close together, each electron can occupy the volume around either proton. The resulting increase in the uncertainty of the electron's position leads to a decrease in the uncertainty of its momentum and so to a *decrease* in its zero-point energy. Thus, there is a binding energy associated with this diatomic configuration, and the resulting attractive force is responsible for the formation of the hydrogen molecule. By contrast, when the electron spins are parallel, the Pauli exclusion principle operates to limit the volume accessible to each electron. In this case the effective force is repulsive.

More relevant to the present paper is the force that arises when some of the normal modes of a zero rest-mass field such as the electromagnetic field are excluded by boundary conditions. This is the famous Casimir (1948) effect. If one places in Minkowski space–time two parallel flat perfect conductors, then the boundary conditions on the conductors ensure that normal modes whose wavelength exceeds the spacing of the conductors are excluded. If now the conductors are moved slightly apart, new normal modes are permitted and the zero-point energy is increased. Work must be done to achieve this energy increase, and so there must be an attractive force between the plates. This force has been measured, and the zero-point calculation verified. This agreement with experiment is important, since it shows that calculations with the zero-point energy of a continuous field does have some correspondence with reality, although the total energy associated with modes of arbitrarily high frequency is infinite. The Casimir effect show that *finite differences between different configurations of infinite energy do have physical reality*. Another example of this principle for zero-point effects is the Lamb shift, which we shall consider shortly.

6 Zero-Point Noise and Damping

We shall need to understand the properties of zero-point fluctuations when the associated density of states is large, for example in a continuous field. In such a case the zero-point fluctuations can be an important source of *noise* and *damping*, these two phenomena being related by a fluctuation-dissipation theorem. We shall now consider various examples of these zero-point effects.

X-ray Scattering by Solids

The noise associated with zero-point fluctuations was discovered in 1914 by Debye during his study of X-ray scattering by solids. His main concern was to calculate the influence of the thermal vibrations of the lattice on the X-ray scattering, but he showed in addition that, if one assumes with Planck that the harmonic oscillators representing these vibrations have a zero-point motion, then there would be an additional scattering which would persist at the absolute zero of temperature. This additional scattering is associated with the emission of phonons by the X-rays, this emission being induced by the zero-point fluctuations. Thus, noise and damping are related together as in Einstein's theory of Brownian motion. Here we have the first example of a 'spontaneous' radiation process, which can be regarded as being induced by the coupling of the radiating system to a field of zero-point fluctuations. We shall meet other examples of this process later.

Einstein Fluctuations in Black Body Radiation

We note in passing here that the interference between the zero-point and thermal fluctuations of the electromagnetic field gives a characteristic contribution to the Einstein fluctuations of the energy in a black body radiation field, namely the term 'linear' in the energy density. This term would be absent for a classical radiation field pictured as an assembly of waves.

The Lamb Shift

An electron, whether bound or free, is always subject to the stochastic forces produced by the zero-point fluctuations of the electromagnetic field, and as a result executes a Brownian motion. The kinetic energy associated with this motion is infinite, because of the infinite energy in the high-frequency components of the zero-point fluctuations. This infinity in the kinetic energy can be removed by renormalizing the mass of the electron (Weisskopf 1949). As with the Casimir effect, physical significance can be given to this process in situations where one is dealing with different states of the system for which the difference in the total renormalized Brownian energy (kinetic plus potential) is finite. An example of this situation is the famous Lamb shift between the energies of s and p electrons in the hydrogen atom; according to Dirac theory, the energy levels should be degenerate. Welton (1948) pointed out that a large part of this shift can be attributed to the effects of the induced Brownian motion of the electron, which alters the mean Coulomb potential energy. This change in electron energy is itself different for an s and p electron, and so the Dirac degeneracy is split. This theoretical effect has been well verified by the observations.

One can also regard the Lamb shift as the change in zero-point energy arising from the dielectric effect of introducing a dilute distribution of hydrogen atoms into the vacuum. The frequency of each mode is simply modified by a refractive index factor (Feynman 1961; Power 1966; Barton 1970).

The Vacuum as a Dissipative System

A physical system containing a large number of closely spaced modes behaves as a dissipator of energy, as well as possessing fluctuations associated with the presence of those modes (the fluctuation-dissipation theorem). Now the vacuum states of the electromagnetic field constitutes an example of a system with closely spaced energy levels, there being $(8\pi/c^3)\,\nu^2\,\mathrm{d}\nu$ modes with frequencies lying between ν and $\nu+\mathrm{d}\nu$. We would expect this state also to possess a dissipative character related to the zero-point fluctuations; indeed, Callen and Welton (1951) proved this in their quantum mechanical derivation of the fluctuation-dissipation theorem.

In this derivation the dissipation is represented simply by the *absorption* of energy by the dissipative system. This can be justified by the expectation that the energy absorbed is divided up among so many modes that we may neglect the possibility that it is later re-emitted into its original modes with its initial phase relations intact. The absorption rate is calculated by second-order perturbation theory and is found to depend quadratically on the external force acting on the system. This enables an impedance function to be defined in the usual way. This function also appears in a linear relation between the external force acting on the system and the response of the system to this force. A familiar example would be the relation between an impressed electric field, the resulting current flow, and the electrical resistance of the system. The rate of energy absorption (and hence the resistance) clearly depends on the coupling between the external disturbance and the system, and on the density of states in the system.

Callen and Welton then proceeded to calculate the quantum fluctuations of the system in its unperturbed state. These fluctuations also depend on the density of states. The procedure is quite general, but for simplicity they restricted themselves to the case where the system is in thermal equilibrium at temperature T, so that its states are occupied in accordance with the Boltzmann distribution $\mathrm{e}^{-h\nu/kT}$. By eliminating the density of states, the following relation is obtained between the mean square force $\langle V^2\rangle$ associated with the fluctuations and the frequency-dependent impedance function, $R(\nu)$:

$$\langle V^2\rangle = \frac{2}{\pi}\int_0^\infty R(\nu)E(\nu, T)\,\mathrm{d}\nu,$$

where

$$E(\nu, T) = \frac{1}{2}\,h\nu + \frac{h\nu}{e^{h\nu/kT} - 1},$$

which is the mean energy of a harmonic oscillator at temperature T. The presence of the zero-point energy $\frac{1}{2}h\nu$ shows that the zero-point fluctuations (as well as the thermal fluctuations) are a source of noise power and damping, and therefore satisfy the fluctuation-dissipation theorem.

This theorem can be interpreted in three ways.

1 It shows how the damping rate R is determined by the equilibrium fluctuations $\langle V^2 \rangle$ (Nyquist relation).

2 It shows how the noise *power* $\langle V \rangle^2$ is determined by the absorption coefficient R (Kirchhoff relation).

3 It shows how the damping rate is proportional to the density of states in the dissipative system. In fact, the mean square force $\langle V^2 \rangle$ is proportional to the mean square electric field, which in turn is proportional to the energy in the dissipative system. This energy is given by the Planck distribution plus the zero-point contribution. This is the content of the fluctuation-dissipation theorem when $R(\nu)$ is regarded as proportional to the density of states, that is as proportional to $\nu^2\,\mathrm{d}\nu$. When the dissipative system is the vacuum ($T=0$), this remains true. Thus, the impedance of the vacuum is proportional to ν^2.

Radiation Damping

An important illustration of these ideas was given by Callen and Welton, who showed that the radiation damping of an accelerated charge could be interpreted in this way. The non-relativistic form of this damping force is

$$\frac{2}{3}\frac{e^2}{c^3}\frac{\mathrm{d}^2 v}{\mathrm{d}t^2},$$

and if the charge is oscillating sinusoidally with frequency ν, then the associated impedance function turns out to be

$$\frac{2}{3}\frac{e^2}{c^3}\,\nu^2.$$

This is precisely of the expected form (that is, proportional to ν^2, the impedance of the vacuum). Thus, the dependence of the damping force on $\mathrm{d}^2v/\mathrm{d}t^2$ is simply a manifestation of the density of states for the vacuum electromagnetic field.

A similar analysis can be carried out for the radiation damping of a perfect conductor accelerating through Minkowski space–time. The radiation emitted by such a conductor could be computed in terms of the currents that must flow to guarantee that the conductor is perfect. However, a more elegant method, and one that is more in harmony with the approach taken in this paper, is to adapt the Casimir effect to a moving boundary. This method is based on the requirement that the zero-point fluctuations must vanish on the boundary. This requirement leads to a finite change in the zero-point energy, which can be extracted from the infinite total energy by special techniques. The two-dimensional problem can be solved exactly by de Witt's point-splitting method (de Witt 1975, Fulling and Davies 1975), and one again finds a radiation damping force which in the non-relativistic limit is proportional to $\mathrm{d}^2v/\mathrm{d}t^2$.

The Zero-Point Noise of an Electric Circuit

As another illustration of these ideas, I can mention Weber's (1954) discussion of the zero-point noise of an electric circuit. One can quantize such a circuit and show that its zero-point fluctuations represent a measurable source of noise. For example, a beam of electrons passing near such a circuit would develop a noise component from this source. Associated with this noise is a damping of the electron beam caused by 'spontaneous' emission by the beam into the circuit. The zero-point noise is thus physically real, and is unaffected by a change in the origin of energy which might be introduced to remove in a formal way the infinite total energy associated with the zero-point fluctuations of arbitrarily high frequency.

The Damping of Zero-Point Fluctuations

The work of Callen and Welton and of Weber was generalized by Senitzky (1960), who studied the damping of a quantum harmonic oscillator coupled to a loss mechanism (reservoir) idealized as a system whose Hamiltonian is unspecified but which possesses a large number of closely spaced energy levels. Senitzky assumed that the coupling was switched on at time t_0, so that for $t < t_0$ the harmonic oscillator executed its zero-point motion undisturbed by the reservoir. After t_0, however, the fluctuating forces exerted by the reservoir would damp out the zero-point motion of the oscillator by a now familiar mechanism. To avoid a conflict with the Heisenberg uncertainty principle, we require that the zero-point fluctuations of the reservoir also introduce sufficient noise into the oscillator to restore its zero-point motion to the full quantum mechanical value $\frac{1}{2}h\nu$. Senitzky showed that after a few damping times the zero-point motion of the oscillator would be effectively driven by the zero-point motions of the reservoir. This is a beautiful quantum example of the relation between noise, damping, and equilibrium which Einstein discovered in 1905.

We now apply these ideas to the interaction of a two-level atom with the vacuum electromagnetic field.

If the atom is initially in an excited state, we expect it to make a transition to the ground state, at the same time emitting one or more quanta of radiation. The question arises whether this transition is induced (stimulated) by the zero-point fluctuations of the vacuum or is a reaction to spontaneous radiation emitted by the atom. As I mentioned in the introduction, this question has caused some confusion in the literature, and the physical reality of the zero-point fluctuations has been challenged by Pauli and other physicists. The matter has now been resolved. There is in fact no unique answer because, as is often the case in physics, one can substitute one set of variables for another in the calculation owing to the coupling between different degrees of freedom which it is necessary to introduce in order for the transition to take place. In the present case one can interchange at will operators for the atom and for the field, since these commute. One can in this way change the relative contributions of spontaneous and stimulated emission to the total transition rate. A good account of this question has been given by Cohen-Tadjoui (1966), who points out that if we demand that each contribution be governed by a hermitian term in the interaction Hamiltonian, so that according to the precepts of quantum mechanics each contribution is physically real, then the spontaneous and stimulated contributions are precisely equal.

If we adopt this representation, then we also have to say that *when the atom is in the ground state these two processes are equal and opposite* (Fain and Khanin 1969). In other words, the atom in its ground state is continually radiating to and absorbing energy from the zero-point fluctuations of the vacuum electromagnetic field at equal rates, and these two processes undergo complete destructive interference. As in the classical case considered by Planck, the radiation rate is determined by the *noise power* in the zero-point fluctuations of the atomic dipole moment, and the absorption rate is determined by the *noise power* in the zero-point fluctuations of the electromagnetic field. Each of these powers is $\frac{1}{2}h\nu$ per mode, so the two processes come into equilibrium without any net transfer of energy.

From this point of view, the hydrogen atom has a stable ground state in which the electron does not fall into the proton only because the system is kept 'pumped up' by the zero-point fluctuations of the vacuum electromagnetic field. This raises the question of deciding whether this mechanism constitutes a 'detection' of these zero-point fluctuations. This is a purely semantic matter, but we shall follow the usual convention, and regard an atom as detecting an ambient noise power only when the atom is thereby sent into an excited or ionized state. On the other hand, we should regard the Lamb shift and 'spontaneous' decay of the atom as representing a detection of the zero-point fluctuations. Again, it is a noise power that is being detected, since the energy shift and the decay rate can be combined into a complex quantity which is proportional to the Fourier transform of an autocorrelation function (Barton 1970; Agarwal 1974, 1975).

With these ideas in mind, we can now reconsider Einstein and Stern's attempt to include zero-point motion in the Einstein–Hopf derivation of the fluctuation formula for black body radiation. The damping force F arising from uniform motion V of the harmonic oscillator is given by (Einstein 1909)

$$F=\eta_\nu V,$$

where

$$\eta_\nu=\frac{3}{2c}\left(\epsilon_\nu-\frac{1}{3}\nu\frac{\mathrm{d}\epsilon_\nu}{\mathrm{d}\nu}\right).$$

Now the zero-point fluctuations of the radiation field possess an energy $\frac{1}{2}h\nu$ per mode, and there are $(8\pi/c^3)\nu^2\,\mathrm{d}\nu$ modes between ν and $\nu+\mathrm{d}\nu$. Hence these fluctuations have a spectrum

$$\epsilon_{ZP} = \frac{4\pi h}{c^3} \nu^3 \, d\nu.$$

We now note that *for this spectrum the friction co-efficient η_ν vanishes.* The fundamental reason for this is clear. In the vacuum state there is no rest-frame defined (Lorentz-invariance of the vacuum), and so a uniform inertial motion cannot be damped out by a frictional force. Consistency therefore requires that the zero-point fluctuations should have a Lorentz-invariant spectrum, which a straightforward calculation shows must have the form $\nu^3 \, d\nu$.

It follows that Einstein's result for η_ν relates only to those fluctuations associated with the Planckian thermal component of the wave field (although, as we have seen, allowance must be made for the interference between the thermal waves and the zero-point fluctuations). This conclusion follows also from Einstein's derivation using the inverse of the Boltzmann relation $S = k \ln W$, since the entropy associated with the zero-point fluctuations vanishes, and so there can be no zero-point fluctuations in the entropy.

7 Response Theory

Response theory was initiated in 1931 by Onsager, who extended Einstein's theory of Brownian motion by taking into account the perturbation of the dissipative system by the system being dissipated. He related this perturbation to the equilibrium fluctuations of the unperturbed dissipative system by an ansatz which was very much in Einstein's spirit. He assumed that if, as a result of these fluctuations, the system at one instant deviated appreciably from its mean configuration, then on average it would regress back to the mean at the same rate as if the deviation had been produced by an external perturbation. This average regression would represent an irreversible approach to equilibrium and so would determine the generalized friction co-efficient associated with the response of the system to the external perturbation. In the approximation envisaged, the perturbation would have a linear effect, and as an example we may consider an electric current as the response to an impressed voltage. The linear coefficient of this response, the resistance, not only would govern the rate at which the current would die irreversibly away after the voltage is removed, but also would govern the rate of dissipation associated with Joule heating. With Onsager's ansatz we would expect this resistance to be determined by the equilibrium fluctuations of the unperturbed system, and so we would have a further reason for obtaining a fluctuation-dissipation relation.

Of course, the friction coefficient determines only the out-of-phase response of the system. However, we would also expect the in-phase response, that is the reactance of the system, to be determined by the equilibrium fluctuation spectrum. This follows from the dispersion relations which are a direct consequence of the causality requirement that the response of the system should not precede the disturbance to it. The reactance at any frequency can then be determined by an integral of the friction over all frequencies; the friction is, of course, itself determined by the fluctuation spectrum.

One can also calculate the total response directly: this was first done by Kubo (1957). Kubo used the fact that an averaging of an observable over a system in thermal equilibrium involves multiplying the observable by $e^{-H/kT}$, where H is the Hamiltonian of the system and T is its temperature. This is very similar to multiplying the observable by a quantum mechanical time-evolution operator, with the temperature acting as the reciprocal of an imaginary time. By exploiting this analogy, and using complex variable methods, Kubo arrived at the following fluctuation-response relation:

$$\sigma_{ab}(\nu) = P(\nu, T) \int_{-\infty}^{\infty} e^{-i\nu t} \langle j_a(t) j_b(0) \rangle \, dt,$$

where the complex conductivity σ_{ab} is given in terms of an external field $E_b(\nu)$ by

$$j_a = \sigma_{ab} E_b,$$

$P(\nu, T)$ is the Planckian function (with zero-point contribution),

$$P(\nu, T) = \frac{1 - e^{-h\nu/kT}}{2\nu},$$

and $\langle j_a(t) j_b(0) \rangle$ represents the quantum thermal correlation function of the unperturbed current fluctuations in the system. Note, in particular, that by microscopic reversibility,

$$\langle j_a(t) j_b(0) \rangle = \langle j_a(0) j_b(t) \rangle,$$

and so

$$\sigma_{ab} = \sigma_{ba};$$

these are just Onsager's reciprocal relations.

8 Zero-Point Fluctuations and Accelerated Observers

So far in our discussion of quantum fluctuations, we have implicitly assumed that they are being evaluated in an inertial frame of reference. To make our first contact with modern developments, we ask the apparently innocent but Einsteinian question: How would they appear to an accelerating observer? In particular, if the observer were carrying with him a detector in the form of a harmonic oscillator or an atom, what would this detector register?

In order to answer this question, we recall Planck's classical discussion of the interaction between electromagnetic radiation and a harmonic oscillator, and the corresponding quantum analogue. For an inertial oscillator the absorption rate is proportional to the noise power in the electromagnetic field, and the emission rate is proportional to the noise power of its own dipole oscillations. Because of the present application we must now make explicit an important detail. The absorption rate is proportional to the *positive frequency component* of the noise power (see e.g. Mandel and Wolf 1965). For an accelerated oscillator, we can imagine in a crude fashion that the acceleration results from the oscillator being pulled by a rope. The effect of the rope might be to distort the oscillator appreciably, but we are not really interested in effects of this type, which depend on the structure of the oscillator and which, in principle at least, could be calculated in a straightforward manner. We are interested in the *kinematical* effects arising purely from the fact that the world-line of the oscillator is no longer a geodesic. We shall therefore assume that our oscillator is so robust that we may neglect any structure-sensitive effects arising from the action on it of an external force.

We shall therefore suppose that its absorption and emission rates remain proportional to the noise powers in the field and its own oscillations, respectively. However, the absorption rate now depends on the positive frequency components of the noise as measured by the *accelerating* detector. Because of our robustness assumption we may suppose that, if the oscillator is initially in its ground state, the noise power of its own zero-point oscillations is $\frac{1}{2}h\nu$, where ν is the frequency variable conjugate to the proper-time of the oscillator. *However, this is no longer true for the noise power in the field.* This noise power must now be evaluated along the accelerated world-line of the oscillator, which deviates from a geodesic. By the Wiener–Khinchine theorem, this noise power is the Fourier transform of the autocorrelation function of the field *evaluated along the accelerated world-line*. Now we would expect such an autocorrelation function to be the same along all timelike geodesics in a vacuum electromagnetic field, since the vacuum state is Lorentz-invariant and we can transform any timelike geodesic (in Minkowski space–time) into any other by a Lorentz transformation. But there is no reason to expect the autocorrelation function to be the same along a non-geodesic (that is, accelerating) world-line, which of course would take us outside the Lorentz group.

We now come to the most remarkable feature of this situation. If the world-line is that of a *uniformly* accelerated observer, in the sense of special relativity, then the additional noise power in the field has a *precisely thermal* spectrum, with the temperature T given by

$$kT = \frac{\hbar a}{2\pi c},$$

where a is the observer's proper acceleration. This result was discovered only in 1976, by W. G. Unruh (although a closely similar result was obtained by P. C. W. Davies in 1975). It should not be regarded as a casual algebraic coincidence, because, as we shall see, *the quantum fluctuations about this mean state are also precisely those derived by Einstein for thermal radiation at temperature* T.

Uniform acceleration can be defined by reference to an instantaneous inertial frame relative to which the accelerated observer has zero velocity. At any instant we may define the *proper* acceleration as the acceleration relative to such an inertial frame. If this proper acceleration is independent of time, then we say that the acceleration is uniform. An example of such motion is provided by a classical electric charge subject to a uniform, static, electric field. A good discussion of the properties of this motion has been given by Rindler (1966). In particular, he shows that if the motion is confined to, say, the z-axis, then the world-line is a rectangular hyperbola in the z, t plane, with asymptotes $z = \pm ct$. These asymptotes are the world-lines of light-rays, and represent *event horizons* for the accelerating observer. No event occurring outside the region bounded by these asymptotes can communicate with an observer confined to this region. This inability of the light to catch up with the observer is compatible with special relativity (indeed, demanded by it) because the observer is accelerating. The velocity of light is invariant only relative to an inertial observer.

To bring out this, and other, features of uniformly accelerated motion, it is sometimes helpful to use a non-inertial frame of reference attached to the accelerating observer. To arrive at a coordinate system that is as extensive as possible, we may consider a family of accelerating observers, one for each rectangular hyperbola with asymptotes $z = \pm ct$. The natural coordinate system to use is then the co-moving one in which, along each hyperbola, the space coordinate is constant while the time coordinate is proportional to the proper time as measured from an initial instant $t = 0$ in some inertial frame. It turns out to be convenient to choose a different constant of proportionality along each hyperbola, namely, the associated proper acceleration. When one calculates the vacuum noise along the world-line of an accelerating observer, one obtains the thermal result already mentioned. One also finds the Einstein fluctuations associated with this thermal state.

Another aspect of this thermal state is the role played by the horizon. We might expect it to be important from the following simple physical consideration. In order to establish that he is observing an essentially thermal distribution, an observer must maintain his acceleration at a constant rate for a time long enough for him to observe several oscillation cycles of the ambient radiation field at frequencies near the peak of the Planck distribution. It follows immediately from dimensional considerations that during this time the observer's change in velocity, as measured in a single inertial frame, is close to the velocity of light. Thus, the asymptotic part of the motion plays a crucial role in the establishment of the thermal distribution, and this part is closely related to the occurrence of an event horizon.

The role of the horizon can be demonstrated more precisely in connection with our previous remark that we are justified in regarding the Rindler state as truly thermal only if *the fluctuations about the equilibrium state satisfy Einstein's relation*. In conventional thermal physics, the randomness that underlies these fluctuations would arise partly from the radiation being produced by a very large number of excited atoms which would be radiating independently, and partly from a quantum origin. As Jordan showed, Einstein's relation can be derived statistically for *any* quantum radiation field with random phases, the linear term arising from the interference between the zero-point motion and the excited motion. However, a quadratic term involving only the zero-point motion is absent, so that in the vacuum state there would be no fluctuations of the zero-point energy. This is consistent with the conventional procedure of setting the entropy of the vacuum state to zero.

The situation is quite different for fluctuations evaluated along a Rindler hyperbola. The inertial vacuum state is itself a pure state, but *it involves the zero-point motion of modes which lie entirely beyond the Rindler event horizons.* These modes are correlated with the modes of the Rindler space–time, but they cannot be detected by a Rindler observer. Accordingly, the correlation is broken; the Rindler observer detects a purely random field and so obtains the Einstein relation for the fluctuations. This then guarantees the consistency of the thermal Rindler state. For example, a mirror coupled to the thermal field would develop the appropriate fluctuations in its translational motion. We can therefore associate the usual non-zero entropy with this thermal state.

We finally consider the particle description which a Rindler observer would give for the quantum wave-field he observes. It was first pointed out by Fulling in 1973 that, since the Rindler metric is static, it is possible to construct creation and annihilation operators for the field using normal modes defined with respect to Rindler time τ. One can also define a vacuum state in the usual way, as that which is destroyed by all those annihilation operators. This state will be invariant under the group that preserves the Rindler metric (consisting of the transformations $\tau \rightarrow \tau + \tau_0$, and displacements and rotations in the x, y plane). Fulling's main point was that this quantization procedure is *inequivalent* to the usual one carried out in an inertial frame. In particular, the vacua are different (as we have already mentioned), and a positive-frequency state for an inertial observer may not be a positive-frequency state for a Rindler observer.

It was Unruh (1976) who pointed out that the Fulling–Rindler particles that correspond to an inertial vacuum have a thermal distribution (cf. also Davies 1975). He pointed out in addition that a Rindler detector would detect these particles. In this sense, Fulling–Rindler particles are 'real' particles. There is no particular mystery about this. It corresponds to my earlier remark that the noise along a

Rindler hyperbola is greater than the noise along a geodesic, and that it is this excess noise that excites a Rindler detector.

9 Do Zero-Point Fluctuations Produce a Gravitational Field?

We now wish to comment on the unsolved problem of the relation between zero-point fluctuations and gravitation. If we ascribe an energy $\frac{1}{2}h\nu$ to each mode of the vacuum radiation field, then the total energy of the vacuum is infinite. It would clearly be inconsistent with the original assumption of a background Minkowski space–time to suppose that this energy produces gravitation in a manner controlled by Einstein's field equations of general relativity. It is also clear that the space–time of the real world approximates closely to the Minkowski state, at least on macroscopic scales. It thus appears that we must regularize the zero-point energy of the vacuum by subtracting it out according to some systematic prescription. At the same time, we would expect zero-point energy *differences* to gravitate. For example, the (negative) Casimir energy between two plane-parallel perfect conductors would be expected to gravitate; otherwise, the relativistic relation between a measured energy and gravitation would be lost. Similarly, the regularized vacuum energy in a curved space–time would be expected to gravitate, where the regularization is achieved by subtracting out the Minkowski contribution in a systematic way.

This procedure is needed in order to obtain a pragmatically workable theory. The difficulty with it is that existing theory does not tell which is the fiducial state whose energy is to be set to zero. It is no doubt an intelligent guess that one should take Minkowski space–time for this fiducial state, but the awkward point is that this (or any other) choice is not *prescribed* by existing theory. Clearly, something essential is missing.

Another way of looking at this problem is in terms of the cosmological constant. If the vacuum energy density is a local quantity, one would expect that, for reasons of symmetry, it would have to have the form $\lambda g_{\mu\nu}$, and so correspond to the 'cosmological' term in Einstein's field equations. One is reminded of the related problem that, in grand unified and supergravity type theories, one would expect to have a cosmological term of order 10^{120} times greater than any allowed by observational cosmology. Some very fine-tuned cancellations seem to be required to achieve agreement with observation. One sees occasional claims that some varieties of superstring theory lead naturally to the vanishing of the cosmological constant, but no agreement has yet been reached on this point. We here face a fundamental problem of outstanding importance. Its solution may still require a radical change in our theories beyond our present imagining.

References

Agarwal, G. S. (1974), 'Coherence in Spontaneous Emission in the Presence of a Dielectric', *Phys. Rev. Lett.* 32: 703–6.

—— (1975), 'Quantum Electrodynamics in the Presence of Dielectrics and Conductors. I: Electromagnetic Field Response Functions and Black Body Fluctuations in Finite Geometries', *Phys. Rev.* A11: 230–42.

Barton, G. (1970), 'Quantum Electrodynamics of Spinless Particles Between Conducting Plates', *Proc. Roy. Soc.* A320: 251–75.

Callen, H. B. and Welton, T. A. (1951), 'Irreversibility and Generalised Noise', *Phys. Rev.* 83: 34–40.

Casimir, H. B. G. (1948), 'On the Attraction Between Two Perfectly Conducting Plates', *Kon. Ned. Akad. Wetensch.* B51: 793–6.

Cohen-Tannoudji, C. (1986), 'Fluctuations in Radiative Processes', *Physica Scripta*, T12: 19–27.

Davies, P. C. W. (1975), 'Scalar Particle Production in Schwarzschild and Rindler Metrics', *J. Phys.* A8: 609–16.

Einstein, A. (1909), 'Zur gegenwärtigen Stand des Strahlungsproblems', *Phys. Zeit.* 10: 185–93.

—— and Stern, O. (1913), 'Einige Argumente für die Annahme einer molekularen Agitation beim absoluten Nullpunkt', *Ann. d. Phys.* 40: 551–60.

Fain, V. M. and Khanin, Y. I. (1969), *Quantum Electronics*, vol. I: *Basic Theory*, Pergamon, Oxford.

Feynman, R. P. (1961), *Solvay Institute Proceedings*, Wiley (Interscience), New York, pp. 76–9.

Fulling, S. A. and Davies, P. C. W. (1975), 'Radiation from a Moving Mirror in Two Dimensional Space–Time: Conformal Anomaly', *Proc. Roy. Soc.* A348: 393–414.

Heisenberg, W. (1925), 'Über quantentheoretische Umdeutung kinematischer und mechanischer Beziehungen', *Zeit. f. Phys.* 43: 172–98.

Hellman, H. (1927), *Einführung in die Quantenchemie*, Deutsche, Leipzig.

Kubo, R. (1957), 'Statistical Mechanical Theory of Irreversible Processes I', *J. Phys. Soc. Japan*, 12: 570–86.

Mandel, L. and Wolf, E. (1965), 'Coherence Properties of Optical Fields', *Rev. Mod. Phys.* 37: 231–87.

Power, E. A. (1966), 'Zero-Point Energy and the Lamb Shift', *Am. J. Phys.* 34: 516–18.

Rindler, W. (1966), 'Kruskal Space and the Uniformly Accelerated Frame', *Am. J. Phys.* 34: 1174–8.

Senitzky, I. R. (1960), 'Dissipation in Quantum Mechanics: The Harmonic Oscillator', *Phys. Rev.* 119: 670–9.

Unruh, W. G. (1976), 'Notes on Black-Hole Evaporation', *Phys. Rev.* D14: 870–92.

Weber, J. (1954), 'Vacuum Fluctuation Noise', *Phys. Rev.* 94: 215–17.

Weisskopf, V. F. (1949), 'Recent Developments in the Theory of the Electron', *Rev. Mod. Phys.* 21: 305–15.

Welton, T. A. (1948), 'Some Observable Effects of the Quantum-Mechanical Fluctuations of the Electromagnetic Field', *Phys. Rev.* 74: 1157–67.

de Witt, B. S. (1975), 'Quantum Field Theory in Curved Spacetime', *Phys. Rep.* 19C: 295–357.

7
The Vacuum and Unification

I. J. R. AITCHISON

1 Introductory Survey

It is a common observation that fundamental physics seems to have progressed, over the last hundred years or more, in the direction of ever greater unification. It is fair to describe this movement as progress, because it does appear that the drive towards unification has, in many diverse ways, acted as a powerful heuristic device in constructing phenomenologically successful theories. Less widely noted, perhaps, is the crucial role that the *vacuum* (as currently understood) has played, and still plays, in many aspects of the unification programme. This is what I want to examine, very briefly, in the present contribution.

'Unification' is a very broad notion, and it will help to have some sort of chart to navigate by. Referring to Fig. 1, in which historical time is increasing down the page (roughly), it was traditional to distinguish three fundamental conceptual categories: 'Matter', 'Force(s)', and 'Space–time'. We shall be concerned with unification within and between these categories.

Our first example (Section 2) will be the unification, under the 'Force(s)' column, of electricity and magnetism achieved by Maxwell. This is now a well studied topic, so we can be very sketchy. Our purpose is to emphasize Maxwell's clear dependence on a literal mechanical model of the vacuum, in the formative stages of his theory.

As time went by, of course, this *mechanical ether* came to seem less and less necessary, or plausible, and the electromagnetic *field* emerged as a new, non-mechanical, concept; its vibrations were supposed not to require the existence of any underlying mechanical contraption. Gravity was also naturally regarded as a field theory. A

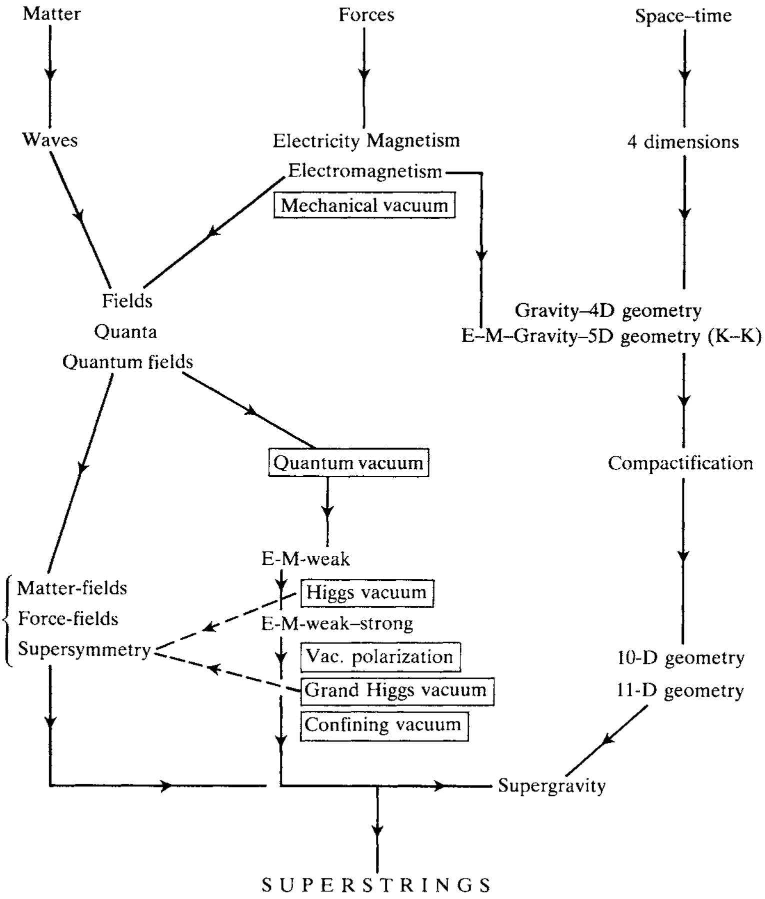


FIG. 1 Themes in unification.

clear distinction remained, however, between the field and its sources: the latter, of course, were electric charges for electromagnetism, and masses for gravity, both of which were thought of as associated with matter.

With the advent of quantum mechanics, this distinction between force-fields and their matter sources began to break down. On the one hand, matter, instead of being discrete and particle-like, revealed (in electron diffraction, for example) continuous, wavelike aspects—such as were characteristic of wave-fields. On the other, the energy of electromagnetic radiation was found to be quantized. These developments eventually led to the notion of a *quantum field*, which—until the appearance of strings!—remained our most fundamental concept. All particles were understood to be quantized excitations of appropriate quantum fields. Nevertheless, what appeared to be one fundamental distinction between 'matter-fields' and 'force-fields' still persisted: the typical matter quanta (electrons, quarks, etc.) are *fermionic*, while the force quanta (photons, gluons, $W^{\pm}$, Z^{0}, gravitons) are *bosonic*.

From its very beginning, with Dirac's treatments of spontaneous decay and of pair creation, quantum field theory provided a rich and powerful conceptual framework for discussing the vacuum. The state with no visible particles in it is one in which all the quantum fields are unexcited: it is the ground (or lowest energy) state of the field system. This *quantum vacuum* need by no means be featureless—any more than the ground state of a complicated many-body system need be. In Section 3 we shall consider how specific features of the quantum vacuum, established or hypothetical, played a crucial role in several recent unification 'moves'. We shall briefly recall the well-known vacuum polarization effects which lead to the possibility of equal strengths (at very short distances) for the weak, electromagnetic, and strong interactions; and the way in which, via the 'Higgs mechanism', the vacuum takes the blame for the wide differences in the observed ranges of these forces.

The Higgs mechanism is not without its problems (Section 4). One—the so-called 'hierarchy problem'—provides some motivation for a further unification move, of a potentially very fundamental nature: the unification of fermions and bosons via *supersymmetry*. If, as suggested above, we think of 'matter' as fermionic and 'force' as bosonic, supersymmetry would provide a unification of matter-fields and force-fields.

Thus far we have followed the 'Matter' and 'Force(s)' columns of Fig. 1. The other problem with the Higgs mechanism engages us with the remaining column, since it involves cosmology—and that was related to 'Space–time', of course, by Einstein's General Relativity. The category 'Space–time' itself accepts the unification (in some sense) of space and time achieved in Einstein's Special Theory of Relativity. His General Theory gave a geometrical interpretation to gravity, supplying for the first time a link between the 'Force(s)' and the 'Space–time' columns, and inspiring the even grander vision of a geometrical basis for all forces.

The earliest proposal for such geometrical unification was Kaluza's (1921) suggestion that gravity could be unified with electromagnetism (then the only other identified force) by formulating general relativity not in four dimensions but in five. This suggestion was further developed by Klein in 1926, and formed a major theme in Einstein's work on unified field theory. In the developed theory, the crucial idea was that for some unexplained reason the fifth dimension was *compactified*, that is, curled up so as to have the topology of a circle, rather than an infinitely extended line. If the radius of the circle was small enough, direct observational evidence for the fifth coordinate might be impossible, all observed phenomena involving merely averages over positions on the circle.

How may this idea be described within the framework of quantum field theory? One plausible way of approaching quantum gravity is to regard the fundamental field of classical general relativity—the metric tensor—as a quantum field. It would then be subject to quantum fluctuations, which would include the possibility of different space–time geometries. In the absence of matter sources, the ground state of this quantum metric field system—the gravitational vacuum—should yield four-dimensional Minkowski space, M^4. In the modern version of Kaluza–Klein theory, one hopes that the vacuum of the five-dimensional general relativity is not Minkowski space, M^5, but rather the product $M^4 \times S^1$ of the four-dimensional M^4 with a circle S^1. Such an outcome is, of course, very *un*symmetrical as regards the four coordinates of our own M^4, and the compactified coordinates of S^1. Thus the vacuum is here, as in the Higgs mechanism, associated with a *loss* of symmetry—which, however, then allows a unification move to be made! In Section 5 we shall briefly discuss a possible way in which this vacuum-breaking of the five-dimensional symmetry might be understood dynamically.

We have now arrived midway down the 'Space–time' column, opposite the various forces known today. Gravity and electromagnetism are no longer the whole story, and five dimensions are not enough to provide for a geometrical unification of all presently identified forces: at least ten dimensions seem to be needed. Attempts to find a realistic ten- (or even eleven-) dimensional *supergravity theory* (unifying gravity with unified matter-force) were widely pursued in the early 1980s. The basic idea was still the same as in the original

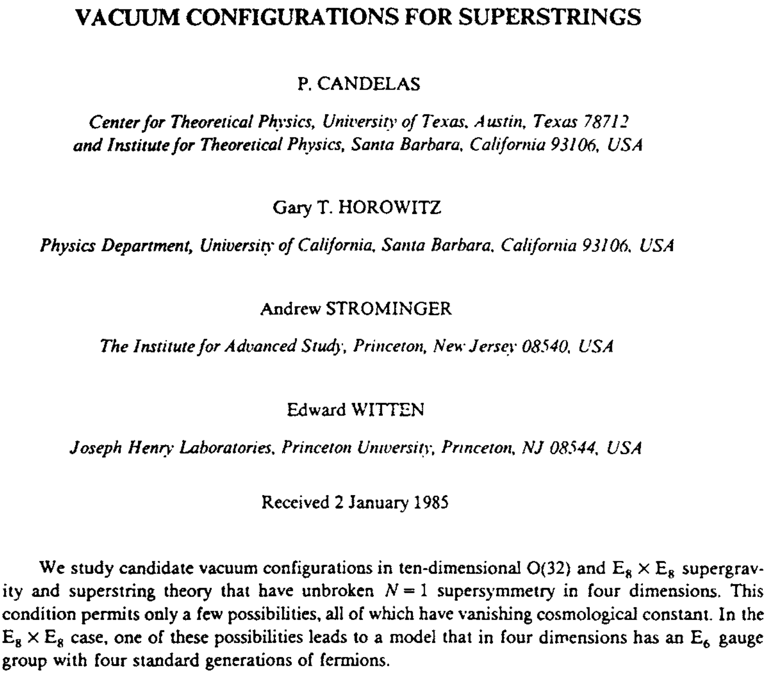
Kaluza–Klein picture, but now the vacuum configuration in the compactified dimensions had to be that appropriate to the group structure, and to the particle multiplet structure of the phenomenologically well established 'standard model' for elementary particles.

Then in 1984 came the discovery by Green and Schwartz of a consistent quantum theory of all forces including gravity (technically, the first 'anomaly-free' theory) based on a framework going beyond the local quantum field concept. This was a (super)*string* theory, in which the fundamental object has a finite extension, of order the Planck length ($\sim 10^{-33}$ cm). In such theories, the whole of our mundane physics at distances much greater than this tiny size—geometry, symmetries, and matter content—is viewed as merely the ground state of the superstring system: in short, as its *vacuum*. Figure 2 shows part of the title page of a paper that generated much additional enthusiasm for the superstring idea. These considerations, which have brought us to the end of Fig. 1, will be briefly described in the concluding section.

2 The Mechanical Vacuum of Maxwell's Electromagnetic Theory

As indicated in Fig. 1, to the classical physicist matter and force were clearly separated. The nature of matter was intuitive, based on everyday macroscopic experience; force, on the other hand, was more problematical. Contact forces between extended bodies were easy to understand, but forces that seemed capable of *acting at a distance* caused difficulties. As Maxwell reminded his audience in a Royal Institution lecture entitled 'On Action at a Distance',

> so far was Newton from asserting that bodies really do act on one another at a distance, independently of anything between them, that in a letter to Bentley, which has been quoted by Faraday in this place, he says:– 'It is inconceivable that inanimate brute matter should, without the mediation of something else, which is not material, operate upon and affect other matter without mutual contact, as it must do if gravitation, in the sense of Epicurus, be essential and inherent in it . . . That gravity should be innate, inherent, and essential to matter so that one body can act upon another at a distance, through vacuum, without the mediation of anything else, by and through which their action and force may be conveyed from one to another, it is to me so great an absurdity that I believe no man who has in philosophical matters a competent faculty of thinking can ever fall into it. (Maxwell 1890, ii: 311 ff.)



VACUUM CONFIGURATIONS FOR SUPERSTRINGS

P. CANDELAS

Center for Theoretical Physics, University of Texas, Austin, Texas 78712
and Institute for Theoretical Physics, Santa Barbara, California 93106, USA

Gary T. HOROWITZ

Physics Department, University of California, Santa Barbara, California 93106, USA

Andrew STROMINGER

The Institute for Advanced Study, Princeton, New Jersey 08540, USA

Edward WITTEN

Joseph Henry Laboratories, Princeton University, Princeton, NJ 08544, USA

Received 2 January 1985

We study candidate vacuum configurations in ten-dimensional O(32) and $E_8 \times E_8$ supergravity and superstring theory that have unbroken $N = 1$ supersymmetry in four dimensions. This condition permits only a few possibilities, all of which have vanishing cosmological constant. In the $E_8 \times E_8$ case, one of these possibilities leads to a model that in four dimensions has an E_6 gauge group with four standard generations of fermions.

FIG. 2 Title page of an influential string paper.
Source: Nuclear Physics, B258: 46, North-Holland, Amsterdam, 1985.

Newton could find no satisfactory mechanism for the transmission of the gravitational force between two distant bodies, though he certainly tried hard to find one in terms of pressure forces in an intervening 'medium'.

The nineteenth century saw the precise formulation of the more intricate force laws of electromagnetism. Here too the distaste for action-at-a-distance theories led to numerous mechanical, or fluid mechanical, models of the way electromagnetic forces, and light, are transmitted. Because of the 'rotatory' character of the action of a current on a magnet, as discovered by Oersted, 'many minds', as Maxwell put it, 'were set a-speculating on vortices and streams of aether whirling around the current' (1890, ii: 317 ff.). In his own series of papers in 1861–2 entitled 'On Physical Lines of Force', Maxwell made essential use of vortex models, as he struggled to give physical and mathematical substance to Faraday's empirically won concepts of lines of force. Here, for example, is a quotation from the first paper in the series, which was entitled 'The Theory of Molecular Vortices Applied to Magnetic Phenomena':

> Let us now suppose that the phenomena of magnetism depend on the existence of a tension in the direction of the lines of force, combined with a hydrostatic pressure; or in other words, a pressure greater in the equatorial than in the axial direction: the next question is, what mechanical explanation can we give of this inequality of pressures in a fluid or mobile medium? The explanation which most readily occurs to the mind is that the excess of pressure in the equatorial direction arises from the centrifugal force of vortices or eddies in the medium having their axes in directions parallel to the lines of force. (Maxwell 1861a: 165)

The existence of the 'medium' is quite taken for granted; the problem is to imagine a dynamics for it which could account for the observed electromagnetic phenomena.

The vortex model was significantly elaborated in the second paper in the series, entitled 'The Theory of Molecular Vortices Applied to Electric Currents'. Here is Maxwell's introduction of the famous 'idle wheels', which allowed adjacent vortices to rotate smoothly in the same sense:

> We know that the lines of force are affected by electric currents, and we know the distribution of those lines about a current; so that from the force we can determine the amount of the current. Assuming that our explanation of the lines of force by molecular vortices is correct, why does a particular distribution of vortices indicate an electric current? A satisfactory answer to this question would lead us a long way towards that of a very important one, 'what is an electric current?'
>
> I have found great difficulty in conceiving of the existence of vortices in a medium, side by side, revolving in the same direction about parallel axes. The contiguous .portions of consecutive vortices must be moving in opposite directions; and it is difficult to understand how the motion of a part of the medium can coexist with, and even produce, an opposite motion of a part in contact with it.
>
> The only conception which has at all aided me in conceiving of this kind of motion is that of the vortices being separated by a layer of particles, revolving each on its own axis in the opposite direction to that of vortices, so that the contiguous surfaces of the particles and of the vortices have the same motion.
>
> In mechanism, when two wheels are intended to revolve in the same direction, a wheel is placed between them so as to be in gear with both, and this wheel is called an 'idle wheel'. The hypothesis about the vortices which I have to suggest is that a layer of particles, acting as idle wheels, is interposed

between each vortex and the next, so that each vortex has a tendency to make the neighbouring vortices revolve in the same direction with itself. (Maxwell 1861b: 288)

In an imaginative but strictly economical stroke, Maxwell now proposed to put these new hypothetical entities to good use: he suggested that the motion of the 'particles, acting as idle wheels' constituted an electric current. With this idea he was able to explain the phenomenon of induced currents, as illustrated in his famous diagram reproduced here as Fig. 3.

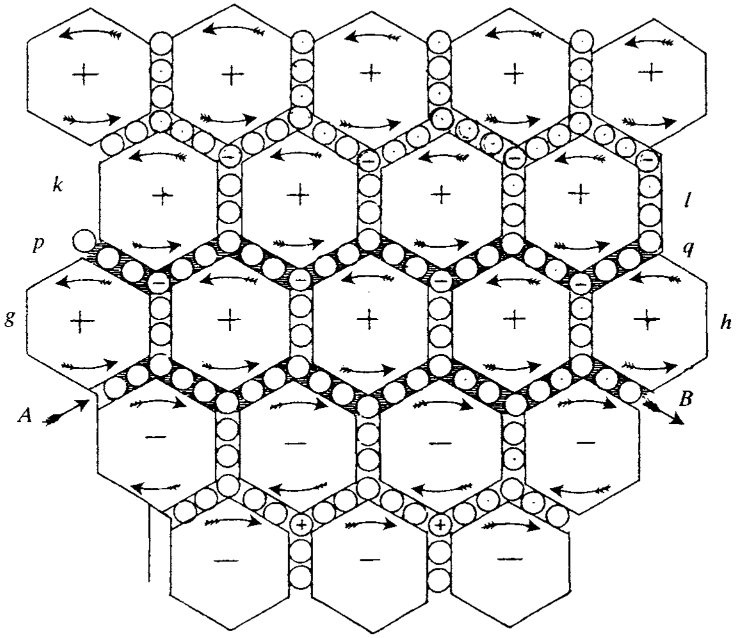


FIG. 3 Maxwell's vortex ether.

Referring to this diagram, Maxwell wrote:

We shall in the first place examine the process by which the lines of force are produced by an electric current.

Let AB, Plate VIII, p. 488, fig. 2, represent a current of electricity in the direction from A to B. Let the large spaces above and below AB represent the vortices, and let the small circles separating the vortices represent the layers of particles placed between them, which in our hypothesis represent electricity.

Now let an electric current from left to right commence in AB. The row of vortices *gh* above AB will be set in motion in the opposite direction to that of a watch. (We shall call this direction +, and that of a watch −.) We shall suppose the row of vortices *kl* still at rest, then the layer of particles between these rows will be acted on by the row *gh* on their lower sides, and will be at rest above. If they are free to move, they will rotate in the negative direction, and will at the same time move from right to left, or in the opposite direction from the current, and so form an *induced* current. (Maxwell 1871: 291)

The third paper of the series was entitled 'The Theory of Molecular Vortices Applied to Statical Electricity', and it is here that the great unification of light and electromagnetism was proposed:

The velocity of light in air, as determined by M. Fizeau, is 70 843 leagues per second (25 leagues to a degree) which gives

$$\begin{aligned} V &= 314\,858\,000\,000 \text{ millimetres} \\ &= 195\,647 \text{ miles per second} \end{aligned}$$

The velocity of transverse undulations in our hypothetical medium, calculated from the electromagnetic experiments of MM. Kojlrausch and Weber, agrees so exactly with the velocity of light calculated from the optical experiments of M. Fizeau, that we can scarcely avoid the inference that *light consists in the transverse undulations of the same medium which is the cause of electric and magnetic phenomena.* (Maxwell 1862: 22)

Similar considerations were advanced in his Royal Institution Lecture, the peroration of which is worth reproducing in full:

But if the luminiferous and the electro-magnetic media occupy the same place, and transmit disturbances with the same velocity, what reason have we to distinguish the one from the other? By considering them as the same, we avoid at least the reproach of filling space twice over with different kinds of aether.

Besides this, the only kind of electro-magnetic disturbances which can be propagated through a non-conducting medium is a disturbance transverse to the direction of propagation, agreeing in this respect with what we know of that disturbance which we call light. Hence, for all we know, light also may be an electro-magnetic disturbance in a non-conducting medium. If we admit this, the electro-magnetic theory of light will agree in every respect with the undulatory theory, and the work of Thomas Young and Fresnel will be established on a firmer basis than ever, when joined with that of Cavendish and Coulomb by the keystone of the combined sciences of light and

> electricity—Faraday's great discovery of the electromagnetic rotation of light.
>
> The vast interplanetary and interstellar regions will no longer be regarded as waste places in the universe, which the Creator has not seen fit to fill with the symbols of the manifold order of His kingdom. We shall find them to be already full of this wonderful medium; so full, that no human power can remove it from the smallest portion of space, or produce the slightest flaw in its infinite continuity. It extends unbroken from star to star; and when a molecule of hydrogen vibrates in the dog-star, the medium receives the impulses of these vibrations; and after carrying them in its immense bosom for three years, delivers them in due course, regular order, and full tale into the spectroscope of Mr Huggins, at Tulse Hill.
>
> But the medium has other functions and operations besides bearing light from man to man, and from world to world, and giving evidence of the absolute unity of the metric system of the universe. Its minute parts may have rotary as well as vibratory motions, and the axes of rotation from those lines of magnetic force which extend in unbroken continuity into regions which no eye has seen, and which, by their action on our magnets, are telling us in language not yet interpreted, what is going on in the hidden underworld from minute to minute and from century to century.
>
> And these lines must not be regarded as mere mathematical abstractions. They are the directions in which the medium is exerting a tension like that of a rope, or rather, like that of our own muscles. The tension of the medium in the direction of the earth's magnetic force is in this country one grain weight on eight square feet. In some of Dr. Joule's experiments, the medium has exerted a tension of 200 lbs. weight per square inch.
>
> But the medium in virtue of the very same elasticity by which it is able to transmit the undulations of light, is also able to act as a spring. When properly wound up, it exerts a tension, different from magnetic tension, by which it draws oppositely electrified bodies together, produces effects through the length of telegraph wires, and when of sufficient intensity, leads to the rupture and explosion called lightning.
>
> These are some of the already discovered properties of that which has often been called vacuum, or nothing at all. They enable us to resolve several kinds of action at a distance into actions between contiguous parts of continuous substance. Whether this resolution is of the nature of explication or complication, I must leave to the metaphysicians. (Maxwell 1890, ii: 322–3)

This is very 'concrete' language. Did Maxwell really believe in such a literal mechanical vacuum? In the first paper of the series referred to above, he claimed that the theory of molecular vortices was not a 'mechanical illustration ... to assist the imagination', or an 'analog[y]' but 'a theory, which if not *true*, can only be proved to be erroneous by experiments which will greatly enlarge our knowledge of this part of physics'. (Maxwell 1861*a*). He ended the second paper in the series on a more guarded, but similar note:

> We have now shewn in what way electro-magnetic phenomena may be imitated by an imaginary system of molecular vortices. Those who have been already inclined to adopt an hypothesis of this kind, will find here the conditions which must be fulfilled in order to give it mathematical coherence, and a comparison, so far satisfactory, between its necessary results and known facts.
>
> Those who look in a different direction for the explanation of the facts, may be able to compare this theory with that of the existence of currents flowing freely through bodies, and with that which supposes electricity to act at a distance with a force depending on its velocity, and therefore not subject to the law of conservation of energy.
>
> The facts of electro-magnetism are so complicated and various, that the explanation of any number of them by several different hypotheses must be interesting, not only to physicists, but to all who desire to understand how much evidence the explanation of phenomena lends to the credibility of a theory, or how far we ought to regard a coincidence in the mathematical expression of two sets of phenomena as an indication that these phenomena are of the same kind. We know that partial coincidences of this kind have been discovered; and the fact that they are only partial is proved by the divergence of the laws of the two sets of phenomena in other respects. We may chance to find, in the higher parts of physics, instances of more complete coincidence, which may require much investigation to detect their ultimate divergence. (Maxwell 1861*b*: 347–8)

However, by 1864, in the paper (Maxwell 1865) usually cited for 'the Maxwell equations', Maxwell had begun to sidestep the issue of the detailed mechanism of the electromagnetic vacuum. Significantly, the paper was entitled 'A Dynamical Theory of the Electromagnetic Field', in which the aim was to set down a complete set of field equations that would account for all the observed phenomena. Later, in 1873, Maxwell put his field equations into Lagrangian form (the appreciation of generalized mechanics in Britain having been accelerated by the publication in 1867 of the *Treatise on Natural Philosophy* by Maxwell's two friends, the 'Northern Wizards' William Thomson (Lord Kelvin) and P. G. Tait). Many physicists came to agree with Larmor 'that the [construction of a satisfactory Lagrangian] really involves in itself the solution of the whole problem' (quotation taken from Siegel 1981). After all, the same mathematical equations can describe, when suitably interpreted, the dynamics of systems of masses, springs, and dampers, or of inductors, capacitors, and resistors.

For further discussion of the attitude of Maxwell and his contemporaries to mechanical ether models, I must refer to the work of historians (e.g. Siegel 1981). Lest we smile too easily at the naïvety of these great physicists, I show in Fig. 4 a picture, published in 1980, not wholly dissimilar to Fig. 3. Even the title doesn't sound totally unfamiliar:

A COLOR MAGNETIC VORTEX CONDENSATE IN QCD

J. AMBJØRN AND P. OLESEN

Nuclear Physics B170[FS1] (1980) 265–282

The point is, surely, that in trying to cross from conceptual *terra* that is relatively *firma* to new islands of knowledge lying beyond the troubled waters of conflicting theories, partial data, and unexplained phenomena, the theorist must contrive a bridge out of whatever material comes to hand. It may be flimsy, but its very lightness can be tactically advantageous. Once some contact with other land has been established, heavier structures can be assembled. In the formative stages, too much weight should not be placed on the new bridge; eventually, the bridge itself may be discarded, as the waters are forced back, and the islands united with the growing mainland of knowledge. At all events, the enormously fertile heuristic value, for some minds at least, of having a tangible physical model in view when formulating new theories is surely demonstrated by Maxwell's great work of unification—and in his case it was a model, as he said, of 'what has often been called vacuum, or nothing at all'.

In this spirit we shall now consider our own latter-day version of Maxwell's vacuum, and its role in unification. For us, it will be not a mechanical object—thanks partly to Maxwell's own work in creating a field theory of electromagnetism—but rather a quantum-field-theoretic state, since this sort of stuff is all we have, at present, from which to make our models.

3 The Quantum Vacuum and 'Grand Unification'

I gave an introductory account of the vacuum in quantum field theory in 'Nothing's Plenty' (Aitchison 1985), to which I must refer for background. In this section I shall have to recapitulate some of what I said there, before carrying the argument forward in the following sections.

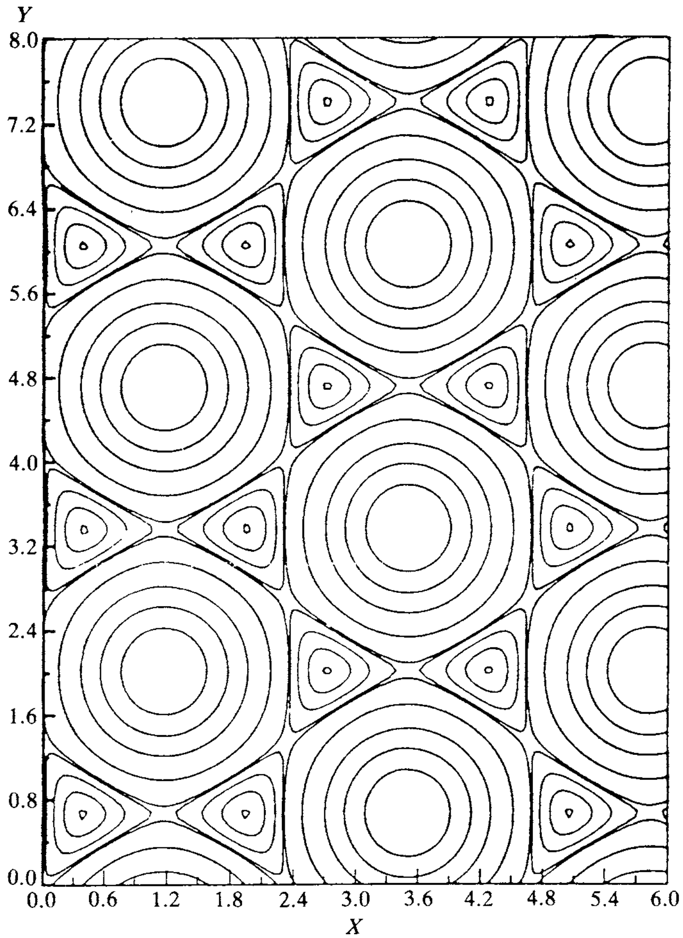


FIG. 4 A modern vortex vacuum.
Source: Nuclear Physics B170, 265 (1980).

A free (non-interacting) quantum field $\hat{\phi}(\mathbf{x}, t)$ is conventionally represented as a sum of an infinite number of *modes*. States with particles present correspond to the states of the field in which some modes are excited; the vacuum state $|0\rangle$ is the state in which all modes are unexcited, and the field is in its ground state. Simple though this sounds, one important caveat should at once be entered: an absolute distinction between a state with particles in it and one without can be sustained only in *flat* space–times. A counter accelerating uniformly with respect to a vacuum defined over flat Minkowski space–time will record a flux of particles, distributed in energy according to a Planck spectrum with an effective temperature proportional to the acceleration (Davies 1975; Unruh 1976; see also Fulling 1973).

Ignoring from now on this effect of the space–time background, we recall two important features of the quantum field theory vacuum, as they relate to unification.

The Effect of Fluctuations

Although the vacuum state is (normally) one in which the average value of all quantum fields is zero ($\langle 0|\hat{\phi}|0\rangle = 0$), the mean square values of the fields are in general not zero: that is, there are fluctuations—and this is true even at absolute zero. These fluctuations are observable: those in the electromagnetic field account for most of the Lamb shift, in hydrogen, while those in charged-particle (matter) fields give rise to the phenomenon of *vacuum polarization*, which accounts for a further contribution (small in hydrogen, but large in muonic helium) to the Lamb shift. In quantum electrodynamics, the effect of the vacuum is partially to screen a test charge at distances $\gtrsim \hbar/m_e c$, where m_e is the mass of electron. In other words, at distances smaller than this size ($\sim 4 \times 10^{-13}$ m) the effective electric charge on a test body will appear to increase. This is the usual effect of polarization in all normal dipolar media (including, in this respect, the vacuum of quantum electrodynamics (QED)). A major surprise was the discovery by Gross and Wilczek (1973) and Politzer (1973) that in non-Abelian gauge theories (such as those now believed to describe the strong and weak interactions of quarks and leptons) vacuum polarization produces a net *anti*-screening effect: the effective strong and weak 'charges' *decrease* at small interparticle separation.

It is conventional to discuss this phenomenon in terms of 'fine-structure constants' α, α_s, and α_w for the electromagnetic, strong, and weak interactions, respectively. These are (up to factors like $4\pi\epsilon_0\hbar c$) just the squares of the appropriate charge: thus, $\alpha = e^2/4\pi\epsilon_0\hbar c$. Figure 5 shows the way in which α increases, while α_s and α_w decrease, as the interparticle separation r decreases.

If things worked out in the way suggested by the figure, the basic strengths (as measured by these fine-structure constants) of all three different interactions could actually be the same at some appropriately small distance, r_u (the 'grand unification' scale).

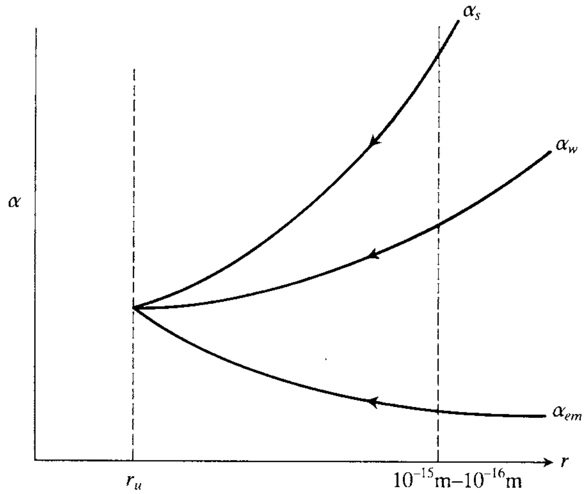


FIG. 5 Predicted convergence of the different interaction strengths at the very short distance r_u.
Source: Contemporary Physics 26: 333–91.

Actually, Fig. 5 is an oversimplification; but proper calculations can be done which lead to qualitatively the same conclusion. The predicted value of r_u comes out to be of the order of 10^{-31} m, a distance which, though apparently very tiny, is still uncomfortably large from one point of view. If, at this distance, all forces are equivalent, presumably the traditional distinction between (strongly interacting) quarks and (non-strongly interacting) leptons breaks down, and transitions can occur between quarks and leptons. These transitions would allow the proton to decay (e.g. to $\pi^0 e^+$)—yet it is known to be extremely stable. Of course, the $q \leftrightarrow l$ transitions occur only over the tiny volume within, say, a sphere of radius $\sim r_u$. Detailed calculations show that the proton lifetime goes roughly as $(r_p/r_u)^4$ where r_p is the proton radius. Though this leads to a very long lifetime ($\gtrsim 10^{31}$–10^{32} years), present limits on the observed proton lifetime are very hard to reconcile with the simplest grand unification schemes.

The above has illustrated the role of the quantum vacuum in uncovering a symmetry (at r_u) where none appeared to exist (at $r \sim 10^{-15}$–10^{-16} m). The second feature of the quantum field theory vacuum I need to recall is the way it can (perhaps!) account for a striking *lack* of symmetry.

The Vacuum as the Cause (or Explanation) of Asymmetry

Despite the possible equality (at r_u) in fundamental strengths of the weak, electromagnetic, and strong forces, there still remains a profound asymmetry between them: their different *ranges*. As is well known, in quantum field theory, the range R of a given force is inversely proportional to the mass M of the quantum of the associated force field: $R = \hbar/Mc$. As far as is known, the photon is massless, and the range of the electromagnetic force is infinite (in this sense). By contrast, the weak force quanta $W^{\pm}$, Z^0 have masses of the order of 90 GeV/c^2, with an associated force range $\sim 10^{-18}$ m. How can these forces possibly be unified? (We leave aside for the moment the question of the strong force.)

In more physical terms, this question can be rephrased as: Do we know of any circumstances in which the *photon behaves as if it had a mass*? Here we are engaged in fragile bridge-building, trying to reach a desirable unification aim with the aid of (possibly obsolescent) concrete physical analogies and models. The answer to this question is: Yes, in a superconductor. For the present purpose, the crucial property of a superconductor is the one revealed in the Meissner effect. If a substance that becomes superconducting at temperatures below the critical temperature T_c is placed in a static external applied magnetic field at a temperature $T > T_c$, the lines of magnetic flux will permeate the material; but as the specimen is cooled down below $T = T_c$ the flux is abruptly expelled from the interior of the sample. However, it is not wholly banished, but penetrates a characteristic distance λ in the surface layers of the material. In fact, the magnetic field varies approximately as $B = B_0\, e^{-x/\lambda}$ where x is the distance into the material measured normal the surface; λ is called the screening length (see Fig. 6).

Why does the phenomenon mean that the photon behaves as if it had mass? We can interpret things this way because the *range* of the B-field is exponentially attenuated in the superconductor; instead of an essentially infinite-range electromagnetic field, the B-field has range of order λ. Remembering the relation between 'range of force' and 'mass of quantum', we can say that, inside a superconductor at $T < T_c$, the photons in the B-field behave as if they had a mass $m = \hbar/\lambda c$.

Consider now a superconductor with two holes in it, such that lines of B-field pass through the holes (Fig. 7). The flux lines will each

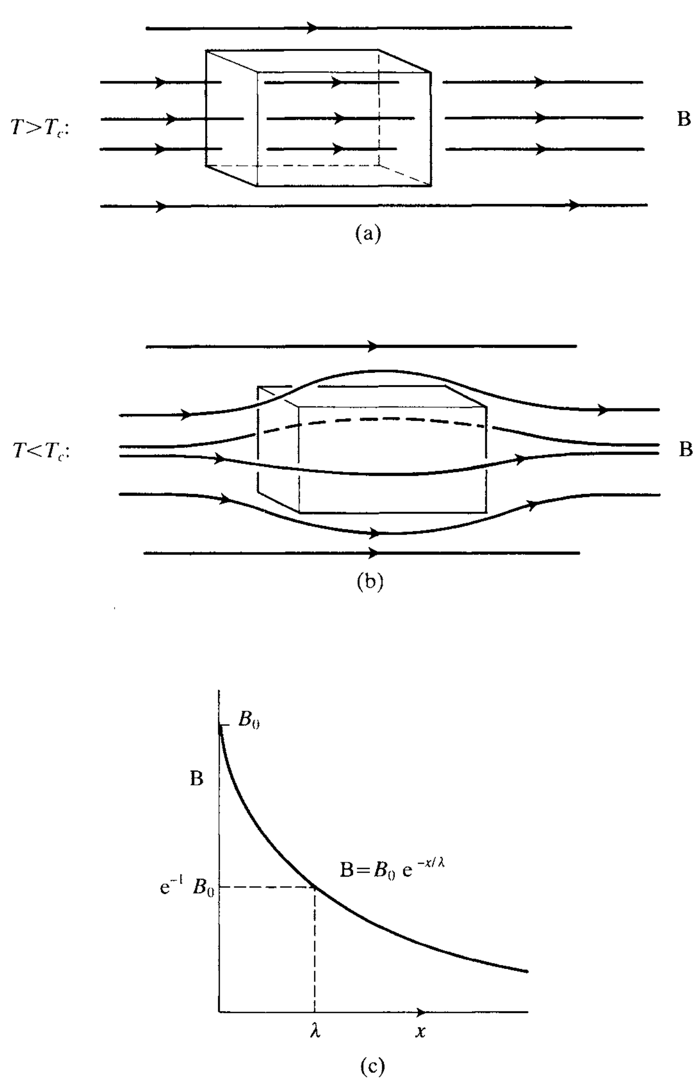


FIG. 6

(a) A superconductor in a field **B** at a temperature T above T_c; the field passes into and through the specimen.

(b) Below T_c, the **B** field is excluded from the interior of the specimen.

(c) Exponential penetration of the **B** field into the surface layers of the specimen, for $T<T_c$.

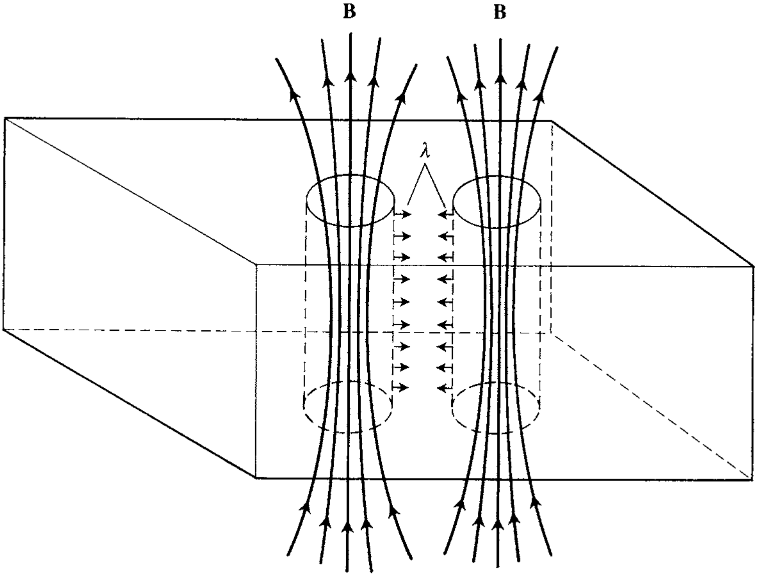


FIG. 7 Two flux bundles passing through holes in a superconductor, below T_c.

penetrate a small distance λ into the body of the superconductor, transverse to the walls of the cylindrical holes. If the distance between the holes is much greater than λ, there will be no mutual disturbance of the two flux bundles. But if the distance apart is of order λ, their 'leakages' into the superconductor will begin to overlap and affect each other: the bundles will *interact* via an effective force of range $\sim \lambda$.

This provides an analogy for the action of the weak force. We want it to 'penetrate' only a very short distance, $\sim \hbar/Mc$ (where M is the mass of the W or Z boson) into the physical vacuum. Thus, we shall suppose that the latter behaves like a *superconductor as far as the weak force is concerned.* The matter particles themselves, being regions of high concentration of weak charge (and weak field energy), may exist in 'pockets' of 'normal vacuum', in which the weak superconductivity has broken down (much as real superconductivity breaks down under a sufficiently strong applied external field). As a weakly interacting matter particle passes through the weak superconducting vacuum, it 'breaks down' the weak superconductivity up to a distance of order $\hbar/Mc$ around its trajectory; if two such particles pass within this sort of distance of each other, they will interact. Equivalently, we say the particles interact via the exchange of a W or Z quantum, of mass M (Fig. 8).

What kind of a mechanism could lie behind this conjectured behaviour? In a real superconductor, the reason the B field is zero in the interior of the sample is that, in the surface layer of thickness $\sim\lambda$, currents are set up which produce an internal B field opposing the applied field; and below T_c, this B field cancels the external one within the body of the superconductor. These *screening currents* are of course carried by the electrons in the material, so this phenomenon cannot be understood without reference to the charged current-carrying matter in the superconductor. Carrying over the same mechanism to the weak superconducting vacuum therefore requires the existence of a 'weak current-carrying matter field', which provides the necessary screening currents throughout the physical vacuum. This is a remarkable proposal, since we are now *not* talking about a fluctuation phenomenon; this matter field has got to be genuinely present in the vacuum. Normally it is assumed that the average value of all fields in the vacuum is zero: $\langle 0|\hat{\phi}|0\rangle = 0$, for all quantum fields. For the unification of QED and weak interactions to work, according to the above scheme, we have to reckon with a situation in which this is no longer true. The field (or fields) that have a non-zero vacuum value are called generically Higgs fields, after P. W. Higgs, one of the originators of this mechanism (Higgs 1964) whereby normally massless quanta acquire mass. The corresponding 'abnormal' vacuum $|\tilde{0}\rangle$ may be called a Higgs vacuum, and is such that $\langle\tilde{0}|\hat{\phi}|\tilde{0}\rangle \equiv f \neq 0$, for some Higgs field(s) $\hat{\phi}$.

There is actually nothing inherently unreasonable in the idea that the state of minimum energy (the vacuum) may be one in which some field quantity has a non-zero average value. Plenty of condensed-matter physics examples exist which display this feature—for example, the ferromagnet below its transition temperature, where there is a net alignment of the atomic spins. However, it remains a *conjecture* that something like this actually happens in the weak interaction case; at present, no dynamical basis for the Higgs mechanism exists, and it is purely phenomenological.

The form of the relation between a non-zero value of $f \equiv \langle\tilde{0}|\hat{\phi}|\tilde{0}\rangle$ and the mass of the force quantum is easily guessed on dimensional grounds. In units $\hbar = c = 1$, a boson field has the dimension of a mass. Also, if the weak charge g were 'switched off', we would expect the

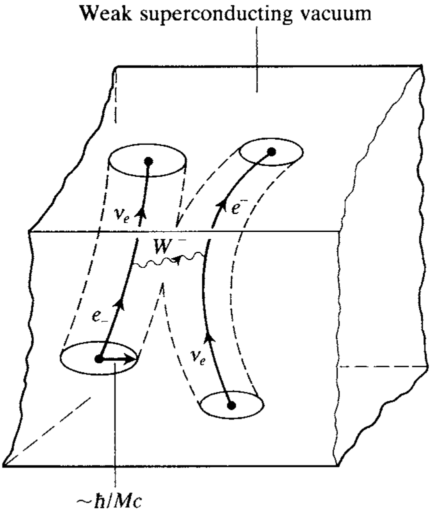


FIG. 8 A chunk of vacuum illustrating its 'superconducting' property with respect to the weak interaction.
Source: Contemporary Physics 26: 333–91.

effect to vanish. Thus we would guess that the effective mass of the W or Z quantum in the Higgs vacuum would be of order gf, which is correct. Inserting numerical factors, the known value of g, and the vector boson masses, one finds that a vacuum expectation value $f_W \approx 250$ GeV is required.

A Higgs vacuum must also be invoked in the context of the 'grand unification' of the three non-gravitational forces. In Fig. 5 we were led to the idea that the forces had a similar strength at distances of order $r_u \sim 10^{-31}$ m. As usual, we may translate this distance into an equivalent mass via $M_u = \hbar/r_u c$, which comes out at $M_u c^2 \sim 10^{15}$GeV. The grand unified force field will contain some quanta ('U') with this mass M_u. It is the exchange of these quanta, in fact, that can mediate proton decay. We require a second, independent, Higgs mechanism to generate mass for the ultra-heavy quanta. The corresponding f-value, f_u, is presumably of order 10^{15} GeV.

Finally, we note that yet another non-zero vacuum value appears to be required, associated with the dynamics of the strong force between quarks, described by the theory called quantum chromodynamics ('QCD'). Technically, this is somewhat different from the previous two examples, and is referred to as a 'Goldstone' rather than a 'Higgs' mechanism (see e.g. Aitchison 1982: ch. 6). The associated f-value is $f_\chi \sim 90$ MeV. We remark that this f-value does not imply anything about the range of the QCD force: it does not serve to give the gluons mass. It would take us too far afield here to explain why this is so—and why, in fact, the strong force apparently has the remarkable property of forbidding the existence of free isolated quarks ('confinement'). Nevertheless, this latter property can probably again be understood in terms of vacuum structure (see Section 6.4 of Aitchison 1985). It is this (QCD) vacuum that Fig. 4 is supposed to represent—Maxwell's vortices come around again!

After all this, I need hardly emphasize the centrality of hypotheses about the vacuum to the whole edifice of unification of the non-gravitational forces. Are the above pictures models, or are they true? As always, and as Maxwell reminds us, experiment must be the judge. The average values f of Higgs fields are constants—but these quantum fields can presumably also support excitations, as usual, which correspond to massive quanta ('Higgs bosons'). Very little is known about their mass, or masses, but they may be in the region from about 100 GeV to 1 TeV. The discovery of a Higgs boson—a quantum vibration of the Higgs vacuum—would be of enormous importance. The study of its properties should provide clues to the *real* mechanism of mass-generation.

4 Two Problems with the Story so Far

Certain rather elaborate properties of the vacuum have been posited, which allow unification of electromagnetic, weak, and strong forces. In this section I note two problems with this general picture, which themselves indicate (possibly) the need for further unification.

The Hierarchy Problem

This problem relates to the two very different values of the gauge boson masses introduced in the unification of the electroweak interactions ($M_W, M_Z \sim 90$ GeV/c^2) on the one hand, and of all three non-gravitational interactions ($M_u \sim 10^{15}$ GeV/c^2) on the other. (Equivalently, we note the disparity between $f_W \sim 250$ GeV and $f_u \sim 10^{15}$ GeV.) It seems, in the first place, most implausible that, as Fig. 5 in fact presupposes, one would encounter no new physics at all as one passed through the 13 orders of magnitude separating these energies (or, in Fig. 5, distances). Yet this is not the real problem. Figure 5 refers to the behaviour of the fine-structure constants of the theories, as a consequence of vacuum polarization effects. These effects turn out to induce a *logarithmic* variation of the 'constants' with distance scale (or, equivalently, energy scale). That is why the convergence of the three constants to a unified value takes so many orders of magnitude in scale to achieve. However, vacuum fluctuations have another effect: they can cause shifts in the *masses* of particles, much as an electron, moving through a crystal, generally behaves as if it had an 'effective' mass different from its mass in free space. These mass shifts depend much more strongly on the distance (or energy) scale than do the fine-structure constants. In fact, the mass shifts will generally be of the same order as the energy scale (not its logarithm). Here is where the problem arises: mass shifts induced by grand unified interactions will inevitably tend to pull all masses (e.g. those of W and Z bosons) up to that of the 'grand unified' bosons, M_u. How can the hierarchy of scales be maintained?

The gauge boson masses themselves arise from non-zero vacuum values of Higgs fields. Thus, to explore the problem we really have to consider the separate electroweak and grand unified Higgs fields, which are denoted generically by $\hat{H}$ and $\hat{\Sigma}$ respectively, where $\langle\tilde{0}|\hat{H}|\tilde{0}\rangle = \mathrm{f_W} \approx 250$ GeV and $\langle\tilde{0}|\hat{\Sigma}|\tilde{0}\rangle = f_u \approx 10^{15}$ GeV. We might imagine 'insulating' the $\hat{H}$ from the $\hat{\Sigma}$ in the assumed fundamental interactions of the theory, by forbidding all couplings between them. In this way, the two widely different energy scales would have no 'communication' with each other. But such couplings will actually be induced by perturbative corrections (see e.g. Ross 1985: 181 ff), which *require* the existence of fundamental $\hat{H}$–$\hat{\Sigma}$ couplings if the theory is to be renormalizable. These couplings are typically of the form $\hat{H}^2\hat{\Sigma}^2$: and when $\hat{\Sigma}$ is replaced by its vacuum value in this expression, we see (recalling that the mass term for a scalar field is $m^2\hat{\phi}^2$) that contributions to the $\hat{H}$ mass are inevitably generated, which are of order 10^{15} GeV/c^2. In turn, this mass will feed into a huge mass correction for the W and Z particles.

Of course, matters are more complicated than sketched here. In particular, there are various terms in the '$\hat{H}^2\hat{\Sigma}^2$' potential, and perhaps cancellation can be arranged between them, and/or between them and a 'bare' mass term—but these cancellations have to be repeated in each separate order of perturbative correction.

If we regard such delicate cancellations (two numbers of order

10^{15} GeV/c^2 conspiring to produce one of order 10^2 GeV/c^2) as unsatisfactory in a fundamental theory, we have a problem: the 'hierarchy problem'. Two ways out have been proposed. One is to regard the electroweak Higgs field, at least, as not elementary, but as *composite*, i.e. as a bound state of some as yet unknown matter-fields. It is then no longer point-like, but has a characteristic size, which provides a physical cut-off to the scale of corrections to its mass (Ross 1985: 279 ff). If this size is of order $\gtrsim \hbar/(1\ \text{TeV/c})$, then all perturbative mass shifts will be of order $\lesssim 1$ TeV/c^2, and no unnatural cancellations will have to be appealed to in holding the Higgs mass, and vector–boson masses, at the same level, i.e. $\gtrsim 1$ TeV/c^2. This hypothesis (of compositeness) is actually very analogous to the 'superconductor' picture sketched in the previous section: the Higgs field for a real superconductor is precisely not an elementary field, but a bound state of two electrons (a 'Cooper pair'). It is therefore a rather attractive proposal. It is also phenomenologically exciting, since it seems clear that, if true, it would imply the existence of a rich new *spectroscopy* of bound states in the TeV region, such as could be found by future accelerators. Unfortunately, however, there are technical problems with the detailed implementation of the idea (Ross 1985: 280 ff), and no fully acceptable model along these lines (generally called 'technicolour theories') seems to exist.

The other strategy is to arrange for the perturbative corrections, which generated the necessity for the undesirable $\hat{H}^2\hat{\Sigma}^2$ potential, to cancel among themselves. This is a very radical suggestion, since some of these corrections involve bosons, and others fermions, in the intermediate states (in the sense of quantum mechanical perturbation theory). But it is precisely this difference which, qualitatively, suggests the possibility of a cancellation: in the language of Feynman graphs, boson loops and fermion loops enter with opposite signs! But more than this sign difference is needed: clearly, the boson and fermion coupling constants must also be all interrelated. In fact, there has to be a *supersymmetry* (Ross 1985: sec. 10.5) between the bosons and fermions in the theory. Thus, our search for unification of the force, and attendant problems consequent upon proposed vacuum structure, have led—in one scenario at least—to the possibility of an even more profound unification.

Of course, the snag with this is that we do not observe *any* symmetry between fermions and bosons! Such a symmetry would imply, for instance, the existence of a spinless particle degenerate in mass with the electron—indeed, to every known 'elementary particle' there should correspond a degenerate 'superparticle' of appropriately different spin. Nothing like this, of course, is seen; hence supersymmetry has to be a 'broken' symmetry, the superparticles somehow acquiring different masses from the normal particles. However, in that case the delicate cancellation between boson and fermion loops, which solved the hierarchy problem, is lost. Thus, if this solution to the problem is to be preserved, the mass of the superparticles cannot be too high—in fact, probably not much more than about 1 TeV/c^2. It seems, therefore, that on either 'leg' of these two responses to the hierarchy problem, one must anticipate new physics around the 1 TeV energy region, which is a remarkable prediction.

We turn now to another problem associated with Higgs-type vacua.

The Cosmological Constant Problem

This is an even worse 'fine-tuning' problem. The difficulty arises from the *vacuum energy density* which is carried by non-vanishing Higgs fields. In a normal vacuum, almost by definition, the energy density is zero and all fields have zero average values. But in a Higgs (or Goldstone) vacuum there are some fields $\hat{\phi}$ with constant remnant values, which we have denoted generically by the symbol f. Such constants carry no kinetic energy, but they will in general contribute a term $V(f)$ to the potential energy of the vacuum. On dimensional grounds, we expect this vacuum potential energy density to be of order f^4. (Any other widely different value would necessitate the introduction of a new dimensionless fundamental constant.) This energy density is presumed constant throughout all of space—and it would have remarkable cosmological implications (Linde 1974); Ross 1985: sec. 12.7), as we now discuss.

Assuming a homogeneous and isotropic expanding universe, the metric following from Einstein's equations of General Relativity (GR) is that introduced by Robertson and Walker, which involves the quantity $R(t)$, a time-dependent factor setting the scale of the universe. For zero curvature, the dynamical equations of GR give

$$\left(\frac{\dot{R}}{R}\right)^2=\frac{8\pi}{3}G_N\rho \tag{1}$$

where G_N is Newton's constant ($\sim 10^{-38}$ GeV^{-2} in units $\hbar = c = 1$), and ρ is the energy density. Generally ρ is considered to be either radiation- or matter-dominated. In particular, it has no constant part, which would correspond to a 'cosmological constant' term. To see the effect of such a term, suppose ρ is a constant and set

$$\frac{8\pi}{3} G_N \rho = \wedge \, . \tag{2}$$

Then it is easy to see from (1) that the scale size grows *exponentially* with time, as

$$R(t) = R_0 \exp[\wedge^{1/2}(t - t_0)]. \tag{3}$$

Current observations are consistent with matter-dominated expansion, which grows as $t^{2/3}$. This gives a very severe bound on the possible magnitude of $\wedge$:

$$\wedge^{1/2} \lesssim 10^{-42} \text{ GeV}. \tag{4}$$

On the other hand, we have just argued that a Higgs vacuum leads precisely to a constant ρ, of order f^4, where f is order 10^{15} GeV for grand unification, and 250 GeV for electroweak unification. From (2), these give

$$\wedge_{gu}^{1/2} \sim 10^{11} \text{ GeV}. \tag{5}$$

$$\wedge_{ew}^{1/2} \sim 10^{-14} \text{ GeV}. \tag{6}$$

Even (6) is 28 orders of magnitude away from (3).

Of course, this vacuum potential energy density is a constant, and we might feel that we are free to cancel it away by subtracting the appropriate constant value from V. But again, this is another (and worse) fine-tuning problem.

Actually, exponential growth in R—which is called *inflation*—is believed to be a highly desirable feature in the very early stages of the evolution of the universe (see e.g. Ross 1985: sec. 12.8; Guth and Steinhardt 1984; Aitchison 1985: sec. 6.4.4). A typical value of f required (via $\rho \sim f^4$) for this very early inflation is of order 10^{16} GeV, not very different from the 'grand unified' f_u. Perhaps this is significant.

At all events, the fact remains that the vacuum value of V had better be pretty nearly zero *now*. The idea that the value of the (effective) potential energy can change as a system evolves is familiar in a condensed matter context. A system may cool down, for example, so as to pass through a critical temperature where a *phase transition* occurs. Such a change in phase can be modelled in terms of an appropriately chosen V. Applying this idea to the expanding and cooling universe, one can arrive at a possible form of potential $\hat{V}(\hat{\phi})$ shown in Fig. 9—where only one scalar field $\hat{\phi}$ is in play. We imagine

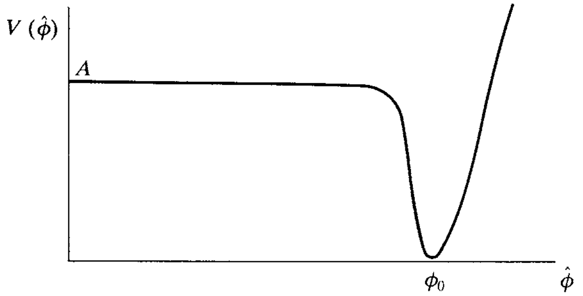


FIG. 9 Potential energy density versus Higgs field, modelling slow 'roll-over' phase transition.

that at $t \approx 0$ the field system starts at A near the origin, with a value of $\hat{V}$ appropriate to early inflation. Then the field 'rolls' slowly down the shallowly sloping part of $\hat{V}$ (the universe inflating meanwhile), before making a sharp transition to the equilibrium value ϕ_0, for which $\hat{V}(\phi_0)$ (and hence $\wedge$) is zero. Note that ϕ_0 is non-zero, and could perhaps be identified with f_u.

Supersymmetry may be able to help here, once again, since it turns out that in a (globally) supersymmetric theory the value of the potential, and hence of $\wedge$, has to be zero at its minimum (cf. the Abstract in Fig. 2). But of course, supersymmetry cannot be exact, as we have seen—and if the scale of its breaking is ~ 1 TeV we are back with a gross violation of (3).

Even if such ideas can reconcile the Higgs mechanism with cosmology, there will remain the energy density associated with the QCD (Goldstone) vacuum parameter f_χ (Section 3). It is very hard to see what possible principle can arrange for this energy to be zero 'naturally'. Are we here beginning to face problems akin to the absurd 'rigidity' of the nineteenth-century ether? There is little doubt that Higgs (and Goldstone) fields are our modern equivalents of those earlier mechanical contrivances populating the plenum, albeit very subtly. As in those days, we must push forward with such concepts as we have, until perhaps some new Einstein appears to explain why it is really not necessary to look at things this way at all.

The discussion has now brought us most of the way down the first two columns of Fig. 1, and gravity and cosmology have been mentioned; it is time to tackle the third column.

5 Kaluza–Klein Theories

It is obvious that I cannot possibly provide here an adequate introduction to this topic: once again, I want to direct my remarks at the possible role of the vacuum in this type of geometrical unification. An excellent recent compendium is provided by Appelquist *et al.* (1987); despite its title, this volume includes many historical papers, including those by Kaluza and Klein.

Classical general relativity employs ten basic fields, the components of the metric $g_{\mu\nu}^{(x)}$, where x stands for (x^0, x^1, x^2, x^3). The Einstein–Hilbert action, which leads to the field equations of GR, is

$$I = -\frac{1}{16\pi G_N}\int d^4x\sqrt{-g}\,\mathscr{R} \tag{7}$$

where $g = \det g_{\mu\nu}$, $\mathscr{R}$ is the scalar curvature, and a possible cosmological constant term (which would add $2\wedge$ to $\mathscr{R}$) is omitted. On dimensional grounds (with $\hbar = c = 1$), we see that G_N has dimension $(\text{mass})^{-2}$: in fact, we already gave the value $G_N \cong 10^{-38}\ \text{GeV}^{-2}$. It is interesting to compare this situation with that in the weak interaction; there, Fermi's original theory employed a basic strength parameter G_F which also carried dimension $(\text{mass})^{-2}$. Today, we know that Fermi's constant is not really fundamental, and that its 'dimensionfulness' originates from the mass of the W bosons: indeed, the electroweak theory gives $G_F = \sqrt{2}\, g^2/8{M_W}^2$, where g is the dimensionless weak charge. Could the dimensions of Newton's gravitational constant be explained in a similar way? In other words, could we get a theory of gravity characterized by a fundamental mass (or length) and a dimensionless strength? Could we then unify *all* the forces?

It is clear, first of all, that we do *not* want to involve any kind of Higgs mechanism to generate the mass scale ($\simeq 10^{19}$ GeV) associated with G_N. (By the way, this mass scale is of course the *Planck mass*,

M_P.) To do so would surely imply that gravitational forces extend only over distances of order $\hbar/(10^{19}\ \text{GeV/c}) \simeq 10^{-35}$ m (the *Planck length*, L_P), very far from the truth! Furthermore, if $g_{\mu\nu}(x)$ is indeed the fundamental field of gravity, the quanta of its quantized version $\hat{g}_{\mu\nu}(x)$ have spin 2, and this seems quite different from the spins of the quanta of the other force-fields. (But could the vacuum somehow reconcile even these differences?) These considerations seem to rule out any obvious inclusion of the gravity fields in the unification framework as so far developed: something new is needed.

Given that we abandon Higgs, where can a gravitational length (L_P) or mass (M_P) scale come from? We could suppose that this is indeed a natural length scale associated with the most fundamental objects of all (as is done in string theory, Section 6). Alternatively, we might suppose—having heard of the Kaluza–Klein idea—that such a length might represent the effective size of a *highly compactified fifth dimension.*

The introduction of a fifth coordinate dimension does allow a geometrical unification of gravitation and electromagnetism, as Kaluza showed. The metric field g_{MN} $(x^0, x^1, x^2, x^3, x^4)$ now has 15 components. These can be regarded as comp.:ising 10 for the standard $g_{\mu\nu}$ of GR, 4 for the vector potential A^μ, and 1 scalar field σ. There is much more to this than mere arithmetic. First, a special coordinate transformation involving the 'fifth' coordinate x^4 turns out to correspond to a gauge transformation on A^μ—allowing the latter to be interpreted geometrically. Second, the five-dimensional version of the action (6), when expanded in terms of $g_{\mu\nu}$, A^μ, and σ, contains the four-dimensional action (6), a Maxwell action for A^μ, and the usual kinetic term for σ, provided that all dependence of these fields on the fifth coordinate x^4 is dropped. (Kaluza also, in fact, assumed that σ was a constant.) Thus, a remarkably unified geometrical picture emerges.

Why can the dependence on x^4 be ignored? Assuming that the dimension is compactified to an S^1 of radius R, we can Fourier analyse the fields according to

$$F(x, x^4) = \sum_n F^{(n)}(x)\, e^{inx^4/R} \tag{8}$$

where F is $g_{\mu\nu}$, A_μ, or σ and x is (x^0, x^1, x^2, x^3). In the quantum version of this theory, we interpret n/R as a momentum-like quantity, and expect that the $n \neq 0$ harmonics in (8) correspond to excitations of mass n/R. If R is of order the Planck length, these are of order the Planck mass, and thus are totally ignorable at conventional energy scales.

The idea was carried further by Klein in 1926, who related the electromagnetic charge to Newton's constant and R. In modern terms, we would say that, since transformations in the fifth coordinate are associated with gauge transformations on the vector potential A^{μ}, the charges on the $n \neq 0$ modes in (7) must be determined. Indeed, the nth harmonic carrries a charge

$$e_n = n(16\pi G_N)^{1/2}/R; \tag{9}$$

the charge is quantized because the fifth dimension is compact (n is integer to ensure periodicity in (7)), the elementary unit being $(16\pi G_N)^{1/2}/R$. The corresponding fine-structure constant is $4G_N/R$, so that if we identify this with 1/137, $R \sim 100L_P \sim 10^{-33}$ m.

All this is very liberating—but what mechanism can we imagine for the required compactification of the fifth dimension? It is possible that vacuum dynamics can supply the answer. The essential idea is a variant of the *Casimir effect* (Aitchison 1985: sect. 3.2). In the simplest version of this effect, we consider the change in the *zero-point energy* ($\sim \Sigma_k \frac{1}{2}\hbar\omega_k$) of the radiation field when we simply position two large conducting plates a distance d apart *in vacuo*. The zero-point energy (z.p.e.) can be regarded as being associated with vacuum fluctuations in the fields. The introduction of the plates forces the fields to have nodes on the surfaces of the plates; thus the wave numbers, and hence the frequencies of the allowed modes, will depend on the separation d. Changing d will change the z.p.e. by a finite amount. In physical terms, this means that work has to be done to establish the required boundary conditions when the plates are moved. This work can be regarded as a potential energy function which depends on d. In the present case, $V(d)$ is negative, indicating an attractive force between the plates (which has been measured). Instead of this 'plate' geometry, we might consider the effect of introducing a spherical shell, of radius r. In the electromagnetic case this produces a positive $V(r)$—so that it cannot, for example, cancel repulsive Coulomb forces, were we to regard the system as a model for the electron, say (as Casimir himself originally hoped it might be). Suppose, however, we consider a gravitational analogue, and interpret r as the radius of the Kaluza–Klein (K–K) compactified fifth dimension. Vacuum fluctuations of the gravitational fields, subject to the boundary conditions at $x^4 = r$, will produce an effective potential $V(r)$, which might indicate a tendency—on energetic grounds—for r to decrease (compactify) 'spontaneously'.

The calculation of $V(r)$ in a 5-dimensional Kaluza–Klein (quantized) gravity theory was done by Applequist and Chodos (1983), to one loop order. Subtracting off a constant part (corresponding to an—infinite!—cosmological constant), they found that V, the change in the z.p.e. per unit three-dimensional volume, was negative and proportional to r^{-4}.

Thus, according to this calculation, there *is* a dynamical tendency for the fifth coordinate to shrink. Unfortunately, it collapses indefinitely—but the calculation is not reliable as soon as r is order L_P (any more than the electromagnetic one would be for the plate separations of order an atomic spacing). There would be a large prize attached to a successful calculation of the *equilibrium* value $r = R$ (if such exists), since from (9) this would be tantamount to a calculation of the electromagnetic charge! A number of proposals were made in order to achieve stability. For example, Rubin and Roth (1983*a*) discussed whether, at high enough temperatures, thermal pressure effects could counterbalance collapse; and Rubin and Roth (1983*b*) and Tsokos (1983) noted that the addition of massive fermions can change the sign of V at short distances.

But of course, the simple *K*–*K* model, in which only electromagnetism is considered, apart from gravity, is not satisfactory. To interpret the weak and strong gauge forces in terms of compactified geometry, we must go to more dimensions than five. The standard model $SU(3)_c \times SU(2)_L \times U(1)_y$ needs at least an additional *seven* dimensions, beyond our own M^4. The extra dimensions constitute what is called the 'internal manifold', since they are the geometrical origin of the 'internal' symmetries. A significant new feature now appears, as soon as the internal manifold has more than one dimension: it will in general have *curvature*. Candelas and Weinberg (1984) exploited the tendency of the (positive) energy associated with such a curvature to produce stability against collapse. They considered manifolds of the form $M^4 \times S^N$, where S^N is an N-sphere and N is odd. They included both bosonic and fermionic matter-fields, and were able to find a V with a stable minimum. Unfortunately, about 10^4 or 10^5 different species of matter-fields were required to produce gauge couplings (via the generalization of (9)) of the right order of magnitude. (Further possibilities along these lines are collected in Appelquist *et al.* 1987.)

There is, however, something of a conceptual problem associated with having an internal manifold with curvature, such as an S^N. The trouble is that oppositely-handed versions of our low-energy leptons would then exist—which, since they have not been observed, must be presumed to be very massive. Now fermions of only one handedness (or 'chirality') can introduce certain diseases into the theory, called 'anomalies'. It is believed that anomalies have to be cancelled for the theory to make sense. Indeed, the anomalies associated with leptons of *both* chiralities would cancel each other out. The snag is that the anomalies due to our familiar low-energy leptons are *already* cancelled by corresponding contributions from the quarks. This is surely too big a coincidence to be accidental—though exactly what its significance is remains obscure. At any rate, the heavy leptons of opposite chirality are not needed, and it is believed that the fermion spectrum in higher-dimensional theories should be chiral (i.e. of one handedness only).

An interesting way out recalls the Cremmer–Scherk (1976) model of spontaneous compactification. If additional non-Abelian gauge fields *in a monopole configuration* are added as a background to the higher-dimensional theory, then the index theorem (which relates chirality to topological charge) allows the fermion spectrum to be chiral (Ranjbar-Daemi *et al.*, 1983).

This suggestion naturally raises questions. Where do these background gauge fields come from, and why are they automatically in a monopole configuration? Is such a configuration *energetically preferred*? Here I cannot resist throwing out a no doubt wild idea. A remarkable new type of gauge-field structure, in ordinary quantum mechanics, has recently been discovered by Berry (1984). It occurs in theories in which the dynamical degrees of freedom separate into two classes—'fast' and 'slow'—distinguished by their time rates of change, or equivalently by the characteristic energy level differences for the two types of excitation; one thinks, for example, of electronic and nuclear degrees of freedom in a molecule. In one formulation of such a situation, the 'fast' degrees of freedom (which of course are coupled to the 'slow' ones) are integrated out, and an effective Lagrangian (or Hamiltonian) for the dynamics of the slow coordinates is obtained. This Lagrangian includes, quite characteristically, a gauge-field contribution which is induced by the process of integrating out the fast coordinates (Wilczek and Zee 1984; Li 1987*a*, 1987*b*; Kuratsuji and Ida 1985*a*, 1985*b*, 1987; a recent review is provided in Aitchison 1987). Even more remarkable, this gauge field frequently appears precisely in a monopole-like configuration!

It is tempting to wonder whether the coordinates x^0, x^1, x^2, x^3 of our own space–time could be regarded as 'slow' in the above sense, while some of those associated with a compactified internal manifold $x^4, x^5, \ldots$ would be 'fast'. Integrating fields over the fast internal coordinates might then induce an effective monopole-like gauge structure, which would allow chiral fermion multiplets. A simple model somewhat along these lines has been suggested by Duncan and Segrè (1987). Of course, we have to explain why some compactification occurs, as usual; maybe there is some kind of self-consistency here, since the Cremmer–Sherk mechanism will come into play once background monopole fields exist, and the Berry phenomenon may generate such fields if some internal dimensions were compactified. To begin to address such issues would seem to require a generalization of the Appelquist–Chodos z.p.e. calculation, to allow for Berry-type phenomena.

Though a slight detour, it is worth mentioning here another application of the 'adiabatic' approximation in a vacuum (specifically, Casimir energy) context. Instead of starting with a five-dimensional Einstein action, as in the original *K–K* approach, one could consider a theory *without* any explicit gravitational action, consisting only of matter fields in a (five-dimensional) space–time background. A simple scalar field Lagrangian would then involve a kinetic piece $g^{MN}\,\partial_M\hat{\phi}\,\partial_N\hat{\phi}$, for example, where g^{MN} is the 5-D metric. Integrating out the fifth coordinate yields an effective action which is essentially the Casimir energy (Myers 1987), and which can be expanded in powers of derivatives acting on the fields—which we know to include the 4-D $g^{\mu\nu}$ and a Maxwell field A^{μ}. The quantity setting the scale in this expansion is the radius R of the compactified dimension—or, equivalently, the associated mass $1/R$. Thus, the expansion is in powers of $(R\partial)$ where ∂ is a typical field gradient. 'Adiabaticity' in this context means that $\partial \ll R^{-1}$, or that the field gradients are small compared with R^{-1}. In this case only the first few terms in the expansion need be considered.

Toms (1983) did such a calculation for a 5-D theory with scalar and spinor fields. It is clear that the expansion of the effective action, in powers of derivatives, must yield a sequence of terms the form of which is severely constrained, *a priori*, by general covariance. In fact, the lowest-order term is a cosmological constant, and the next is an Einstein curvature term together with a Maxwell action (both second-order in derivatives). Thus, a 4-D gravitational action has been 'adiabatically induced'—and its contribution is the natural geometric quantity, the curvature of the parameter manifold. The coefficient of the Einstein term has to be related to G_N; and this fixes the magnitude of the radius R to be (Toms 1983) some 100 Planck lengths.

The idea was pushed to higher dimensions by Awada and Toms (1984), who showed that the 4-D Einstein–Yang–Mills action could be induced, in adiabatic approximation, from an effective (Casimir) action obtained by integrating out higher dimensions with a non-Abelian K–K metric. In all cases, one generates—as usual!—an unacceptably large cosmological constant term.

Returning to the main line of historical development, we have reached, in the era of supergravity and non-Abelian K–K theories, roughly the years 1983–4. At that point, a paper by Green and Schwartz (1984) appeared, which triggered an enormous switch of emphasis towards *(super)string* theories. They showed how to construct, via a superstring theory, a quantum theory of gravity which was free of (gravitational) anomalies, the latter being cancelled by anomalies associated with the *other* interactions, provided the internal gauge group was SO(32). This was a quite remarkable 'reason' for having all the forces!

String theories abandon the 'point particle and local field' framework that has served us since Maxwell's time. They postulate as fundamental objects things with *extension*—in the case of strings, extension in effectively one dimension. Thus, in place of the space–time trajectory or world-line of a point particle, we have a world-sheet of a string. Naturally, the extension had better be quite small—and it is indeed supposed to be of the order of the Planck length L_P. This implies, as we have seen in other contexts, that excitations of the quantized string will have a mass $\sim M_P$. Consequently only the ground—or *vacuum*—state of the string is likely to come in (though the excited states play a role in anomaly cancellation).

Bosonic strings can be consistently formulated only in 26 dimensions; supersymmetric strings can get by with 10. The latter (via the Green–Schwartz mechanism) allow for consistent quantization of gravity; indeed, they may well be free of all divergences. Soon after the Green–Schwartz discovery, a new kind of superstring theory was found, the 'heterotic' type (Gross *et al.* 1985; Freund 1985). This too successfully describes a quantum gravity, and has an alternative internal gauge group, $E_8 \times E_8$. It is this type that was particularly favoured by the analysis of the paper in Fig. 2, since it can be compactified so as to yield a four-dimensional world not unlike our own.

String theory is developing extremely rapidly, and is by no means a fully deductive theory as yet. In particular, no complete 'second-quantized' version of the theory is yet to hand. As one way of understanding what is meant by this, consider the path-integral method of quantization for a point particle. One passes, characteristically, from a picture in which a classical particle moves from A to B along one specific (classical) path, to one in which one has to sum over *all* paths, weighted by the factor $e^{iS/\hbar}$, where S is the action. By this process one effectively moves from the classical variables $x(t)$ to the quantum operator $\hat{x}(t)$. For a string, the place of the classical path is taken by the world-sheet. To 'first-quantize' such a theory, one has to imagine summing over all possible world-sheet configurations, beginning with a stated initial one and ending with a final one. By this process, one passes from the classical string coordinate $x_\mu(\sigma, \tau)$ to the quantum operator $\hat{x}_\mu(\sigma, \tau)$—here σ and τ label the world-sheet and μ runs over the 26 dimensions, for a bosonic string. What about 'second quantization'? In ordinary quantum field theory, one introduces field operators $\hat{\phi}(x)$—or, equivalently, one works with generating functionals which involve integrating over all possible field configurations, with some measure $D_\phi(x)$. In string theory the analogue of the latter step would seem to be some integration of the form $D\Phi[x_\mu(\sigma, \tau)]$, where the fields are defined with string arguments. The precise form of this measure is not known, at least for interacting strings.

Until we have an analogue of the 'Feynman rules' for string theories, we cannot answer some obvious questions that arise in connection with superstring-inspired unification. In particular, even assuming we had one unique string theory, we need to explain *why* the low-energy world in which we actually live is its ground state. In other words, why is the compactification scheme found by the authors of Fig. 2, say, energetically preferred? There are very many solutions of the classical field equations of superstring theories—all of which are candidate ground states, or vacua.

All the same, it is pretty mind-blowing to reflect that our own four-dimensional world, its geometry, its matter content, and its force, may all be just the dynamically determined ground state or *vacuum*—of a superstring! Vacuum dynamics, and unification, can hardly be taken further.

References

Aitchison, I. J. R. (1982), *An Informal Introduction to Gauge Field Theories*, Cambridge University Press.

—— (1985), 'Nothing's Plenty', *Cont. Phys.* 26: 333–91.

—— (1988), 'Berry's Topological Phase in Quantum Mechanics and Quantum Field Theory, *Physica Scripta*, T23: 12–20.

Applequist, T and Chodos, A. (1983), 'Quantum Dynamics of Kaluza–Klein Theories', *Phys. Rev.* D28: 772–84.

——, ——, and Freund, P. G. O. (1987), *Modern Kaluza–Klein Theories*, Addison-Wesley, Reading, Mass.

Awada, M. A. and Thoms, D. J. (1984), 'Induced Einstein–Hilbert–Yang–Mills Action from Quantum Effects in Generalized Kaluza–Klein Theory', *Phys. Lett.* 135B: 283–7.

Berry, M. V. (1984), 'Quantal Phase Factors Accompanying Adiabatic Changes', *Proc. Roy. Soc. London* A392: 45–57.

Candelas, P. and Weinberg, S. (1984), 'Calculation of Gauge Couplings and Compact Circumferences from Self-consistent Dimensional Reduction', *Nucl. Phys.* B237: 396–441.

Cremmer, E., Julia, B., and Scherk, J. (1978), 'Supergravity Theory in 11 Dimensions', *Phys. Lett.* 76B: 409–12.

—— and Scherk, J. (1976), 'Spontaneous Compactification of Space in an Einstein–Yang–Mills–Higgs Model', *Nucl. Phys.* B108: 409–16.

Davies, P. C. W. (1975), 'Scalar Particle Production in Schwarzschild and Rindler Metrics', *J. Phys.* A8: 609–16.

Duncan, M. J. and Segrè, G. (1987), 'A Simplified Model for Superstring Compactification', *Phys. Lett.* 195B: 36–40.

Freund, P. G. O. (1985), 'Superstrings from 26 Dimensions?' *Phys. Lett.* 151B: 387–90.

—— and Rubin, M. A. (1980), 'Dynamics of Dimensional Reduction', *Phys. Lett.* 97B: 233–5.

Fulling, S. A. (1973), 'Non-uniqueness of Canonical Field Quantization in Reimannian Space–Time', *Phys. Rev.* D7: 2850–62.

Green, M. B. and Schwartz, J. H. (1984), 'Anomaly Cancellations in Supersymmetric D = 10 Gauge Theory and Superstring Theory', *Phys. Lett.* 149B: 117–22.

Gross, D. J., Harvey, J., Martinec, E., and Rohn, R. (1985), 'Heterotic String', *Phys. Rev. Lett.* 54: 502–5.

—— and Wilczek, F. (1973), 'Ultraviolet Behaviour of Non-Abelian Gauge Theories', *Phys. Rev. Lett.* 30: 1343–6.

Guth, A. H. and Steinhardt, P. J. (1984), 'The Inflationary Universe', *Sci. Amer.* 250 (5): 90–102.

Higgs, P. W. (1964), 'Broken Symmetries and the Mass of Gauge Bosons', *Phys. Rev. Lett.* 13: 508–9.

Kaluza, K. (1921), 'Zum Unitätsproblem der Physik', *Sitzungsber. d. Berl. Akad.*: 966–73. An English translation appears in Appelquist *et al.* (1983), pp. 61–8, with the title 'On the Unity Problem of Physics'.

Klein, O. (1926), 'Quantum Theory and Five-dimensional Theory of Relativity', *Z. f. Phys.* 37: 895–904.

Kuratsuji, H. and Iida, S. (1985*a*), 'Semiclassical Quantization with a Quantum Adiabatic Phase', *Phys. Lett.* 11A: 220–2.

—— (1985*b*), 'Effective Action for Adiabatic Process', *Prog. Theor. Phys.* 74: 439–45.

—— (1987*a*), 'Geometrical Meaning of Quantum Adiabatic Phase: The Case of a Non-canonical System', *Phys. Rev. Lett.* 56: 1003–6.

—— 1987*b*), 'Adiabatic Theorem and Anomalous Commutators', *Phys. Lett.* 184B: 242–6.

Lu, H.-Z. (1987*a*), 'Induced Gauge Fields in a Non-gauged Quantum System', *Phys. Rev. Lett.* 58: 539–42.

—— (1987*b*), 'Conversion of Isospin to Spin Degrees of Freedom: Induced Monopole Fields in Adiabatic Processes', *Phys. Rev.* D35: 2615–18.

Luciani, J. F. (1978), 'Space–Time Geometry and Symmetry Breaking', *Nucl. Phys.* B135: 111–30.

Linde, A. D. (1974), 'Is the Lee Constant a Cosmological Constant?' *J.E.T.P. Lett.* 19: 183–4.

Maxwell, J. C. (1861*a*), 'On Physical Lines of Force. Part I: The Theory of Molecular Vortices applied to Magnetic Phenomena', *Phil. Mag.* 21: 162–75.

—— (1861*b*), 'On Physical Lines of Force. Part II: The Theory of Molecular Vortices applied to Electric Currents'. *Phil. Mag.* 21: 281–91, and 338–48.

—— (1862), 'On Physical Lines of Force. Part III: The Theory of Molecular Vortices Applied to Statical Eelectricity', *Phil. Mag.* 22: 12–24.

—— (1865), 'A Dynamical Theory of the Electromagnetic Field', *Phil. Trans. Roy. Soc.* 155: 459–512.

—— (1890), 'On Action at a Distance', in W. D. Niven (ed.), *The Scientific Works of James Clerk Maxwell*, Cambridge University Press, pp. 311–23.

Myers, E. (1987), 'Interpretation of the Energy of the Vacuum as the Sum over Zero-Point Energies', *Phys. Rev. Lett.* 59: 165–8.

Politzer, H. D. (1973), 'Reliable Perturbative Results for Strong Interactions?' *Phys. Rev. Lett.* 30: 1346–9.

Ranjbar-Daemi, S., Salam, A., and Strathdee, J. (1983), 'Spontaneous Compactification in Six-dimensional Einstein–Maxwell Theory', *Nucl. Phys.* B214: 491–512.

Ross, G. G. (1985), *Grand Unified Theories*, Benjamin-Cummings, Menlo Park, California.

Rubin, M. A. and Roth, B. (1983*a*), 'Temperature Effects in Five-dimensional Kaluza–Klein Theory', *Nucl. Phys.* B226: 444–54.

—— and —— (1983*b*), 'Fermions and Stability in Kaluza–Klein Theory', *Phys. Lett.* 127B: 55–60.

Siegel, D. M. (1981), in G. N. Cantor and M. J. S. Hodge (eds.), *Conceptions of Ether*, Cambridge University Press, ch. 8.

Toms, D. J. (1983), 'Induced Einstein–Maxwell Action in Kaluza–Klein Theory', *Phys. Lett.* 129B: 31–5.

Tsokos, K. (1983), 'Stability and Fermions in Kaluza–Klein Theories', *Phys. Lett.* 126B: 451–4.

Unruh, W. G. (1976), 'Notes on Black-Hole Evaporation', *Phys. Rev.* D14: 870–92.

Wilczek, F. and Zee, A. (1984), 'Appearance of Gauge Structure in Simple Dynamical Systems', *Phys. Rev. Lett.* 52: 2111–14.

Witten, E. (1981), 'Search for a Realistic Kaluza–Klein Theory', *Nucl. Phys.* B186: 412–28.

8

Making Everything Out of Nothing

ROBERT WEINGARD

1 Descartes' Theory of Space

I want to begin with Descartes. This may seem strange, since Descartes argues that a vacuum is impossible. But by this he meant simply that two things could not be separated by a distance and yet have, literally, nothing between them. If there were, literally, nothing between two objects, they would be in contact. If there is a distance or extension between two objects, then that distance is a property of something, namely an extended substance.

Indeed, Descartes held that there is only what he calls 'a distinction of reason' between space and matter. Really, the whole universe, with the exception of God and mental substances (minds), is a single substance extended indefinitely in three dimensions. That it is extended in three dimensions is its principle attribute. This means, on the one hand, that all there is to the substance is that it is so extended, the extension constituting or defining the substance; and on the other hand that all the phenomena of the material world are to be explained solely in terms of this extension.

What we ordinarily call 'matter' and 'empty space' are therefore just different regions of the one all-embracing extension. In Descartes' sense of matter, there cannot be space without matter; but in our ordinary usage of 'matter' and 'physical object', there can be. Indeed, in that ordinary sense we can plausibly say that on Descartes' view everything is made of empty space; i.e., everything is reduced to geometrical properties of space. I think we can fairly say, then, that, as we use the word, not only does Descartes think a vacuum can exist, but really, he thinks that all there is is the vacuum.

Descartes wants to reduce physical objects to geometrical properties of space. But his conception of space as an extended substance imposes severe limitations on him. First, since extension in three dimensions is the principle attribute of space, he has to identify quantity of matter with the volume of the region of space that matter occupies. How then can he explain phenomena like condensation or melting where a given quantity of matter appears to change its volume? Second, Descartes, like his contemporaries, takes for granted that space is topologically and metrically Euclidean. How then can different regions of space be geometrically different enough so that they can be identified with different material substances like, say, gold and banana ice cream?

To solve this second problem, Descartes focuses on the property of shape. Although he thinks there is a single continuous extended substance, he conceives of it as being divided up into little regions of different shapes, which behave as particles that can move with respect to each other. In effect, he thinks of the universe as being continuously filled with particles of several distinct shapes. Thus, although he attacks the atomists' distinction between matter and the void, Descartes in practice appropriates much of the atomists' explanations.

For example, the above problems of condensation and melting are given atomistic solutions. Water consists of distinctly shaped regions of space, i.e. particles. In steam they are relatively far apart, but they become more densely packed when the steam is condensed to water. At the same time, air particles, with their own distinctive shape, which were evenly distributed throughout the steam, are displaced from the region where the water condenses and fill up the region where there is no longer steam. The point is, the quantity of matter in the region occupied by the steam has not changed during condensation: its pieces have merely been rearranged.

2 Some Problems for Descartes' Geometrical Theory of Everything

It is well known that such a theory faces severe difficulties. Let me focus on two of these that are relevant to our story. First, consider space at some instant of time. It is divided up into many regions of different shapes. But what distinguishes one region (particle) from another at that instant? Space is homogeneous and isotropic, its only property being that it is extended in three dimensions. I do not see how, then, on geometrical grounds, space at a given instant can be considered to be divided up into particles one way rather than another. Rather, what makes one region a particle, instead of part of two particles, is that it preserves its shape relative to other regions during motion. And that this occurs does not follow from the fact that the region has a certain shape at a given time. Besides shape, then, Descartes has to bring something like causal powers into his explanatory scheme: namely, that some regions (particles) preserve their shape when colliding with other particles, while other regions, say a bunch of air molecules, do not.

Second, Descartes has a problem about motion. Since he thinks of the single extended substance as an incompressible fluid, he knows that one portion of the fluid cannot move unless it moves into a place vacated by the simultaeous motion of an adjacent portion of the fluid. Because of this, he thinks that all motion of the fluid takes place along closed paths. One portion of the fluid moves in step with all the other portions along the closed path so that, as it were, no empty places appear. For Descartes, however, a portion of the fluid is nothing but a portion of extended substance. How then can a portion of extension move, leaving a place (that is extended) for an adjacent portion of extension to occupy?

The point is that, for Descartes, there is no 'real distinction' between the extension of a region of extended substance and the substance itself of which it is the extension. So a portion of the substance cannot move, leaving an extended place for another portion of the substance to occupy. This is not to say, however, that one cannot work out a theory of motion using a relational theory of place. I am only saying that Descartes cannot do this because for him the extension (or distance) between two points is a property of a substance, not a relation between two things.

3 Solving Descartes' Two Problems

Both of the above problems arise because Descartes does not have a rich enough notion of geometry. Because he knows only about Euclidean geometry, he does not have sufficient geometrical resources for his project. So he is forced to use the shape of a region of space as his basic explanatory tool. And this, as we saw, leads him outside the sphere of geometry. However, we know that Euclidean geometry is not the only conceivable geometry, and by allowing space to have a variable curvature and topology, we can solve the above problems.

If we identify each elementary particle with a region of space with a distinctive curvature and/or multi-connected topology, different particles will be distinguished in a purely geometrical way. At any instant, whether a region of space is a particle, a part of two different particles, etc., will depend solely on the geometry of that region at that instant.

That takes care of our first problem. As for the second one, Descartes' problem arises because, to put it crudely, he identifies the principle property of space—extension—with the stuff of space. If these two can somehow be separated, then the properties will be free to move while the stuff—extension—does not. Again, I think we can do this with our space of variable curvature and topology by conceiving of the motion of particles on analogy with the motion of the letters on the illuminated news sign in Times Square. The letters move, not by anything moving, but by the sequential blinking of the lights of which the sign is composed. Similarly, we can conceive of the motion of a particle as the motion of a geometrical property. Regions of space do not move, but the geometry of adjacent regions change in such a way that a pattern of geometrical properties—the pattern we identify with a certain particle—moves. Just as in the Times Square sign, only a pattern moves, not the 'material' instantiating the pattern.

4 Extending Descartes' Programme—Space–Time and the Mind

We have taken care of Descartes' two problems by allowing space to have inhomogeneous time-dependent metrical and topological properties. However, we still have not explained the physical world solely in terms of the geometry of space, for we have also used time as an additional ingredient in our explanations. A complete geometrization of physics must give a geometrical explanation of time as well. And as we all know from relativity, we can do this by making time an additional dimension. We thus take four-dimensional space–time, rather than three-dimensional space, as our fundamental geometrical entity.

But independently of our Cartesian desire to geometrize time, relativity tells us that it is necessary to geometrize time if we are to geometrize the rest of the universe, in particular, if we are to give a geometrical account of gravity. For example, suppose we try to account for gravity in terms of the curvature not of four-dimensional space–time, but of three-dimensional space. Galileo's principle tells us that the trajectory of a particle in a gravitational field depends only on its initial position and velocity, but not on its constitution. In an affinely connected space, one and only one geodesic passes through any point in a given direction. Thus, in three dimensions the initial speed of the particle would be irrelevant, its trajectory being determined by the initial direction of motion only. However, in four-dimensional space–time, the initial three-space velocity does determine a direction in space–time, allowing the geometrizing of gravity.

We have just seen that, although Descartes does not include time as part of the geometry, we can extend geometry to include it. Descartes also thinks that the mind cannot be given a geometrical explanation, and we ought to question that as well. His arguments are really arguments against a physicalist account of mind, and almost no philosopher today accepts them. Either the arguments simply do not work, as a majority believe, or, according to a minority, they work too well and there are no minds; that is, eliminativism is correct.

However, some non-eliminativists have felt that the existence of mind poses problems for the space–time view of the universe. This is because they think we know from our own conscious experience that we persist through time as a subject of change, the same consciousness having different thoughts and sensory experiences at different times. And this is supposed to be incompatible with four-dimensional space–time in which no entity can change its space–time position just because time is part of the geometry.

What we call 'change' in space–time is simply, as Russell argued long ago, differences at different times (different space–time regions). For an entity, whether a banana or a mind, to persist or move through time, and therefore through space–time, would require a time dimension outside of space–time with respect to which this change or movement took place. And therefore mind cannot be explained solely within the framework of space–time.

But the quick reply to this is that what we know from our own conscious experience is that it seems to us that we persist through time, that we are the same consciousness moving from one time to another. However, this fact, it seems clear to me, does not require that anything persists or moves through time (space–time). It simply requires that each instantaneous state of consciousness contain the thoughts, for example that it is perceiving things now which it did not perceive yesterday; that it used to think things it is not thinking now; etc. (Of course, we are to understand a token reflexive account of 'now' here.) And this is compatible with the mind being composed of a series of instantaneous states of consciousness. That is, the 'mind' is a four-dimensional entity (or an aspect of a four-dimensional entity like the history of a person's brain), each three-dimensional slice of which is an instantaneous state of consciousness.

5 Two Questions

My own feeling is that we can conceive, in broad outline anyway, how everything—minds, bananas, and stars—could be composed out of the vacuum in the sense of curved empty space–time. But two questions arise. The first concerns what 'empty' means here; the second whether, and in what sense, contemporary physics fills in the details of the broad outline. Our major concern is with this second question, but let us begin by turning briefly to the first one.

Since we are interested in constructing everything out of curved empty space–time, 'empty' does not mean that there are no physical objects. Intuitively, we mean that such objects are nothing more than aspects of the geometry of space–time. But now, let us carry out our discussion within the framework of general relativity. Then the standard view is that space–time is empty if and only if the energy momentum tensor field **T** is everywhere zero. But it might be objected that, while $\mathbf{T}=0$ means there is no non-gravitational energy, it does not mean there is no energy. For **T** does not include gravitational energy. That is, the objection is claiming that the gravitational field, even though it is an aspect of the geometry, is a physical field. And thus, a space–time containing a gravitational field but nothing else can not be considered empty, i.e. a vacuum.

Assuming that there is a gravitational field in a region S of space–time if and only if the curvature tensor field $\mathbf{R}\neq 0$ throughout S, only a flat space–time will be empty according to the above suggestion. But the curvature tensor is a function of the metric **g**. That is really the physical field of general relativity, and in that sense even flat space–time is not empty. On this view, then, according to general relativity there cannot be a vacuum, empty of all particles and fields, because any space–time will at least contain a metric field.

On the other hand, will anyone who takes this view accept anything as a vacuum? Is there any level of geometric structure which, if it alone existed, would count as being completely empty? My view is that this issue is, at least in part, a matter of definition. Here, I just want to mention that some people might not accept my equating geometry with the vacuum (or as one sense of vacuum).

The second question concerns the extent to which contemporary physics shows us how to carry out the details of the programme of constructing everything out of geometry. As we have seen, in general relativity time and gravity are geometrized by extending the geometry of the universe to the four dimensions of space–time and identifying gravity with the metric curvature of space–time. Let us next consider electromagnetism and ask, Can we give a geometrical account of the electromagnetic field within the framework of general relativity? There are two answers to this which I want to briefly look at: the already unified field theory of Rainich, and the theory of Kaluza and Klein.

6 The Already Unified Field Theory

This approach starts from the combined Einstein–Maxwell equations of standard general relativity,

$$R_{uv}-\tfrac{1}{2}g_{uv}R=\frac{k}{c^2}\,(F_u^wF_{wv}-\tfrac{1}{2}g_{uv}F_{wz}F^{zw}), \tag{1}$$

$$D_{vv}F^{uv}=0, \tag{2}$$

$$F_{uv/w}+F_{wu/v}+F_{vw/u}=0. \tag{3}$$

Here F is treated non-geometrically, as a field that exists in addition to the geometry of space–time. Since **F** carries energy and momentum, it is a source of the gravitational field and space–time is curved in its presence, in accordance with the general relativistic field equation (1), while (2) and (3) are Maxwell's source-free equations.

Now note that (1), with the anti-symmetry of **F**, implies that the curvature-invariant R satisfies

$$R = R^u_u = 0, \tag{4}$$

so that (1) becomes

$$R_{uv} = \frac{k}{c^2}\,(F^w_u F_{wv} - \tfrac{1}{2} g_{uv} F_{wz} F^{zw}). \tag{5}$$

Thinking of **F** as a non-geometrical field, we interpret (5) as **F** imposing a condition on the Ricci tensor R_{uv} and thus on the metric **g**. Given suitable boundary conditions, we solve for the g_{uv} in terms of the given F_{uv}. However, in 1925 Rainich showed that the content of (5) can be expressed in purely geometric terms by (4) along with (6),

$$R^w_u R^v_w = \tfrac{1}{4}\delta^v_u (R_{wz} R^{wz}), \tag{6}$$

in the following sense. The Ricci tensor R_{uv} can be expressed in terms of an anti-symmetric tensor **F** as in (5) if and only if the purely geometric conditions (4) and (6) are satisfied.

This suggests that we regard the use of **F** in (5) as just a convenient device for expressing the purely geometric conditions (4) and (6). On this view, then, a region of space–time containing an electromagnetic field is nothing more than a region of space–time with a distinctive kind of curvature. This curvature is specifed partly by (4) and (6). But that takes into account only the anti-symmetry of **F** and the way it couples to gravity. We also need to translate Maxwell's equations (2) and (3) into purely geometric conditions. In particular, for the identification to work, we need to show that any given geometry corresponds to a unique F that satisfies Maxwell's equations. (Note that we don't have to show that each **F** corresponds to a unique geometry, any more than the mind–body materialist has to show that each mental state or event corresponds to a unique physical state or event.) For non-null regions of space–time ($R_{wx}R^{wx} \neq 0$), and null regions of dimension less than 4, this has been done. But there are null regions of dimension 4 whose geometry does not uniquely determine **F** (Geroch 1966). This is a problem for the already unified theory, to which we will return shortly.

I need to mention one other aspect of this approach. This concerns the fact that it translates the source-free Maxwell's equations into conditions on space–time geometry. How then does the theory account for electromagnetic fields that originate from sources, namely charged particles and their currents? The answer of Misner and Wheeler (1957) is to give a geometrical account of the sources as well as the fields. Along the lines mentioned earlier in connection with Descartes, they suggest (developing an idea of Einstein and Rosen 1935) that a charged particle is a region of space–time with a multi-connected topology.

If we think of the plane of a sheet of paper as a three-dimensional spatial slice, then a multi-connected region can be represented by the intrinsic geometry of a handle as depicted in Fig. 1. If electric field lines thread their way through the handle, then the flux of the electric field **E** over the surface S surrounding one end of the handle will be non-zero and so will look like a charged particle. But really, div**E** = 0 and the field is sourceless. We just don't recognize that the region is multi-connected, so we think S completely encloses a region of space.

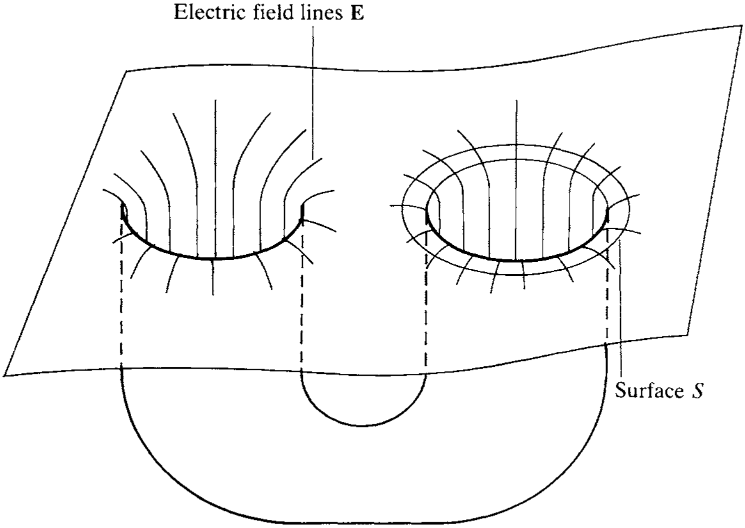


FIG. 1

7 Problems with the Already Unified Theory

I want to consider three problems for the already unified theory. The first, which has already been mentioned, is that there are null gravitational fields, $R_{uv}R^{uv} = 0$, which do not determine a unique **F**. So there are distinct electromagnetic fields that are compatible with the same curved geometry. It has been argued by Geroch (1966) that this is not a problem, since any **F** that we can measure involves test charges and so, because of the presence of these charges, will be non-null. So we will never meaure (detect) an electromagnetic field that is not uniquely determined by the geometry.

But I think that this response to the null field problem does not work. For the theory allows that we could detect electromagnetic fields that would have been null except for the perturbation of the test charge. If we detect such a field, we would accept that as detection of a null field, just as we always correct for the effects of our measuring devices. And that would be evidence that such null fields exist where we have not made any measurements, not that there are only perturbed null fields associated with measurements; that is, the response works only if we accept that, in general, measurements tell us only what exists during measurements. And this is a view not even supported, I would argue, by quantum mechanics.

The second problem is that the Misner–Wheeler strategy for constructing 'charged particles' cannot be carried through within general relativity. There are no time-independent, regular particle-like solutions to the general relativistic field equation $R_{uv}=0$. And the third is that we now know that there is a close connection between the electromagnetic and weak interactions. But that connection is not one that can be understood in terms of the curvature of four-dimensional space–time.

However, it turns out that on our second approach towards a geometrical account of the electromagnetic field we can understand the connection between the electromagnetic and the weak force. So let us now turn to the Kaluza–Klein theory.

8 The Kaluza–Klein Theory

The basic idea of the original Kaluza–Klein theory is to get the extra degrees of freedom for the electromagnetic field by going to five-dimensional relativity. The metric tensor then has 15 independent components, instead of the 10 it has in four dimensions. That leaves us five new components to work with, four of which can be the components of the electromagnetic four-vector potential.

Letting $u=1,\ldots,4$, $i=1,\ldots,5$, the admissible coordinate transformations are assumed to be

$$\begin{aligned}\bar{x}^5&=x^5+f^5(x^u),\\ \bar{x}^u&=f^u(x^u),\end{aligned}\tag{7}$$

and the five-dimensional metric is not a function of x^5. Here we are taking the x^u to be coordinates of four-dimensional space–time, while x^5 is a coordinate for the extra space dimension. Then, from the expression for an arbitrary coordinate transformation,

$$\bar{\gamma}_{5u} = \frac{\partial x^i}{\partial \bar{x}^5} \frac{\partial x^j}{\partial \bar{x}^u} \gamma_{ij}, \tag{8}$$

we see that, under a pure four-dimensional space–time transformation $\bar{x}^5 = x^5$,

$$\bar{\gamma}_{5u} = \frac{\partial x^v}{\partial \bar{x}^u} \gamma_{5v}, \tag{9}$$

while under a pure x^5 transformation, $\bar{x}^u = x^u$ and

$$\bar{\gamma}_{5u} = \frac{\partial f^5}{\partial \bar{x}^u} \gamma_{55} + \gamma_{5u}. \tag{10}$$

Thus, the mixed components γ_{5u} transform as a four-vector under ordinary four-dimensional space–time transformations, but by (10) pick up the gradient of a scalar function under a pure x^5 transformation. Assuming further that the topology of five-dimensional space–time is $M^4 \times S^1$ and that x^5 coordinizes the circle S^1, an x^5 coordinate transformation is a $U(1)$ gauge transformation along the circle. And we can identify the γ_{5u} with the components of the electromagnetic four-vector potential.

However, the components of the metric **g** of our familiar four-dimensional space–time cannot be identified with the 1–4 space–time components γ_{uv} of the five-dimensional metric. Instead, the relation between γ and **g** is given by

$$\gamma = \begin{bmatrix} g_{uv} + A_u A_v \phi^2, & A_u \phi \\ A_u \phi, & \phi \end{bmatrix}. \tag{11}$$

To get the splitting of the five-dimensional metric γ into a four-dimensional space–time metric and four-dimensional vector field, we have to use a frame that gives non-zero cross terms γ_{u5}, $u < 5$. Therefore, there is no single series of four-dimensional slices of the five-dimensional space–time that can be identified with 'observable' four-dimensional space–time. Instead, the observable metric **g** is an abstraction from the different slices.

This is analogous to the rotating disc frame in flat four-dimensional space–time. Although space–time is flat, you can easily see that three-dimensional space relative to the disc is curved. Measuring rods tangential to the disc contract relative to those that lie along a radius. Because the world-lines of the disc are not orthogonal to any single series of spacelike hypersurfaces, space relative to the disc is an abstraction from the actual hypersurfaces of space–time.

Note what this means. Suppose we start with Einstein's empty space–time equations in five dimensions. If we express this in terms of the four-dimensional metric **g** and the potentials A_u, it splits into the coupled Einstein–Maxwell equations in four dimensions. In four dimensions, the space–time metric **g** is coupled to the electromagnetic field **F**, and is therefore curved. From the point of view of four-dimensional relativity, the curvature of space–time is due to the energy of the electromagnetic field. But from the five-dimensional perspective, the electromagnetic energy–momentum tensor is not fundamental. It is, instead, a geometrical term that arises in rewriting the five-dimensional Einstein equations. Instead of a causal account, we thus have a geometrical explanation for why the electromagnetic field is associated with curved four-dimensional space–time.

In the Kaluza–Klein theory, it is assumed that the radius l of the extra compact space dimension is very small, perhaps on the order of the Planck length. That explains why we never notice the fifth dimension as a dimension of space. (We certainly do notice it, because we notice the electromagnetic field.) Within classical Kaluza–Klein theory there is no explanation for the smallness of this radius, but in the quantum version there is. This is interesting to us because here is a place where the quantum mechanical vacuum connects with the classical vacuum, i.e. with geometry.

The aspect of the quantum mechanical vacuum or ground state relevant here is vacuum fluctuations. While the vacuum expectation $\langle 0|\mathbf{E}|0\rangle$ of the electric field **E** is zero, the mean square deviation $\langle 0|E^2|0\rangle \neq 0$. Thus there is a (divergent) vacuum energy density, the so-called zero-point energy. If we consider two parallel conducting plates a distance d apart, in otherwise empty space, the plates impose a boundary condition on the zero-point modes. They must vanish on the plates. Taking the energy density ρ of the vacuum as our reference energy, we find that the difference $\rho-\rho_d$, when the plates are a distance d apart, is finite after introducing a suitable regularization, such as a high-frequency cut-off of the order of the interatomic distance of the atoms of the conductor. This leads to a finite energy per unit area of the plates of the form

$$-\frac{c}{d^3}, \qquad c>0,$$

which gives an attractive force,

$$f=-\frac{3c}{d^4},$$

between the plates. This is the well-known Casimir effect.

In the five-dimensional Kaluza–Klein theory, the finite 'circumference' of the compact fifth dimension imposes a boundary condition on the vacuum fluctuations of the high-dimensional gravitational field. Analogous to the Casimir effect, we get a cut-off-independent zero-point energy per unit volume of the form

$$-\frac{c'}{l^4}, \qquad c'>0.$$

Again, this gives rise to an attractive 'force' which causes the fifth dimension to contract. The force gets stronger as l decreases, which drives l towards the Planck length.

We saw that in the already unified field theory there is the problem of giving a geometrical account of the sources of the electromagnetic field. In four-dimensional relativity there are no time-independent particle-like solutions. But they do exist in five-dimensional relativity. The reason is, roughly, that the Euclidean version of a time-dependent four-dimensional solution can be made into a time-independent five-dimensinal space–time, just by adding $-dt^2$ to the metric (or line element). A simple example is based on the Euclidean Taub–Nut solution. This can be written

$$ds^2=H[dx^5+4m(1-\cos\theta)d\phi]^2+\frac{1}{H}(dr^2+r^2d\theta^2+r^2\sin^2\theta\, d\phi^2) \tag{12}$$

$$H=1+\frac{4m}{r}.$$

Since from (11) we recognize the general Kaluza–Klein metric as

$$ds^2=\Phi(dx^5+A_u dx^u)^2+g_{uv}dx^u dx^v,$$

we see that A_ϕ is proportional to $4m((1-\cos\theta)$, which is a Dirac monopole. In terms of the unit vectors $\hat{\boldsymbol{\phi}}$ and $\hat{\mathbf{r}}$,

$$A_\phi = \frac{4m(1-\cos\theta)}{r\sin\theta}\,\hat{\boldsymbol{\phi}} \qquad \text{and} \qquad \mathbf{B} = \text{curl}\,\mathbf{A} = \frac{4m\hat{\mathbf{r}}}{r^2}.$$

Here we have a particle-like electromagnetic field embedded in a perfectly regular geometry, albeit a five-dimensional one, and assuming that x^5 is suitably periodic. While this is an advance over the already unified theory, the problem of geometrizing sources is not completely solved in Kaluza–Klein because only some sources can be so treated, but not all.

But the Kaluza–Klein solutions have an independent philosophical interest that is relevant to our discussion. Because of analogies with two-dimensional creatures trying to understand three dimensions, as in flatland and sphereland, and the fact that the fourth spatial dimension is of the same type as the other ones, most philosophers[1] would argue that the interpretation of the extra dimension in Kaluza–Klein as a real, physical space dimension is perfectly intelligible. What some of them do question, after granting its intelligibility, is whether we must interpret the extra space dimension this way, or can regard it as just a mathematical device for finding four-dimensional theories.

There are, I think at least two things we can say here. First, this is a problem that affects all physical theories that postulate unobservable entities or properties. One can ask the same question about atomic theory. Does it really describe very small entities, or can we regard it as just a useful formal device for describing and manipulating macroscopic phenomena? And since it is the same problem, the same moves in support of one side or the other can be given.

But second, there are Kaluza–Klein space–times which are not plausibly interpreted in terms of four-dimensional space–time fields (Gross and Perry 1983). Our monopole is one of them. There are also magnetic dipole solutions, dipoles that do not come from moving currents and have zero angular momentum. (This comes from the Euclidean Kerr metric.) These are structures that really require the geometry of the four space dimensions to be understood (Gross and Perry 1983).

In the case of the monopole, for the Dirac string to be a gauge artefact (i.e. to be unobservable), the extra space dimension must be periodic. Since the radius of the fifth dimension is related to the

[1] But not all: see e.g. Van Cleve (1987).

electric charge in Kaluza–Klein theory, this is just the geometrical version of Dirac's charge quantization condition. Similarly, it is the non-trivial topology of the four-dimensional spatial slices that makes the magnetic dipole stable, that is prevents the annihilation of the monopole–antimonopole pair that make up the dipole.

Just to forestall misunderstanding, compare this with the embedding of three-dimensional spherical space, S^3, in four-dimensional Euclidean space. If physical space has the structure of S^3, it might be argued that we can explain the properties of space by the hypothesis that it is a three-dimensional hypersphere in a four-dimensional Euclidean space. Maybe this is right. But if our evidence is just that the geometry of space is S^3, we do not have to make this additional hypothesis. The hypothesis that space is S^3 stands on its own, and a higher dimension embedding space is not required for its intelligibility. In the above Kaluza–Klein examples, however, the extra space dimension does play a crucial role in the understanding of the three-dimensional monopoles and dipoles.

9 Extensions of Kaluza–Klein

One of the problems with the already unified theory is that, by explaining electromagnetism in terms of the curvature of four-dimensional space–time, it could not shed any light on the connection between electromagnetism and the weak interaction. However, we can extend the Kaluza–Klein theory so that we can include not only the weak interaction, but the strong interaction as well. We do this by adding more than one compact space dimension. Namely, we add enough compact space dimensions so that the $SU(3) \times SU(2) \times U(1)$ gauge symmetries are realized as isometries in the extra dimensions. And then the gauge fields emerge from the metric in a way analogous to the five-dimensional theory.

Unfortunately, this beautiful theory faces a very serious problem. It does not admit of chiral fermions. We know that right- and left-handed fermions belong to different representations of $SU(2) \times U(1)$. Since these fermions get mass (those that do) by spontaneous symmetry-breaking, they will be massless modes of a $(4+N)$-dimensional spinor field.

Now there are two possibilities. The first is that chirality is a low-energy effect. The higher-dimensional fermions are vector-like, but for some reason we have not seen the mirror partners of the observed fermions. The second is that the fermions are chiral from the start in the higher dimensions. The problems with the first are obvious. There are no observed mirror fermions; the family structure repeats itself, and as Witten remarks (1981, 1985), the anomalies cancel already, without the mirror fermions. So the second approach is the favoured possibility.

But there are very strong arguments against the possibility of chiral fermions in the higher-dimensional Kaluza–Klein theories. As an example, consider the massless Dirac equation in $4+N$ dimensions:

$$\gamma^u D_u \psi = \gamma^u D_u^4 \psi + \gamma^u D_u^n \psi = 0,$$

so $\gamma^u D_u^n$ acts as a mass operator and we are interested in its zero eigenvectors. In order to have chiral fermions, that is, left- and right-handed fermions transforming according to different representations of the gauge group G, G must admit of complex representations. But when the gauge fields come from the metric, the zero modes of $\gamma^u D_u^n$ transform according to a real representation of G.

We can get an idea of how the gauge fields coming from the geometry can make a difference by noting that

$$(i\gamma^u D_u^n)^2 = -\Sigma D_u D^u + \tfrac{1}{4}R,$$

where R is the curvature scalar of the N-dimensional compact space, and $D_u = \partial_u + gT^a A_u^a$. Because the A_u^a are components of the metric, we get the connection with the geometrical quantity R. It turns out that $-\Sigma D_u D^u$ is a non-negative operator, so in any compact space with $R > 0$ everywhere, the Dirac operator has no zero modes (Witten 1981, 1985).

If we admit elementary gauge fields into our theory, then we can have chiral fermions in higher dimensions. But then, from the point of view of Kaluza–Klein theory, we have lost our motivation for introducing higher dimensions, which was to give a geometrical explanation of the gauge fields.

We can contrast this with string theory. In string theory we also have extra space dimensions, but they are not postulated to explain the gauge fields. Rather, consistent quantization of the string requires extra dimensions. The extra dimensions come about naturally, but the gauge fields do not get a geometrical explanation (at least, not in terms of the geometry of space–time).

10 The Gauge Field as Geometry: Another Point of View

The Kaluza–Klein idea is one attempt to give a geometrical interpretation to the gauge fields. But if we reject Kaluza–Klein, there is still a geometrical option opeւ to us. We can note that the gauge fields form a connection, $T^a A^a_u$, on the fibre bundle whose base space is four-dimensional space–time and whose typical fibre is the internal space of the quantum fields. This can also be regarded as a geometrical interpretation of the gauge fields. In going from Newtonian mechanics to general relativity, we go from a three-dimensional spatial geometry to a four-dimensional space–time geometry in which gravity is the metric curvature of that geometry. Similarly, the unified gauge interaction field strength is the curvature of the connection of an expanded geometry, of which four-dimensional space–time is just a component.

If you like, you can regard calling this geometry an extension of the notion of 'geometry'. But it is a natural extension because of the structural, or formal, similarity to what we already call geometry. However, I want to make two remarks about this extension. The first brings us back to the subject of monopoles; the second is that there is more to the view that the gauge field is geometry than that it can be given a certain mathematical formulation.

First, concerning the idea of the gauge field as part of the geometry, some have thought that there is much more to it than I have indicated above. That is, they have thought that, ontologically, the internal space directions are somehow on a par with the four dimensions of spacetime. As a case in point, the t' Hooft Polyakov monopole has been offered as an example of how this is to be understood.

This monopole is a composite field configuration built out of a SU(2) gauge field and a triplet of scalar fields. Of relevance to our present discussion, the asymptotic structure of the scalar fields is of the form

$$\phi^a = \frac{bx^a}{r}, \qquad r = [(x^1)^2 + (x^2)^2 + (x^3)^2]^{1/2}. \tag{13}$$

The index a refers to the SO(3) internal space. But if we regard the ϕ^a as the components of a spatial vector, then $\boldsymbol{\phi} = b\hat{\mathbf{r}}$ is a radial vector field in three-dimensional physical space. As the spatial vector field $\boldsymbol{\phi}$ changes direction in physical space as we move around the monopole, the internal space vector with components ϕ^a changes direction in the internal space. Thus, it seems that a given direction in physical space is also a direction in the internal space.

But this apparent mixing of internal and physical spaces is an illusion.[2] For take any three scalar fields ϕ^a that form an SO(3) triplet of the internal space. Then we can define the vector field $\phi^a\hat{\mathbf{x}}^a$, and as the vector field changes direction in physical space, it also changes direction in the internal space. But this is artificial. Rather than taking $\phi^1\hat{\mathbf{x}}^1+\phi^2\hat{\mathbf{x}}^2+\phi^3\hat{\mathbf{x}}^3$ as my physical space vector, I could have chosen some other combination, such as $\phi^3\hat{\mathbf{x}}^1+\phi^1\hat{\mathbf{x}}^2+\phi^2\hat{\mathbf{x}}^3$. The point is, the SO(3) triplet ϕ^a, by itself, does not pick out a unique direction in physical space, and so does not establish a correspondence between the internal and physical spaces. It is just a convention as to what direction in physical space I associate with ϕ^a.

In the case of the monopole, it looks as if a unique physical space direction is picked out because of the form of equation (13). It is natural to take $b(x^a/r)$ as the component in the direction $\hat{\mathbf{x}}^a$; but, again, it is just a convention that we do so.

On the other hand, and this is my second point, I do think there is more to the view that the gauge field is geometry than that it can be given a certain mathematical formulation. In classical mechanics, for a system of s particles with kinetic energy T, we can write

$$2Tdt^2=\sum_{i=1}^{s} m_i(dx_i^2+dy_i^2+dz_i^2).$$

If there are m kinematical constraints, we can express the $3s$ rectangular coordinates x_i, y_i, z_i, in terms of $n=3s-m$ generalized coordinates q_j, and write

$$ds^2=2Tdt^2=\sum_{i,j=1}^{n} g_{ij}dq_idq_j.$$

The evolution of the system will then be a geodesic in the n-dimensional space with metric g_{ij}.

We have given a geometrical formulation of the classical mechanics of particles. But I do not think that we would regard the fact that

[2] I thank Tim Maudlin of the Rutgers University Department of Philosophy for this observation.

mechanics can be formulated in this way as telling us about the geometry of physical space (assuming the classical mechanics were correct). This is because this is a reformulation of the mechanics of s particles in three-dimensional Euclidean space, in terms of one 'particle' in $3s-m$ dimensions with a curved metric. Since each formulation is a complete theory of mechanics, the g_{ij} is not an extension of the geometry of Euclidean three space. Rather, we believe that ontologically, the three-dimensional Euclidean space formulation is fundamental. This is because, for example, we believe that there could have been a different number of particles, the constraints could have been different, etc., without affecting three-dimensional space. In the case of the gauge field, however, we are not giving a reformulation of the geometry of space–time. Instead, we are making a genuine addition to it.

Note that it is conceivable that the dimension of physical space depend on the number of particles in it. Perhaps, when there are n particles, space has $3+n$, or $3+3n$, or some other number of dimensions. But then, that would be not a reformulation of mechanics, but a new theory of physical space altogether.

11 A Final Observation About Fermions

Even if we accept the view that the gauge connection is part of the geometry, we have not given a geometrical account of everything. In particular, we have not given a geometrical account of Fermi fields. Let me end with the following observation.

The geometry of space–time is a classical, macroscopic structure. It can (presumably) be given a quantum mechanical account because Bose fields have coherent states, whose expectation values are the classical quantities. We can try to extend space–time geometry to include the electromagnetic field, and the other gauge fields, because they are also Bose fields.

Now it is also true that Fermi fields can exhibit classical, macroscopic behaviour. The He^3 superfluid and superconductivity are two examples. But the basic particles of these phenomena are composite bosons. In elementary particle physics (and chemistry), however, what we see are not composite bosons, but fermions engaging in the basic interactions. These are the quanta of fermion fields which do not admit of 'macroscopic' coherent states.

References

Einstein, A. and Rosen, N. (1935), 'The Particle Problem in the General Theory of Relativity', *Phys. Rev.* 48: 73–7.

Geroch, R. P. (1966), 'Electromagnetism as an Aspect of Geometry? Already Unified Field Theory—The Null Field Case', *Ann. Phys.* 36: 147–87.

Gross, O. J. and Perry, M. J. (1983), 'Magnetic Monopoles in the Kaluza–Klein Theory', *Nucl. Phys.* B226: 29–48.

Misner, C. W. and Wheeler, J. A. (1957), 'Classical Physics as Geometry: Gravitation, Electromagnetism, Unquantized Charge and Mass as Properties of Curved Empty Space', *Ann. Phys.* 2: 525–603.

Rainich, G. Y. (1925), 'Electrodynamics in the General Relativity Theory', *Trans. Am. Math. Soc.* 27: 106–30.

Van Cleve, J. (1987), 'Right, Left, and the Fourth Dimension', *Phil. Rev.* 96: 33–68.

Witten, E. (1981), 'Search for a Realistic Kaluza–Klein Theory', *Nucl. Phys.* B186: 412–28.

—— (1987), 'Fermion Quantum Numbers in Kaluza–Klein Theory', in T. Appelquist, A. Chodos, and P. G. O. Freund (eds.), *Modern Kaluza–Klein Theories*, Addison-Wesley, Reading, Mass., pp. 438–511.

9

Vacuum or Holomovement

B. J. HILEY

1 Introduction

Quantizing the gravitational field presents some formidable problems. The deep link between gravity and space–time implies that, if quantization is to be successfully carried out, then radical changes in our understanding of space–time will be needed. In this paper I will explore some new ideas about space–time that can be motivated through a purely algebraic approach to quantum mechanics and which call into question the notion of absolute locality. In order to bring out these ideas, a brief review of the classical and quantum notions of the vacuum, particularly in their relation to our understanding of space–time structure, is first presented. This sets the context in which the new approach can be introduced. It is centred around the novel notion of the *holomovement*, a notion that provides the ground form from which both space–time and matter itself can be abstracted. Thus, in this approach it is the holomovement that provides the source of all being.

2 The Classical Vacuum

The history of the vacuum has had, in my view, a rather curious evolution. Throughout history, opinions as to its nature seem to swing from one extreme to the other. It has been considered at times to be 'full' or substantive, while at other times it has been treated as 'empty' or void. Even from the earliest days, there has been little agreement. For instance, Parmenides argued that 'emptiness is nothingness, and that which is nothing cannot be'. To him the

I would like to thank David Bohm and Philip Davies for their many helpful discussions on the questions raised in this paper.

vacuum had to be a compact plenum which he regarded as being constituted as one continuous unchanging whole. And it is logic alone that forced him to the conclusion that movement is mere illusion. But surely, movement is more than just illusion. Material substance is perceived as constantly changing, some changes being rapid, others being extremely slow. Is this not more naturally explained by the Democritian atoms moving from one region of space to another not already occupied by other atoms, i.e. into empty regions? Therefore, to conceive of movement of substantive entities, must we not surely have an empty vacuum?

This notion of a void, of the empty vacuum, provided the backcloth for the development of Newtonian physics. Stengthened by the differential calculus, particle mechanics grew from the primitive concepts supplied by Democritus and extended qualitatively by Lucretius. Particles-in-motion would provide a mechanical explanation of all physical processes. Even light, in Newton's view, was particulate in nature, with 'corpuscles' moving through the vacuum, sometimes being reflected and sometimes being transmitted when they reached a transparent boundary. However, their predicted behaviour once they entered the medium was not supported by experiment, and it was wave theory that ultimately triumphed with its simpler explanations of interference and diffraction. The corpuscular theory was abandoned until the event of the quantum theory.

But how could a wave be sustained in 'empty space'? Surely, all our experience of wave phenomena was mechanical in origin and required a medium in which the vibrations could be sustained. The mechanical ethos had become so deeply ingrained that an explanation of electromagnetic fields in terms of vibrations in some kind of substantive ether was strongly advocated. Thus a plenum-like vacuum was reintroduced, and by the end of the nineteenth century it became fashionable to take this ether very seriously, and to seek an explanation of the ultimate source of all physical phenomena, including the atoms themselves, in terms of structures or invariant features of the plenum itself (such as vortices).

It was the failure to detect any movement of the earth relative to this Maxwellian plenum that began to raise serious doubts about the existence of such a 'substance'. Not only had experiments like those of Michelson and Morley failed to detect any such movement through the ether, but the very structure of the special theory of relativity made it appear that any attempt to look for such an ether was doomed to failure. The seemingly inevitable conclusion appeared to be that the vacuum is 'really empty', a notion that has dominated the more recent developments in physics. The reaction against the reintroduction of such an ether or plenum has been so strong that any theory that dared to call on such a notion was for a time deemed to be unacceptable and even preposterous. In the 1960s and 1970s I often came across such a reaction when I tried to discuss de Broglie's use of a 'sub-quantum medium' as a means of providing a possible explanation of the quantum formalism. The objection was not so much against the attempt to find a more physically intuitive explanation of quantum phenomenon, but rather against the introduction of the 'sub-quantum medium'. The retort, 'Surely Einstein has shown us that the vacuum is "empty" and the reintroduction of such an outmoded way of thought will not provide a satisfactory understanding of phenomena', was not uncommon. Yet in relativistic quantum field theory the notion of 'vacuum polarization' had already emerged and was being used quite freely, albeit in a very formal way.

But it is not necessary to evoke quantum field theory. Einstein (1924) himself did not react so strongly against the notion of an ether. What he questioned was the need to interpret Maxwell's equations *mechanically*, i.e. in terms of vibrations of a plenum-like substance. Was it necessary to regard the notion of field as an attribute of substance, or could it be regarded as something in its own right? Einstein (1969) argued that Maxwell's equations successfully accounted for a large number of phenomena and that was not necessary to interpret the field quantities in terms of any deeper structure. Nothing was to be gained by trying to interpret these fields in terms of an underlying substance, as demanded by the mechanical approach. Newtonian mechanics was clearly limited, so why continue to use an outmoded conceptual form? Let the continuous field be regarded as an entity in its own right.

Of course, it would now be possible to take the electromagnetic field itself as an ether and attempt to account for all the properties of matter in terms of this field alone. Indeed, many efforts were made in this direction. However, there were processes that were clearly not electromagnetic in origin. Gravity was one of the more obvious examples, and it soon became evident that something more was needed.

Einstein was the first to realize that the electromagnetic field contained a further limitation, namely, that Maxwell's equations are invariant only in special frames of reference, the inertial frames. But why single out this particular class of frames as privileged? Should we not extend the principle of relativity to include all frames of reference, both inertial and non-inertial? A serious problem in generalizing the ideas to all frames seemed to be that the accelerating frames still contained a vestige of the Newtonian absolute, namely, the space–time continuum itself. Einstein did not see this as a serious problem, because he had noticed that the inertial frames themselves also treated the space–time continuum as an absolute. First of all, it should be noted that the inertial frame provide a manifold of events that are coordinated into a space–time continuum, and since the relations between these events are independent of the actual inertial frame, we can regard this continuum as 'physically real'; Einstein (1924) actually went as far as to call it an 'ether'. But it is not an 'ether' to be explained in mechanical terms; rather, the aim would be to account for this ether in terms of a deeper, new, fundamental field which is to be considered as an entity in its own right.

This Minkowski space–time continuum is also absolute in another sense, namely that, while the continuum has a physical effect in that material processes can be used to measure space and time intervals, the manifold itself is not influenced by physical conditions (Einstein 1960). Thus, not only do we have the limited nature of the electromagnetic field together with its dependence on a set of favoured frames, but we also have a further limitation in that the theory actually relies on an independent and absolute space–time continuum.

The above remarks clearly indicate the limitations of the special theory, of Maxwell's equations and the Lorentz group, and it was therefore necessary to consider some form of generalization. The first key assumption in this generalization was to retain the notion of a field as primary. If we want to change the absolute nature of space–time, material processes must influence the space–time continuum itself so that the fundamental field used to describe the structure of space–time should be shaped by the presence of matter. To achieve this, we would try to express the structure of space–time through some set of differential equations that involves matter itself. But now, not only would these equations be invariant to the Lorentz transformations, but they would also be invariant to a much larger group of transformations in order to remove the special role played by the set of inertial frames. If we chose the larger set of transformations to be the group of all coordinate transformations, then no frame would be privileged. All frames of reference would be equally valid, and we could extend the principle of relativity by requiring the laws of physics to be invariant in *all* frames of reference. In this way one arrives at the principle of general relativity. Following on from this, Einstein was able to include gravity via the space–time metric, and all that remained was to find how space–time is affected by the matter that it contains.

It is not necessary to recall here the details of Einstein's arguments that finally led him to his field equations,

$$R_{\mu\nu}-\tfrac{1}{2}g_{\mu\nu}R=T_{\mu\nu}. \tag{1}$$

All that is necessary is to emphasize that it is the field that is taken as basic, and that an explanation of this field is not to be looked for in terms of any underlying material medium. The field *is* the explanation, and it is this field that characterizes the structure of space–time. To Einstein (1924), terms like 'the gravitational field', 'the structure of space–time', and 'the ether' were all synonymous.

In this context the question of whether the ether, or the vacuum, was 'full' or 'empty' becomes somewhat ambiguous. If we did not regard the gravitational field as substantive, we would conclude that the ether is 'empty' in the sense of containing no matter. But it could also be regarded as 'full' in the sense that the metric field is always present, although in the absence of matter it is only potentially active. I use the word 'potential' because we would observe its presence only through the behaviour of particle probes. But this ambiguity of the notion of 'emptiness' has been commented upon a number of occasions before. Descartes, for example, reminded us that in common speech 'the term empty usually means, not a place where there is no object at all but simply a place where there is no object such as we think there ought to be'. He gives the example of a water jug which is called empty if it contains no *water*; but, of course, it does contain *air*. He goes on to remind us 'that a space containing nothing sensible is "empty" even if it is full of created and self-subsistent matter' (see Smart 1964). Maxwell (1954) sums up this position rather aptly: 'the vacuum is that which is left in a vessel after we have removed everything which we can remove from it.'

Now returning to general relativity, it should be remembered that Einstein (1969) himself did not regard the theory as expressed through the field equations (1) to be complete. To quote from his autobiography, 'The right hand side [of equation (1)] is a formal condensation of all things whose comprehension in the sense of a field theory is still problematical' (Einstein 1969: 75). It is as if this term were an interim '*asylum ignorantiae*' (see Hyland 1979). Indeed, it was not clear how a particle as such could be described in this theory. To complete the theory, Einstein wanted to regard the particle itself as a concentration of field energy or, perhaps, even a singularity in the gravitational field so that gravity, particles, and, of course, electromagnetism could be described by one single field, the 'unified field'. Einstein was unable to fulfil his hopes, and physics took another direction which was forced on it by quantum mechanics.

3 The Quantum Vacuum

In the conventional approach to non-relativistic quantum theory, the idea of a position in space is formally replaced by an operator. Bohr argued that this was necessary because at the quantum level the position of an object becomes somewhat ambiguous, this ambiguity being understood in terms of the uncertainty principle $\Delta x \Delta p \approx h$. Indeed, this result can be shown to follow directly from the properties of the operators. On the other hand, time is not replaced by an operator and continues to be treated as a parameter. Nevertheless, an energy–time uncertainty principle is introduced using an entirely different set of arguments. There is no operator generalization of time in going to the relativistic quantum theory. Instead of attempting to find an operator for time to put it on the same footing as position, progress was actually made by making position a parameter again, while making the field quantities operators. Thus there was a return to a classical Minkowski space–time where both position and time are treated as parameters. In other words, space–time is again assumed to be absolute in the Einstein sense discussed above.

Although it is the field that is again taken as basic, there is no single unified quantum field. Rather, there are a series of individual but interacting fields, each being classified by their spin and rest energy. Furthermore, the particles are not singularities of the field, but are represented by the normal modes of the field with quantized energies (the quanta). The excitation and de-excitation of these modes are interpreted as the 'creation' and 'annihilation' of 'particles'. In this context, a new notion of a 'vacuum' is introduced which is not directly related to the space–time structure. Formally, one introduces a 'vacuum state' $|0\rangle$ through the equation

$$a_k|0\rangle = 0, \qquad (2)$$

where a_k is an annihilation operator of the field.

There are several features of these vacuum states that make it hard to conceive of them as 'empty'. A quantum field always has a residual zero-point energy which appears as a consequence of the non-vanishing of the mean square average of the field in its ground state. Thus, although there are no field excitations present in the form of quanta, there is nevertheless a residue activity which actually has experimental consequences. If we consider the electromagnetic field, for example, then fluctuations in this activity can be interpreted as spontaneous creation and annihilation of virtual photons, or virtual particle–antiparticle pairs (i.e. vacuum polarization). When the electromagnetic field is in interaction with, say, the electron (whether treated as a particle or a field), this vacuum polarization can produce observable changes, as is seen in the hyperfine structure of hydrogen (e.g. the Lamb–Retherford shift).

In particle physics, the notion of the vacuum state has increasingly played an important role. There are many different vacuum states, with notions such as 'false vacuums', tunnelling from one vacuum state to another (Coleman 1977), 'unitary inequivalent vacuum states' that arise in interacting quantum field theories (Emch 1972), and so on. In present theories, these vacuum states are treated as entities in their own right, but no attempt is made to account for their variety in terms of any deeper structure. However, it seems clear that the vacuum states are being treated as if they are 'ground states' of some deeper physical process. Indeed, I would like to suggst that these 'ground states' are actually quantum states of some underlying structure which, as has been indicated, is limited by the special role of the inertial frame and the absolute nature of the space–time continuum, as discussed above.

Some attempts have been made to address the difficulty of this absolute continuum by extending quantum field theory to curved space–times. Perhaps the most illuminating general discussion of this problem from our point of view has been made by Davies (1984). What he made clear was that, if a field theory is constructed in an inertial frame, then the same field analysed from the point of view of a non-inertial frame required a different vacuum state from the one used in the inertial frame, so that each non-inertial frame would require its own vacuum state.

To illustrate what is involved, suppose we consider the plane wave decomposition of a scalar field in an inertial frame. Then we can write

$$\Phi = \sum_k (a_k\phi_k + a_k^\dagger \phi_k^\dagger). \tag{3}$$

We will require a vacuum state $|0\rangle$ defined by $a_k|0\rangle = 0$.

Let us now turn to consider what this field looks like in an accelerating frame. We can again decompose the field

$$\Phi' = \sum_k (\tilde{a}_k\psi_k + \tilde{a}_k^\dagger \psi_k^\dagger) \tag{4}$$

where ψ_k will be the appropriate set of (calculable) modes appropriate to the accelerating frame. (See Birrell and Davies 1982 for details.) The relation between the set of annihilation and creation operators $(a_k, a_k^\dagger)$ and $(\tilde{a}_k, \tilde{a}_k^\dagger)$ can be shown to be the Bogolubov transformation,

$$\begin{aligned}\tilde{a}_k &= \alpha_{kj}^* a_j - \beta_{kj}^* a_j^\dagger \\ \tilde{a}_k^\dagger &= \alpha_{kj} a_j^\dagger - \beta_{kj} a_j\end{aligned} \tag{5}$$

where α and β are determined by the nature of the accelerating frame. In this decomposition there exists a vacuum state $|\tilde{0}\rangle$ defined by

$$\tilde{a}_k|\tilde{0}\rangle = 0,$$

so that if

$$\langle 0|a_j^\dagger a_j|0\rangle = 0, \tag{6}$$

then

$$\langle 0|\tilde{a}_j^\dagger \tilde{a}_j|0\rangle = \sum_i |\beta_{ji}|^2. \tag{7}$$

The meaning of this result is as follows. If we begin initially with two inertial frames and assume the field ϕ is in the ground state, then neither observer will see any ϕ-quanta present. If one of the frames is made to accelerate, then the field relative to this frame must be analysed as in equation (4), while the vacuum state remains $|0\rangle$ (free-field Heisenberg picture). If $\beta_{ij} \neq 0$, the accelerating frame will then find quanta present in the field. Thus, an accelerating field will have a different vacuum state, and the inertial observer's vacuum state will not be empty relative to the accelerating observer's frame.

Although the above argument was developed specifically in terms of an accelerating observer, it can be generalized, in principle, to any non-inertial frame. Thus we see a new kind of ambiguity arising in the notion of the vacuum. What is empty to some observers will 'contain particles' for others, so that the vacuum itself has become a relative concept and we must conclude that there is no unique vacuum.

Davies (1984) has presented a particularly clear discussion of these ideas and has argued strongly in favour of using the 'Copenhagen spirit' of quantum mechanics. Here Bohr's key idea (1961) was that one cannot make a sharp separation between the form of the experimental conditions and the content (meaning) of the results. This implies that there was no unambiguous way of attributing properties to particles except in the context of a particular given experimental situation. In the extension of these ideas to relativistic field theories, one argues that there is no unambiguous way of attributing properties to fields except in the context of a given experimental situation. Since the 'particle' is an excited state of the field, it follows that the notion of 'particle' itself cannot be given an unambiguous meaning except in the context of some experimental situation. Thus in Bohr's view we should not talk about 'particles' existing in their own right. Rather, we must always talk about particles and vacuums in relation to specific arrangements of detectors.

This will, of course, provide a completely consistent interpretation of the formalism in its relation to specified experiments. However, it should be noticed that this whole approach assumes that nothing can be said about quantum processes other than in the context of measurement. But one of the key motivations in extending quantum theory to curved space–times and to gravity is, ultimately, to discuss cosmology in a quantum context. If we are to raise questions about the evolution of the universe, then it seems very difficult to see how we can obtain meaningful answers if we are able to discuss only its evolution in terms of specific experimental arrangements. What we require is a discussion of quantum processes in themselves, so that we need not place undue emphasis on measurement processes.

The Copenhagen interpretation to which Davies appeals actually denies that it is possible to obtain an unambiguous description of an individual process at the quantum level, *even in principle.* Bohr's analysis that led to such a conclusion is subtle and somewhat involved, depending to a large extent on what physicists today would regard as philosophical arguments. It is not possible in this short paper to appraise Bohr's analysis, but, briefly, what emerges from his approach is that quantum processes are so ambiguous that it is not possible to have an ontological description of the individual process. Thus, in some sense quantum physics is reduced to epistemology. In spite of this particular kind of ambiguity, the statistical results of given experimental arrangements are quite unambiguous, and it is these results that should be the concern of physics.

It has generally been assumed that Bohr had reached the correct conclusion and that the only way forward is to extend the quantum algorithm to all areas of quantum physics. To make this extension in a coherent manner, it has generally been assumed that the wave function should be taken as giving the most complete description of the state of a quantum system while the time evolution of this function provides the most complete account of the evolution of the system, in spite of the difficulties to which such a statement leads.

4 Towards a Quantum Ontology

For a number of years, David Bohm and I have been exploring an alternative approach to the quantum theory that does allow a description of the evolution of an individual quantum process in a way that reproduces all the known results of quantum mechanics and in consequence does not put primary emphasis on measurement. The main aim of our work has been to develop and clarify the conceptual structure that is necessary to provide a consistent account of the evolution of these individual processes. The hope is that, by having a clear physical understanding of these processes, it will be possible to open up new approaches that may be relevant to future physics.

In our various publications we have shown that it is possible to provide an internally consistent ontological description of individual quantum processes both for particles and for fields. For example, in the non-relativistic theory, our approach gives the particle ontological status in the sense that every particle has simultaneously a definite, but unknown, position *and* momentum. In consequence, the particles follow well defined trajectories which can actually be calculated for many situations from the theory. The trajectories are unlike classical trajectories in the fact that they are determined not only by a classical potential, but also by a qualitatively new potential which we call the quantum potential. This means, for example, that even in 'empty' space, in which there is no classical potential at all, the particle need not move in a straight line because it can be acted on by a quantum potential. Indeed, it is the appearance of this new potential that produces results that are consistent with quantum phenomena.

The quantum potential itself is determined from the wave function which is assumed to satisfy the Schrödinger equation. In using the wave function in this manner, we do *not* regard it as specifying the most complete description of the state of the system. Rather, we regard it as describing a pair of real coupled fields which together shape the evolution of the individual quantum system. These fields have new properties which are totally different from those of classical fields. We have suggested that new concepts, such as 'active information', are needed in order to comprehend the action of these fields and provide a consistent ontology.

An extensive investigation of this approach has been carried out for various situations such as stationary states, barrier penetration, quantum transitions, etc. The main purpose in studying these examples was to see how the quantum potential determines the evolution of an individual quantum process. Indeed, it became clear that one can provide a clear insight into how individual effects come about, and can even talk about a specific time at which a transition actually takes place, without the need to introduce a measuring instrument. In this way, measurement is reduced to a special subclass of transitions (see Bohm and Hiley 1987).

Recently these ideas have been extended to relativistic quantum boson fields (see Bohm *et al.* 1987). Here the individual field can be given a well-defined meaning, and the quantum aspects of the field stem from the presence of an additional 'super-quantum potential', so that once again the deviations from the classical behaviour arise as a result of the action of this super-quantum potential. In this view, the energy of the excited states of the field are continuously distributed in the field as a whole, while the notion of quanta arise as a result of the nature of the super-quantum potential. For example, in the case of absorption of a quantum of energy from the field, it is the nonlinearity and non-locality of the super-quantum potential that causes the energy to be 'swept in' from the entire field so that it concentrates a single quantum of energy at the absorber. In this way we see that, although the field is continuously distributed, it manifests itself in a discrete, particle-like way.

One of the most important lessons that we learn from this whole approach is that it is indeed possible to produce a consistent ontological account of an individual quantum process. Thus, we have provided a counter-example to the apparently compelling position that Bohr and those who followed him have proposed. What has become clear is that their position is not a necessary consequence of physical phenomena themselves, but arises from adopting a particular philosophical position. In a forthcoming book Bohm and I have analysed their position in detail and have shown that this approach depends upon certain key assumptions that are by no means essential (Bohm and Hiley 1991). If these are dropped, then it is possible to explore new approaches to quantum phenomena which do not make the assumption that the individual process is so ambiguous that it is only through measurement that one can achieve unambiguous results. I believe that this now opens up the possibility of developing new ontological descriptions that will be of particular significance for quantum gravity and quantum cosmology. In this way we will be able to study the evolution of the cosmos without needing to regard measurement as an essential basic element of the theory. But all of this will follow only if we give up the notion that the wave function is a function of state and the thinking that goes with it.

5 Locality, Non-locality, and Pre-space

What I would like to do in the remainder of this paper is to give a qualitative outline of a particular form such a new theory might take, concentrating on those aspects of the proposals that are pertinent to the questions discussed in this paper. In particular, I want to consider some of the deeper ideas that are implicit in our study of the quantum potential approach which I think will survive, although much of the original form of its structure will not.

A superficial analysis of our approach to non-relativistic quantum mechanics might lead to the impression that somehow a new potential is being 'plastered' onto the classical ontology. Such a conclusion would be misleading, because the appearance of this new potential radically alters the nature of the ontology. The quantum potential is not like a classical potential. In classical physics fields act mechanically by transferring energy and momentum from the field to the system. In contrast, the quantum potential depends not on the strength or intensity of a field, but on the *form* of the field. This can be seen most easily by looking at the mathematical expression for the quantum potential, which is $Q = -(\hbar^2/2m)\nabla^2 R/R$ where R is the amplitude of the quantum field. Thus we see that, if the field intensity is multiplied by an arbitrary constant, the potential does not change. This implies that very weak fields can produce large effects. These effects do not come from the transfer of energy and momentum from the quantum field to the particle; rather, there is a redirection of the energy within the particle itself. A direct consequence of this is that systems separated by large distances can be strongly interacting.

In the non-relativistic theory it is straightforward to show that this force is, in general, non-local. This non-locality arises in many-body systems which are characterized by a non-product wave function. In these systems non-locality means that particles separated by very large distances can be 'locked together' so that any change in the behaviour of one particle instantaneously produces corresponding changes in the others. Non-locality does not arise in systems that are described by a simple product of single-particle wave functions. Since these wave functions are a special case of a non-product wave function, we must conclude that, in contrast to classical physics, quantum physics takes non-locality to be basic while locality arises in special cases.

At first sight this non-locality seems to contradict the theory of relativity, which requires that no signal be transmitted faster than the speed of light. But it should be noted that this unexpected and unpleasant feature of non-locality arises from a non-relativistic theory in which there is, of course, no upper limit to the speed of transmission. However, what is even more surprising is that this non-locality persists even in the *relativistic* field theory that we have examined. Here one finds that the field at different points can be coupled non-locally through the quantum potential, and, indeed, it is this coupling that offers an explanation of the photon cascade experiments that were actually used in the Einstein–Podolsky–Rosen (EPR) experiment. In spite of this non-locality, the ensemble average over many field configurations produces results that are identical with those calculated from standard quantum mechanics. Thus, both locality and Lorentz invariance emerge at a statistical level even though they are not present at the level of the individual field.

Furthermore, at the classical level, where the quantum potential is negligible, only locality remains, and in the relativistic domain all individual processes are Lorentz-invariant.

Bohr was implicitly aware of these 'non-local' features of quantum mechanics, but because his approach was based on the assumption that reality was inherently ambiguous, we cannot use such a notion at the level of individual processes, and therefore the question of non-locality, and indeed of locality itself, was not relevant in the quantum mechanics. Locality is a classical concept, and it is only at this level that it has significance. It is true that the quantum field is local, but this field is only part of an algorithm and, contrary to present convention, it was not taken to be a function of state of the individual. In fact, quantum mechanics requires a different way of thinking, and to emphasize this Bohr suggested that the word 'phenomenon' should be used in a new way that was different from that commonly used in physics. He writes: 'I advocate the application of the word *phenomenon* exclusively to the observations obtained under specified circumstances, including an account of the whole experimental arrangement' (Bohr 1961: 64). If we add to this the fact that 'we are not dealing with an arbitrary renunciation of a more detailed analysis of atomic phenomena, but with a recognition that such an analysis is *in principle* excluded', we are naturally led to the notion of the 'wholeness of the quantum phenomenon', which cannot be further subdivided even in principle (Bohr 1961: 62). Thus it was not a question of trying to find a mechanism at the individual level in order to account for non-locality, but it was that the phenomenon itself could be given a *statistical* meaning only within the arena of classical space–time. In so far as Davies is applying the spirit of Bohr's approach to quantum field theory in curved space–time, the space–time continuum in this formalism is thus classical in essence and the vacuum states are merely part of the algorithmic description through which the phenomenon is revealed. The quanta, therefore, have meaning relevant only to the measuring instruments, as Davies correctly remarks.

In contrast to Bohr's position, the interpretation through the quantum potential gives primary relevance to the space–time manifold in which an actual process is assumed to evolve and each individual process can be described unambiguously. Within this context, it is assumed that space and time are to be treated in exactly the same way as in non-quantum physics, both in the non-relativistic and the relativistic theories. That is, space–time has built into it an *a priori* notion of locality, together with a Galilean or Lorentz symmetry. In neither theory is there any need to replace position and time by operators, and therefore one avoids the inconsistency that appears when the standard non-relativistic theory is extended to include relativity in the usual approach, a difficulty that I referred to earlier. But the appearance of the key new notions of non-locality and, perhaps more importantly, wholeness raises the question of whether space–time can ultimately remain as a classical absolute object.

This is not the only reason for questioning the role of space–time. Indeed, there is an informed view that if gravity is to be successfully quantized then the continuous space–time manifold will, in all probability, have to be abandoned as an *a priori* given starting-point. As Wheeler (1967) has so eloquently pointed out, the fluctuations of the metric can become so violent that the very notion of a well-defined and permanent neighbourhood is called into question. It is as if the continuum develops fluctuating topological features of which the 'wormhole' is, perhaps, the best known example. Thus, the appearance of these new fluctuating topologies suggests that the notion of an absolute locality is beginning to break down. However, it seems that this notion has not been directly questioned because the basic differential manifold structure is still retained. But if non-locality is a key factor at the quantum level, then this will not fit naturally into the space–time manifold because the manifold arose specifically as an expression for an absolute notion of locality. Indeed, if the notion of a field is added to the manifold, it provides the ideal description for 'action-through-contact'. Thus, the appearance of any form of 'fluctuating metric' or 'action-at-a-distance' will not fit naturally into the concept of a manifold. But our analysis of quantum theory through the quantum potential suggests that both locality and Lorentz invariance can arise statistically, and therefore a natural suggestion to make is that the space–time manifold should itself be thought of as arising statistically from some deeper, as yet unspecified, structure. The quantum vacuum state provides a hint of this structure, but lacks the necessary relation with space–time. In order to bring out this new possibility, we have called this deeper structure, 'pre-space'.

Before proceeding to examine the possibilities for pre-space, we should pause to consider our attitudes towards the notion of locality in more detail. It is a deeply held belief that locality is essential for the description of physical processes (see Davies 1989), and, indeed, some would even go as far as to say that science itself is not possible without a notion of absolute locality. I want seriously to question this assumption. In doing so I am not going to deny that it has been successful, particularly at the classical level; nor do I pretend that it will be easy to give up such a notion. Indeed, I am not even sure it will be possible ultimately to proceed without such a notion, but I feel strongly that the issue should be debated.

In order to motivate my own approach to locality, let me recall one of the main principles used in the development of both special and general relativity, namely the removal of certain absolutes. Newtonian mechanics was built on the concepts of absolute time and absolute simultaneity. It was through the emergence of the Lorentz group that both of these became relative concepts. The further step to general relativity rested on the fact that this group left space–time itself as an absolute with the inertial frames playing a special role. As was explained earlier, it was the removal of this special role that led to the space–time manifold being defined, not as an *a priori* given absolute, but as a structure that depended on the distribution of matter in it. However, what still remained was the notion of absolute locality.

To Einstein (1971), this notion of absolute locality seemed essential for physics. The appearance of non-locality in quantum mechanics led him to suggest that the theory itself was in some sense incomplete and therefore only provisional. It was necessary to find a deeper theory which would account for quantum phenomena by restoring locality. Einstein himself never succeeded in this aim, and now the Bell inequalities (Bell 1987) have clearly shown that this will not be possible. I feel that non-locality is an essential feature of quantum processes and that we must accommodate it by calling into question the notion of a differential manifold.

The suggestion I would like to make is that we replace the notion of absolute locality by a relational notion of locality. The possibility that locality could be relational becomes apparent only when the object–image relationship in the hologram is considered. The essential point here is that it needs only a small portion of the hologram itself to reconstruct an image of the original object with all its neighbourhood relations intact (although, of course, the image loses some of its original sharpness). The important principle emerging from this consideration is that in the hologram the local features of the original object are carrried non-locally. Thus what is local in one representation (the original object) becomes non-local in another (the hologram), and vice versa. But the relation between these two 'frames of reference' is such that the structural content is the same in both frames. What makes the object frame more fundamental to us is that the object has meaning for us in the classical world which we take to be a preferred frame. Normally this is taken to be a Newtonian frame. Clearly, this is not an absolute frame. It is preferred simply because of its utility and not because it has any absolute significance. Relativity has explained all that.

Relational properties actually play an important role in quantum mechanics. One can see this most clearly in the causal interpretation (see Bohm and Hiley 1987). To explain this idea, let us first recall that measurement in quantum theory is, in general, active and not passive. This point can be brought out most clearly again through the causal interpretation, but it is already implicit in the usual approach. It arises because of the nature of the quantum potential, which is such as to cause the system and the measuring apparatus temporarily to 'fuse' to form an inseparable whole before the apparatus finally evolves into one of its distinct final states. Each of these final states is correlated to a final state of the system which is, in general, different from its initial state. It is as if the measured properties of the particle are, in general, 'created' by the measurement itself. The properties not being measured also undergo uncontrollable and unpredictable changes, and it is this that is ultimately responsible for the uncertainty principle. In other words, in quantum mechanics measurement is a transformation in which both system and apparatus actually participate. Thus, although the particles have intrinsic properties, the properties that are measured have meaning only in the context of the particular apparatus involved. Making different measurements again transforms the system and reveals new properties. It is in this sense that measurement plays an active role and the properties are thus relational.

There is one notable exception to all of this in the causal interpretation, and that involves position itself. It can be shown that position is the only non-relational property in the causal interpretation. This means that the position of a particle can be measured without changing it, and in that sense position can be regarded as an intrinsic property. It is this property that gives space–time its privileged and absolute role in the causal interpretation.

Indeed, one of Heisenberg's objections to the causal interpretation was the special role given to space–time. He pointed out that the non-relativistic transformation theory suggested that it was the operators that carried the essential quantum structure and that the position representation was no more fundamental than, say, the momentum representation. I believe Heisenberg was right to emphasize the importance of the algebraic structure of the operators, and I also believe that his objection to the causal interpretation was right, but for different reasons. For me, it is the need to give up an *a priori* given space–time and to introduce a relational notion of locality that limits the validity of the causal interpretation. This, of course, means that the causal interpretation cannot carry us any further.

Giving up space–time as a basic entity implies also that neither the particle nor the field can be taken as basic any longer. As explained above, this is essentially because the space–time manifold evolved from first taking particles-in-local-interaction and then fields-in-local-interaction as the basic explanatory form for physics. Thus it is clear that, if we are to develop a decriptive form that uses not a notion of absolute locality, but rather a relational notion of locality, some very radical new ideas will be needed.

6 The Holomovement and the Abstraction of Space–Time

I believe that the clue as to which direction to proceed in is already implicit in the algebraic structure of the operator formalism, but in order to develop these ideas it is necessary to introduce a different way of looking at physical processes. First we must break with the traditional view that operators are to be thought of as 'observables' which simply give the results of some sort of measurement. Rather, recall the argument that, in quantum mechanics, measurement should be regarded as a particular case of a more general process of transformation in which both 'system' and 'observing apparatus' participate as a whole. In Bohr's language, it is not possible to make a sharp separation between the two. Thus, every transformation irreducibly connects at least two distinguishable aspects of a total process.

In a reductionist framework one would tend to regard the end-points of the connection as basic and the connection itself as secondary. But here it is the connection that is taken to be basic, while the end-points are abstracted as distinguishable features of the connection. Furthermore, it is important to note that this connection is not passive but active. Thus activity or becoming, or more simply movement, is now taken as basic. Within this basic movement there will be quasi-stable, semi-autonomous features, some of which will have particle-like properties. Thus, particles themselves are merely 'ripples' on the 'sea' of underlying activity. In this outlook there are no ultimate particles out of which all other particles are formed. Rather, we will have quasi-invariant forms which can transform into each other and can be self-organized into hierarchies of larger quasi-stable isolatable systems. The totality from which these features will emerge is called the *holomovement*, a term that was first introduced by Bohm (1980) and later given a preliminary mathematical form by Bohm *et al.* (1970). The prefix 'holo' is chosen to emphasize the indivisibility that is so important in quantum processes.

The term 'holomovement', then, describes the totality of processes containing not only the quasi-stable features which we sense either directly or indirectly with the aid of our detection instruments, but also in the transient, very fast processes to which our present instruments are insensitive. It is the latter that correspond to what quantum mechanics calls the vacuum state and which de Broglie inappropriately called the 'sub-quantum medium'. In our view, therefore, the vacuum itself has a very rich structure of activity to which our instruments cannot respond directly (i.e., it is empty only in the sense of Descartes). It is this activity that is responsible for the zero-point energy. It can also be regarded as the source of the quantum potential, but before we can establish a firm connection between the two, it is first necessary to obtain a better understanding of its deeper structure and to explore how its properties may be revealed indirectly in the behaviour of the quasi-invariant features that are accessible to our instruments.

In order to carry out this task we must find an appropriate mathematical form with which to describe this deeper structure. Our first thought in this direction was to regard each connection as a one-dimensional directed simplex, so that the totality of these one-dimensional connections would form a simplicial complex or multiplex. This approach enabled us to exploit the isomorphism between an abstract cohomology theory and the de Rham cohomology that appears significant in the old quantum theory. What emerged from this work was that the main equations of quantum mechanics could be reformulated in a way that was independent of an assumed continuous space–time manifold. This was actually illustrated in some detail for Maxwell's equations, but one could extend these ideas to include the Schrödinger and Dirac equations. Indeed, the Dirac operator can simply be written as a sum of the boundary and coboundary operator, a fact that was later exploited in lattice quantum chromodynamics (see Becher and Joos 1982).

Although the results were of some interest, the simplicial complex does not capture the essence of activity, nor was it particular to quantum mechanics, as the application to Maxwell's equations illustrates. Quantum mechanics with its product and linear superposition of operators suggests that it is the *algebraic* structure that is essential. In my first attempts to relate this algebraic structure to the basic notion of activity, my attention was drawn to some of the early work of Grassmann (1894), Hamilton (1967), and Clifford (1882). Although these names are very familiar because of their very significant contributions to algebraic geometry, it is not generally realized that what motivated them was not simply a desire to understand static geometry; all of them took activity or becoming as the basic form from which all emerges, and it was through an algebra that they found a natural way to describe activity.

Both Grassmann and Hamilton were uncompromising: mathematics, they claimed, was about thought; it was not about a material reality. It was a way of studying relationships in thought: not a relationship of content, but a relationship of form in which a content could be carried. To quote Hamilton (1831), 'Relations between successive thoughts thus viewed as successive states of one more general and changing thought, are the primary relations of algebra.' Thus, mathematics is to do with ordering forms that are created through thought and hence of thought. The fact that these forms could be 'useful' for helping us order physical phenomena was, to them, almost incidental.

The central question that Grassmann addressed was how one thought became another. Can we ever say that we have an entirely new thought which is completely independent of the old thought? Or is it that they are distinct but related? I like to regard them as opposite poles of an essential relationship that exists between the poles themselves. They are then inseparable in the sense that the old thought contains implications of the new thought and the new thought contains a trace of the old. As Bergson (1922) puts it, 'evolution implies a real persistence of the past in the present'. It is the relationship that is essential. But one can go further. The relationship itself is also of thought, and this, in turn, must be one pole of another essential relationship, so that thought itself can be understood as a structure of relationships ordered in some form of hierarchy. By considering this structuring in thought, Grassmann was able to break out of the straitjacket of our three-dimensional world, and for the first time it became possible to describe mathematically spaces of dimensions higher than three.

Grassmann insisted that his theory of space was but a particular realization of the more general notion that he was working towards. Nevertheless, he saw each point of space as a distinctive form in a continuous process of becoming, and motion was conceived of as a continuous generation of one distinctive form followed by another. To Grassmann the succession of these distinctive forms was not to be regarded as the generation of a series of independent points; rather, each successive point was considered the opposite pole of its immediate predecessor so that pairs of points became essentially related, and it is this relationship that Grassmann called his 'field of extensives'. Indeed, to strengthen the essential link between pairs of points, he wrote the points enclosed in square brackets, $[p_ip_j]$, to emphasize the unity of the poles. Higher-order connections of distinctive forms could then be written as $[p_ip_jp_k]$ and so on.

These extensives are not yet the vectors, bivectors, trivectors, etc., that present mathematics uses in what is called a Grassmann algebra. The original basic forms that motivated Grassmann have long been forgotten or ignored, and only those features that are more appropriate for static visualization have been retained. Thus, the very notion of becoming that Grassmann felt to be so important in his approach to algebras has been lost. In fact, this emphasis on the static forms occurred very early in the development of algebraic geometry, and by the 1880s the notion of activity had already been replaced by the static forms of a vector, bivector, etc., together with the exterior product that we now associate with Grassmann algebras today. In fact, Clifford (1882), in contrasting this static view that had emerged from Grassmann's work with Hamilton's own approach, found it necessary to restate the fact that Hamilton had put the emphasis on activity, and to point out that the quaternion product, for example, was actually based on the notion of transformation or activity. By extending these ideas, Clifford was able to generalize the quaternion algebra to a hierarchy of algebras that bore his name. Could it be more than good fortune that relativistic quantum mechanics uses the Clifford algebra in such a fundamental way?

In order to develop the connection between activity and algebras in the quantum context, it is necessary to be aware of the fact that quantum mechanics can also be formulated in a purely algebraic approach which does not refer to operators on a Hilbert space. This algebraic approach had its origins in Heisenberg's work and was investigated originally in some detail by Segal (1946). (An excellent summary of the algebraic approach to quantum field theories can be found in Haag and Kastler 1964.) In general these approaches are very formalistic, and not immediately transparent; consequently it is not at all obvious how this approach could be related to the intuitive ideas outlined in the previous paragraphs.

In order to bring out this connection, Fabio Frescura and I (1980*a*, 1980*b*, 1984) attempted to show in a less rigorous, but hopefully more intuitive, way how it was possible to present a purely algebraic theory of non-relativistic quantum mechanics. Two algebras emerge in this work. First there is the symplectic (Heisenberg) quasi-algebra defined by the generating set $\{1, D_i, Q_i, E_i\}$ over the complex field through the relations

$$[D_i, Q_j] = \delta_{ij}; \quad [D_i, D_j] = [Q_i, Q_j] = 0; \; E_i^2 = E_i; \\ D_iE_i = 0; \quad E_iQ_i = 0, \tag{8}$$

where D is an algebraic displacement operator (cf. $-i\partial/\partial x$ in the Schrödinger representation), Q the position operator (cf. x), and $E = \Pi_i E_i$ is the primitive idempotent which is intimately connected with the vacuum state.

The second algebra to arise is the Clifford algebra, defined by the set of generating elements $\{1, \Delta_i, \theta_i\}$ over the complex field through the relations

$$\{\Delta_i, \theta_j\} = \delta_{ij}; \quad \{\Delta_i, \Delta_j\} = \{\theta_i, \theta_j\} = 0; \quad P_i^2 = P_i; \\ \Delta_i P_i = 0; \quad P_i\theta_i = 0, \tag{9}$$

where θ is a Grassmann number, Δ its conjugate, and $P = \Pi P_i$ is the primitive idempotent. In this case $P_i = \Delta_i\theta_i$. The algebra of Dirac matrices γ_μ are a sub-algebra in the algebra defined in terms of

$\theta_\mu \pm \Delta_\mu$. The elements of the algebra that correspond to the vectors in Hilbert space are just the minimal left ideals.

Nowhere in this work does there appear the space–time continuum; nevertheless, the algebras themselves carry the main space–time symmetries. We could regard this structure as pre-space. Thus, for example, the quaternion Clifford algebra carries the spatial rotations, while the higher-dimensional Dirac–Clifford algebra carries the Lorentz boosts. The symplectic algebra carries the space displacements and the time displacement. Hence it appears that the space–time continuum is not a prerequisite for these symmetries. If this is, in fact, a correct conclusion, then the next essential question that we must raise is whether the presence of these symmetries will enable us to abstract out a space–time manifold from the algebra itself. This is a difficult task which has not yet been completely solved, but it is possible to illustrate how this may be achieved, in principle, if one starts with a very simple algebra. Of course, this algebra is far too simple to capture all the features that will be necessary in a complete theory, but nevertheless it does give an insight into how this might work.

The algebra that we shall use is a finite polynomial Weyl algebra, C_n^2, generated over the complex field by a set of generating elements $\{1, e_0^1, e_1^0\}$, subject to the relations

$$e_0^1 e_1^0 = \omega e_1^0 e_0^1; \qquad (e_0^1)^n = 1; \qquad (e_1^0)^n = 1, \tag{10}$$

where $\omega = \exp(2\pi i/n)$ (see Davies 1981).

The physical significance of this particular algebra can be seen by letting $n \to \infty$. In this limit the algebra becomes essentially the symplectic algebra with e_0^1 being identified with D and e_1^0 being identified with Q. Thus, when n is finite we can regard this algebra as carrying the structure of a two-dimensional phase space with a finite number of points rather than a two-dimensional continuum.

To abstract this phase space from the algebra, we must first indicate how one abstracts ‘generalized points’ from the algebra. An important clue in how to proceed in this task has been provided by Eddington (1958), who argued that, in a purely algebraic approach to physical phenomena, there are elements of existence defined not in terms of some hazy metaphysical concept of existence, but in the sense that existence is represented by an element in the algebra that contains only two possibilities: existence or non-existence. He assumed that existence is represented by an idempotent in the appropriate algebra. In our view, in which we take activity as basic, it is not so much that points exist but that they *persist*. They persist not in the sense that they are static, but rather that they continually transform into themselves. Clearly, the element that transforms into itself is the idempotent because $P \cdot P = P$. (The product represents succession.) Thus, in the algebra, generalized points are going to be represented by elements of activity that transform into themselves so that the generalized points will correpond to a set of idempotents in the algebra.

Let us now illustrate this procedure in the case of the Weyl algebra C_n^2. Because the algebra is finite, it is possible to find a complete set of pairwise orthogonal primitive idempotents ϵ_i. One such set will be

$$\epsilon_i = \frac{1}{n} \sum_k \omega^{-ik} e_k^0. \tag{11}$$

The ϵ_i will satisfy the relation $\Sigma_i \epsilon_i = 1$ and $\epsilon_i^2 = \epsilon_i$. It is our contention that the set $\{\epsilon_i\}$ constitutes a set of generalized points in our phase space. To show that this set can be used to represent a finite set of points in 'space', let us introduce a 'position' operator X defined by

$$X = \frac{1}{n} \sum_{jk} j\omega^{-jk} e_k^0 = \sum_j j\epsilon_j, \tag{12}$$

so that

$$X\epsilon_j = j\epsilon_j. \tag{13}$$

Thus, each primitive idempotent is labelled by the eigenvalue of the 'position' operator X. If we choose ϵ_0 as the point at the origin, its neighbouring point is ϵ_1 since there exists a unit displacement operator, T, such that

$$\epsilon_1 = T\epsilon_0 T^{-1};$$

so that in general

$$\epsilon_j = T^j \epsilon_0 T^{-j}. \tag{14}$$

It is then not difficult to show that the element T^{-1} is just e_0^1 for this particular set of idempotents. Thus, the set of primitive idempotents forms a linear ordered array that can be regarded as the position space. Furthermore, each point is connected to its neighbour by the element e_0^1.

Of course, this set of idempotents is not unique, and there exist many equivalent sets of idempotents each with its own translation or neighbour relation. Since the algebra is rather trivial, all these representations are 'physically' the same, and there is no way to distinguish them. If we regard the algebra C_n^2 as describing the holomovement, then within this holomovement there exist many possible realizations of this 'space'. In the language of David Bohm (1980), the holomovement is an implicate order within which there are many realizations of the explicate order we call space. With a richer algebraic structure there exists the possibility of inequivalent 'spaces', and clearly some physical criterion will be needed to distinguish between these spaces. Therefore we have the possibility that different observers will be able to abstract different spaces and the spaces will have a different content.

Returning to the Weyl algebra C_n^2, we notice that there is a symmetry between e_0^1 and e_1^0. Clearly, it should be possible to use e_1^0 as a translation operator for a different set of primitive idempotents, $\epsilon^{`}_j$, such that $\epsilon^{`}_j = e_j^0 \epsilon^{`}_0 e_{-j}^0$. This set of points would now represent a space that is, in some sense, dual to the set of points represented by the set ϵ_j. We would expect this set of generalized points to have a particular significance even within this simple algebra. To bring out this relation, we first note that we can introduce an operator $X^{`}$ using the equation

$$X^{`} \epsilon^{`}_j = j \epsilon^{`}_{\mathrm{j}} \tag{15}$$

where

$$X^{`} = \frac{1}{n} \sum_{jk} j \omega^{-jk} e_0^1 . \tag{16}$$

The meaning of these particular generalized points becomes more transparent if we write the X-displacement in the form

$$T_x(a) = \exp(-iaP),$$

with $a = 2\pi/n$ and P the momentum operator. But we can also write a p-displacement as

$$T_p(a) = \exp(iax).$$

Then we can identify $e_0^1 = \exp(2\pi iP/n)$ and $e_1^0 = \exp(-2\pi iX/n)$.

This suggests that $X^{`}$ should, in fact, be identified with P, i.e.

the discrete momentum, so that the dual space is nothing but the momentum space. Thus, this algebra contains two sets of distinguished elements that can represent either the points of a discrete position space or a discrete momentum space. These generalized points are not independent but are related by a similarity transformation,

$$\epsilon'_j = S\epsilon_j S^{-1}, \tag{17}$$

where

$$S = \frac{1}{n}\sum_{ijk} \omega^{j(i-k)} e_k^{j-i}. \tag{18}$$

S is, in fact, a discrete Fourier transformation (see Davis 1981), as one would expect.

The important feature from our point of view is that both sets of points cannot be displayed simultaneously. The reason for this is that the ϵ_j are the eigenfunctions of *X*, while ϵ'_j are the eigenfunctions of *P*. But *X* and *P* do not commute. In fact,

$$[X, P] = \frac{1}{n}\sum_{jkrs} (s-j) r\omega^{r(s-j)}\omega^{-jk} e_k^{j-s}. \tag{19}$$

This means that ϵ_j and ϵ'_j cannot be realized simultaneously. In more general terms, we say that it is not possible to explicate both the momentum space and the position space simultaneously. In other words, the discrete phase space is carried *implicitly* by the algebra. It is in this sense that the algebra is given a primary role, and its linear subspaces then correspond to the position and the momentum spaces which can not be realized together. In this way the position and momentum spaces are not *a priori* given, but are distinctive features carried by the algebra.

In quantum mechanics the points of space are labelled by means of the kets, i.e., $|x_j\rangle$. The relation between this labelling and our algebraic structure emerges through the minimal left ideals. To show how this comes about, let us generate a minimal left ideal, $\mathscr{L}^0$, from ϵ_0. Recall that

$$\mathscr{L}^0 = \mathscr{A}\varepsilon_0,$$

where $\mathscr{A}$ spans all the elements of the algebra C_n^2. The basis of this ideal is

$$\mathscr{L}^0(j)=\frac{1}{n}\sum_k e_k^{-j}. \tag{20}$$

Furthermore,

$$X\mathscr{L}_0(j)=j\mathscr{L}^0(\mathrm{j}), \tag{21}$$

$$\mathscr{L}^0(j+a)=e_0^{-a}\mathscr{L}^0(j), \tag{22}$$

so that if we identify $\mathscr{L}^0(j)$ with $|x_j\rangle$ we have

$$X|x_j\rangle=j|x_j\rangle \tag{23}$$

and

$$|x_{j+a}\rangle=T_x(a)|x_j\rangle. \tag{24}$$

These results are exactly what one would expect from the usual approach to quantum mechanics.

We are now in a position to discuss the relational notion of locality within the algebraic structure. The holomovement is a structure of activity with no *a priori* notion of absolute locality within it. In fact, it should be regarded as a-local. In the example of the Weyl algebra, we chose a particular 'local' order by giving special significance to the element e_0^1. Thus we have essentially imposed a local order on the space. This order is arbitrary since it is possible to make an inner automorphism and produce a new set of 'generalized points' with a new neighbourhood operator. Thus, for example, we can define a new set of generalized points ϵ_i'',

$$\epsilon_i''=Z\epsilon_iZ^{-1}, \tag{25}$$

where Z is some element of C_n^2. These 'space' points will also have a corresponding dual 'momentum' space $\epsilon_i^{``}$. In fact, there are many such inner automorphisims in the algebra. Each will have its own unique order and therefore its own neighbourhood relation. If we use a neighbourhood element to define locality, then there exist many different locality relations, each related to another by an inner automorphism.

We can also bring these ideas out within the present formalism of quantum mechanics by examining the Schrödinger representation in the same spirit. The principle of locality is built into this representation through the fact that the dynamical operators are functions of x and $\partial/\partial x$. Furthermore, it is the requirements of continuity and single-valuedness of *all* physically significant operators that give rise to the correct energy levels and transition probabilities. Thus the Schrödinger representation has built into it a notion of locality, whereas the Heisenberg representation, and hence the algebraic approach, does not. In order to make the latter completely equivalent to the Schrödinger representation, one needs to introduce a neighbourhood operator. This can be done as follows. First, let us consider the Schrödinger representation in the discrete case. Here $\psi(x)$ will correspond to a weighting function on the discrete points, x_j, through the relation $\psi_j = \langle x_j | \psi \rangle$. We then replace $\partial\psi/\partial x$ by $\psi_{j+1} - \psi_j$ and $\partial^2\psi/\partial x^2$ by $\psi_{j+1} - 2\psi_j + \psi_{j-1}$.

Under an inner automorphism, the point labelled by $|x_j\rangle$ will go over into a linear combination,

$$|x'_j\rangle = \sum_k C^*_{kj} |x_k\rangle, \tag{26}$$

so that

$$\psi'_j = \sum_k C^*_{kj} \psi_k. \tag{27}$$

The combination $\psi_{j+1} - \psi_{j-1}$ will then go into

$$\sum_k [C^*_{k,j+1} - C^*_{k,j-1}]\psi_k.$$

It is clear that the difference operator has been replaced by another one that is 'non-local', i.e. one that connects points from all over the original space (as if each point had been 'exploded' into the whole space).

To tie in more closely to the algebraic notation used above, we can define two neighbourhood relations, N^+ and N^-, through the equations

$$\begin{aligned} N^+ |x_j\rangle &= |x_{j+1}\rangle \\ N^- |x_j\rangle &= |x_{j-1}\rangle \end{aligned} \tag{28}$$

Then under the inner automorphism, $|x_j\rangle = \Sigma_k C_{jk} |x'_k\rangle$, these operators become

$$\begin{aligned} N^+ |x'_k\rangle &= \sum_m \sum_j C^*_{jk} C_{j+m,m} |x'_m\rangle \\ N^- |x'_k\rangle &= \sum_m \sum_j C^*_{jk} C_{j-m,m} |x'_m\rangle. \end{aligned} \tag{29}$$

In this case all the points in the space are neighbours of each other, and so the neighbourhood operators no longer have their original simple meaning that they had in equation (28). Thus, in the limit of an infinitely dense array, continuity has lost its simple meaning. But even though the 'space' has lost its 'local' meaning under the transform, the physical content is of the theory as reflected in the measurable quantities is unchanged. Indeed, we have something analogous to a hologram in which the locality relations are carried non-locally throughout the hologram. Non-locality is thus implicitly incorporated into quantum mechanics and is a direct result of associating a basis that is not invariant to an inner automorphism.

It is clear from the above considerations of both the Weyl algebra and the Schrödinger representation that the locality relation is imposed from outside the structure. The Weyl algebra brings this out most clearly, whereas in the case of the Schrödinger representation the choice does not appear arbitrary because it uses the classical space–time continuum. In some sense this arbitrary imposition does not introduce serious limitations because within both structures all representations are equivalent. On the other hand, this arbitrariness implies that the physical relationships that are independent of this change of basis are no longer tied to the position space in which classical events are acted out. It is this feature that has led us to ask whether, in fact, there is a need for a basic *a priori* given space–time manifold. The suggestion then is that perhaps the space–time manifold is an essentially classical construct that has ultimately to be abstracted from the appropriate algebra that describes the holomovement.

The algebra that we have explored above is very simple and contains only equivalent representations. Indeed, as soon as one goes on to consider systems with an infinite number of degrees of freedom, inequivalent representations are the rule rather than the exception. Thus, the general algebra describing the holomovement would contain many 'space–time' representations, each with its own structure of generalized points and, in the continuum limit, its own manifold structure. The set of all these representations would then provide the basis for a statistical space–time to emerge.

The above discourse has been intended to bring out the new possibilities that the algebraic approach offers. In all that has been said, I have not yet specifically introduced any new content. To show that such a possibility exists, consider first the analogy of the metric in general relativity. The canonical diagonal metric $g_{\mu\mu}$ can be transformed into a non-diagonal form,

$$\sum_{\mu} \frac{\partial\phi^{\mu}}{\partial\chi^{\rho}} \frac{\partial\phi^{\mu}}{\partial\chi^{\sigma}} g_{\mu\mu} = g'_{\rho\sigma}. \tag{30}$$

Now as long as a global transformation exists, there is no new content in the theory. However, when the curvature is non-zero, only local transformations can be found, which reduce the metric to diagonal form, and we have the possibility of new content, namely gravitation. A similar argument can be applied to the neighbourhood matrix. In the present theory it is assumed that the matrix can be reduced everywhere by a unitary transformation. However, it may not be possible to make all neighbourhood matrices diagonal together. Indeed, the appearance of non-locality in quantum theory may arise from just such a difficulty. Thus we open up the possibility of new content and therefore new physics.

7 Conclusion

I have tried to argue in this paper that a purely algebraic approach to quantum phenomena opens up the possibility of obtaining a generalized ontology of individual quantum processes which does not require an *a priori* given space–time manifold to describe their evolution. Rather than use the notion of fields-in-interaction on the space–time manifold, we should start with the notion of the holomovement which assumes that the basic form is activity or transformation, and that material processes can be abstracted from this activity in terms of quasi-local, semi-autonomous features of this total activity. It has also been demonstrated that space–time itself is to emerge from this structure. Indeed, it is essential to regard the material process and space–time to be abstracted together, so that the particle content depends on the abstracted space–time in an essential way and vice versa.

The appearance of inequivalent representations in a general algebra suggests that there may be many different 'space–times' with different 'particle' contents and different 'locality' relations. All of these will be present in the algebra together. In order to provide a framework in which to comprehend the multiplicity of 'space–times', it is essential to use Bohm's notion of the implicate order. In these terms, the holomovement is an implicate order, and the abstraction of a particular space–time from within it corresponds to the emergence of a particular explicate order. In this view, different explicate orders can have different 'space–times' with a different 'particle' content. Thus, the appearance particles that resulted from the Bogolubov transformation discussed in Section 2 does not arise merely from the measurement process; rather, the accelerating frame of reference demands a different explicate order, i.e. the abstraction of a different space–time from the algebra of the holomovement. This implies that physical processes themselves determine which particular explicate order is relevant in a given situation, and one of the important questions is to determine what factors generate a particular explicate order.

One of the very novel features in our suggestions is that all the various inequivalent representations are present together in the algebra, so that all the different 'space–times' will be implicit in the holomovement. But only one 'space–time' can be explicated in any one situation; the rest remain implicit. Consequently we no longer have the Descartesian ideal that all of nature can, as it were, be laid out before us for our intellectual perusal. Rather, only a certain limited aspect can, in any one situation, be presented at the expense of the rest. This is, I believe, the deeper truth lying behind Bohr's principle of complementarity.

It must be emphasized that the main purpose of this paper has been to draw attention to the new possibilities latent in the algebraic structure which I believe provides a new ontological approach to quantum processes. It is hoped that this approach, which is at present under further development, will offer a new way to address the problems that arise in the quantization of the gravitational field.

References

Becher, P. and Joos, H. (1982), 'The Dirac–Kähler Equation and Fermions on a Lattice', *Z. f. Phys.* C 15: 343–65.

Bell, J. S. (1987), 'On the Einstein–Podolsky–Rosen Paradox', ch. 2 in *Speakable and Unspeakable in Quantum Mechanics*, pp. 14–21, Cambridge University Press.

Bergson, H. (1922), *Creative Evolution*, Macmillan, London.

Birrell, N. D. and Davies, P. C. W. (1982), *Quantum Fields in Curved Space*, Cambridge University Press.

Bohm, D. (1980), *Wholeness and the Implicate Order*, Routledge & Kegan Paul, London.

—— and Hiley, B. J. (1984), 'Generalisation of the Twistor to the Clifford Algebras as a Basis for Geometry', *Rev. Brasileira de Fisca*, special issue '70 anos de Mario Schönberg': 1–26.

—— and —— (1987), 'An Ontological Basis for the Quantum Theory: Non-Relativistic Particle Systems', *Physics Reports* 144: 323–47.

—— and —— (1991), *The Undivided Universe: an Ontological Interpretation of Quantum Mechanics*, Routledge, London.

——, ——, and Kaloyerou, P. N. (1987), 'An Ontological Basis for the Quantum Theory: A Causal Interpretation of Quantum Fields', *Physics Reports* 144: 349–75.

——, ——, and Stuart, A. E. G. (1970), 'On a New Mode of Description in Physics', *Int. J. Theor. Phys.* 3: 171–83.

Bohr, N. (1961), *Atomic Physics and Human Knowledge*, Science Editions, New York.

Clifford, W. K. (1882), *Mathematical Papers*, Macmillan, London.

Coleman, S. (1977), 'Fate of False Vacuums: Semiclassical Theory', *Phys. Rev.* 15D: 2929–36.

Davies, P. C. W. (1984), 'Particles Do Not Exist: Quantum Theory of Gravity', in S. M. Christensen (ed.), *Essays in Honour of the 70th Birthday of Bryce S. DeWitt*, pp. 66–77, Adam Hilger, Bristol.

—— (1989), 'Why is the Universe Knowable?' in R. E. Mickens (ed.), *Mathematics and Science*, Oxford University Press.

Davies, P. G. (1981), 'The Weyl Algebra and an Algebraic Mechanics', Ph.D. thesis, University of London.

Eddington, A. (1958), *The Philosophy of Physical Science*, University of Michigan, Ann Arbor.

Einstein, A. (1924), 'Über den Äther', *Schw. Naturfor. Gesell.* 105: 85–93.

—— (1960), *The Meaning of Relativity*, Methuen, London.

—— (1969), 'Autobiographical Notes', in *Albert Einstein: Philosopher–Scientist*, ed. P. A. Schilpp, vol. 1, pp. 63–85, Cambridge University Press.

—— (1971), *The Born–Einstein Letters*, Macmillan, London.

Emch, G. G. (1972), *Algebraic Methods in Statistical Mechanics & Quantum Field Theory*, John Wiley, New York.

Frescura, F. A. M. and Hiley, B. J. (1980*a*), 'The Implicate Order, Algebras and the Spinor', *Found. of Phys.* 10: 7–31.

—— (1980*b*), 'The Algebraisation of Quantum Mechanics and the Implicate Order', *Found. of Phys.* 10: 705–22.

—— (1984), 'Algebras, Quantum Theory and Pre-Space', *Rev. Brasileira de Fisica*, special issue, '70 anos de Mario Schönberg': 49–86.

Grassmann, H. G. (1894), *Gasammelte Mathematische und Physikalische Werke*, Teubner, Leipzig.

Haag, R. and Kastler, D. (1964), 'An Algebraic Approach to Quantum Field Theory', *J. Math. Phys.* 5: 848–61.

Hamilton, W. R. (1831), Trinity College Dublin MS Notebook 1492–24.5 fo. 49.

—— (1967), *The Mathematical Papers*, vol. 3: *Algebras*, ed. H. Halberstarn and R. E. Ingram, Cambridge University Press.

Heisenberg, W. (1925), 'Quantum Mechanical Re-interpretation of Kinematic and Mechanical Relations,' *Z. f. Phys.* 33: 879–93.

Hyland, G. J. (1979), 'A Nonlocal Spinor Field Theory of Matter', *Gen. Rel. & Grav.* 10: 231–52.

Maxwell, J. C. (1954), *A Treatise on Electricity and Magnetism*, Dover, New York.

Segal, I. E. (1946), 'Postulates for General Quantum Mechanics', *Ann. Maths.* 48: 930–48.

Smart, J. J. C. (1964), *Problems in Space and Time*, Macmillan, New York.

Wheeler, J. A. (1967), 'Superspace & the Nature of Geometrodynamics', C. M. DeWitt and J. A. Wheeler (eds.), in *Battelle Rencontres, Lectures in Mathematical Physics*, pp. 242–307, Benjamin, New York.

10

Theory of Vacuum

DAVID FINKELSTEIN

1 Introduction

Today the vacuum is recognized as a rich physical medium, subject to phase transitions, its symmetry broken by non-vanishing vacuum values for several important fields akin to the permanent magnetization of a ferromagnet, and supporting the emission, propagation, and absorption of particles. A general theory of the vacuum is thus a theory of everything, a universal theory. It would be appropriate to call the vacuum 'ether' once again, as long as we remember its local Lorentz invariance. The most workable theories of the vacuum today are quantum field theories. In these the vacuum serves as the law of nature, as reviewed below. The structure of the vacuum is the central problem of physics today; the fusion of the theories of gravity and the quantum is a subproblem. Here I develop a quantum–space–time description of the vacuum. Einstein considered such a programme in the 1930s:

> To be sure, it has been pointed out that the introduction of a space–time continuum may be considered as contrary to nature in view of the molecular structure of everything which happens on a small scale. It is maintained that perhaps the success of the Heisenberg method points to a purely algebraic method of description of nature, that is to the elimination of continuous functions from physics. Then, however, we must also give up, by principle, the space–time continuum. It is not unimaginable that human ingenuity will some day find methods which will make it possible to proceed along such a

This paper is based in part on work supported by National Science Foundation Grant PHY-8410463, and by the Georgia Tech. Foundation Award 24AF63. During this writing I am pleasurably indebted to Roger Penrose and the Mathematical Institute, to Harvey Brown and the Philosophy Department, and to Wolfson College, Oxford University, for their hospitality, and to Georgia Institute of Technology for sabbatical leave and support. I thank Ray Sorkin and John Barrett for useful discussions, and S. R. Finkelstein for many helpful criticisms of the manuscript as well.

> path. At the present time, however, such a program looks like an attempt to breathe in empty space. (Einstein 1936: 378)

Einstein's stand is not *for* the continuum but *against* the quantum, which he considers no better than thermodynamics as a fundamental language. After a half-century of breathing exercises, many physicists who accept the quantum are also willing to leave the continuum:

> It may well be that the marriage of gravitation and quantum mechanics requires a few more drastic revisions of our ideas. For example, our description of space–time as a continuum may have to be replaced by a discrete granular structure at extremely short distance. (National Research Council 1986*a*: 97)

> It may be that local Lagrangian field theory is not the correct approach to quantum gravity. Perhaps, as some believe, the basic quantum quantities are not the variables describing a space–time continuum but a more discrete structure. (National Research Council 1986*b*: 74)

Quantum network dynamics (QND), which I sketch here, is such a discrete quantum theory, a generalization of graph theory. It retains and even extends the most basic principles of quantum field theory, such as relativistic locality, quantum superposition, and local Lorentz invariance; but initially it sacrifices unitarity for local finiteness and local relativity. QND is a fusion of quantum theory and space–time theory with a natural trial ψ vector for the vacuum network. This vacuum network has hypercubical symmetry, yet is exactly Lorentz-invariant, and supports an action principle leading to the Dirac equation for the motion of its holes. QND suggests that gravity is a macroscopic quantum effect (so that it should not be canonically quantized). The fundamental questions of QND are:

- What are the fundamental modules of the physical network?
- How are they interconnected in the normal phase?
- What are the disturbances in this topology during the dynamical evolution of the physical network?

The normal phase of the physical network I understand to be the vacuum. Here I propose answers to the first two questions; and a strategy for the third. This note may be regarded as a philosophical version of a more mathematical one already published (Finkelstein 1989), which gives further references.

2 Basic Principles

The main guiding principles that have evolved in this process of theory development are a form of monism and (as mentioned above) locality, superposition, relativity, and finiteness.

2.1 Topological Monism

There is just one activity going on in the world, whose fine structure we must determine. The topology of this activity is the only physical variable there is.

The quantum topology of QND is based upon a quantum set theory as classical topology is based on classical set theory; the underlying elements of quantum set theory anticommute rather than commute. The foundation of Peano's classical set theory as well as of his theory of the natural numbers, which may be read as the classical theory of a discrete future time axis, is an operator ι which forms the unit set $\iota\alpha = \{\alpha\}$ from any set α. In Section 4.5 I adjoin a quantum ι to the usual quantum linear space kinematics and Grassmann product. In QND ι is the causal successor operation.

2.2 Locality

Locality (of the Einsteinian kind) is the principle that fundamental concepts and laws connect events only to their infinitesimal neighbourhood. (More abstractly put, a class or property P of fields on a manifold is said to be local if, for every field f, f is in P if and only if for every point p of the manifold there is a neighbourhood N of p and a field f_N in P, such that $f = f_N$ on N.) For example, any property expressed by a differential equation of finite order with respect to space–time coordinates is a local property; the concept of a scalar field is a local concept; and the concept of a coordinate system is non-local, since it is not enough to verify its 1–1 property one neighbourhood at a time.

Locality convinced Newton that his own beautiful and powerful theory of gravity was merely phenomenological rather than fundamental, leaving it to Einstein to provide the first plausible local theory of gravity. Today experimental verifications of locality and superposition, whose results go beyond any classical local theories according to Bell's theorem, decide in favour of the quantum principle of superposition and against its classical reformulations, which lead to non-local theories. Thus, in judging physical theories, locality has been given more weight than accurate experimental predictions and common sense, when it came down to a choice, and this judgement has been vindicated subsequently by further experimental successes. The most advanced form that locality has taken in this century is the gauge theory of fundamental forces.

The locality principle seems to catch something fundamental about nature. I will not give it up here. Instead I seek out and eliminate some dissonances between locality and the other principles I have mentioned. Localized they become local finiteness, local superposition, local monism, and local relativity. The next paragraphs deal with the first three of these; we return to the fourth later.

2.3 *Local Finiteness*

Having learned that the world need not be Euclidean in the large, the next tenable position is that it must at least be Euclidean in the small, a manifold. The idea of infinitesimal locality presupposes that the world is a manifold. But the infinities of the manifold (the number of events per unit volume, for example) give rise to the terrible infinities of classical field theory and to the weaker but still pestilential ones of quantum field theory. The manifold postulate freezes local topological degrees of freedom which are numerous enough to account for all the degrees of freedom we actually observe.

The next bridgehead is a dynamical topology, in which even the local topological structure is not constant but variable. The problem of enumerating all topologies of infinitely many points is so absurdly unmanageable and unphysical that dynamical topology virtually forces us to a more atomistic conception of causality and space–time than the continuous manifold. Atomism has contended against the continuum principle for millenia, and a thoroughly atomistic conception of space and time dominated Islamic thought in the late Middle Ages. QND reinstates this atomism in a quantum form. The usual space–time manifold describes a special case of high order which is conditional and not truly fundamental in nature.

Continuum theories generally start from mathematically meaningless concepts, like the products of singular distributions. We are more likely to develop a correct theory of nature from a meaningful theory than from a meaningless one. Here we apply Einsteinian locality to the network. Instead of infinitesimal neighbourhoods, we speak of nearest neighbours, finite in number.

2.4 Local Superposition

The global quantum principle of superposition says that the physical properties of the system form not a Boolean algebra but a projective geometry. The fundamental duality of projective geometry between the space and its dual expresses a bimodality of the logic between input and output; if vectors represent inputs (to the endosystem from the exosystem, which includes the experimenter and environment) then dual vectors represent outputs, and the two annul each other for a forbidden transition. The term 'superposition' refers to the linear combination of such vectors.

A maximally informative predicate of the system is represented by a ray of a linear space, which serves as phase space, and a more general predicate by a subspace. Superposition thus leads not only to quantum indeterminacy, but to a continuity in possibilities which replaces the continuity of actualities that we have renounced for the sake of local finiteness: it interpolates an infinite number of possibilities 'between'—in a sense peculiar to quantum theory that is neither topological nor casual, but statistical or logical—two mutually exclusive possibilities. Quantum dilemmas have infinitely many horns, though each experimenter can see only two.

Since fundamental principles must be local, and the superposition principle is supposed here to be fundamental, we infer a principle of local superposition: the independent elements of the world and their causal connections to their neighbours also have projective, not Boolean, logics, subject to quantum superposition. We extend superposition to the fundamental space–time topological connections in this work. It then follows for everything else from the monism of Section 2.1.

Superposition and intensionality We extend the superposition principle to classes or predicates of quanta in this work. The principle of intensionality of classical logics associates a property with any given (finite) collection of entities, namely the property of belonging to that collection. A simple Slater determinant ψ vector describes a set of fermionic quanta and indeed corresponds to a Von Neumann class of quanta, represented by the projection operator upon the subspace spanned by the columns in the determinant. But a superposition of Slater determinants also describes an authentic quantum set of quanta, and yet defines no class of quanta in the Von Neumann sense. Evidently there are many more sets than classes in the Van Neumann logic, which thus violates intensionality by limiting superposition. I extend the Von Neumann logic in Section 4.4 to embrace such superpositions.

The sacrifice and rebirth of unitarity Local superposition permits us to formulate a quantum principle of local relativity. The Lorentz group in its most fundamental spinor form SL_2 acts upon the linear space of ψ vectors describing the immediate successors of an event, and respects their causal connection.

This leads to a most productive contradiction. The principles of local finiteness, relativity, and superposition are inconsistent with the standard principle of unitarity; for there is no finite dimensional unitary representation of the Lorentz group. There are many indications that unitarity is the one to sacrifice. Unitarity already contradicts Einsteinian locality and finiteness separately (if they are taken in their strongest senses), as follows.

- Unitarity is a non-local condition in field theory; it involves an integral over all space.
- Unitarity rests on the orthocomplement or negation operation of the quantum logic, and the orthocomplement of a finite-dimensional subspace is infinite-dimensional. A finitistic quantum logic with an infinite number of possibilities (like that of Section 4.4) cannot have a negation operation.
- Unitarity expresses conservation of probability in time and is meaningful only for entities that persist in time, like stable particles. Since the entities of which we assemble particles in QND are transient events, there can be no unitarity among the basic principles.

The unitary group is reborn in QND as a spontaneous breaking of the linear group of the linear space of quantum theory, a property of the condensed phase, like ferromagnetism.

We return to local relativity in Section 4.5.

2.5 Local Dynamics

The concept of dynamical law as a one-parameter group of global transformations is clearly non-local except in a purely timelike one-dimensional world, but Heisenberg's idea of dynamical law in its most general form, simply as a differential equation for the dynamical variables, is satisfactorily local. Its generalization to networks is a collection of operator equations relating variables at neighbouring events. We may use the action principle of quantum mechanics, despite its global appearance, to define such a local system of equations. A local dynamical theory will possess two algebras of variables: a kinematic algebra (KA) free of dynamical equations, and a dynamical algebra (DA) whose variables obey the dynamical equations. The dynamical algebra is a quotient of the kinematical one modulo the dynamical equations (DE):

$$\mathrm{DA} = \mathrm{KA}/\mathrm{DE}.$$

Quantum theories whose ψ vectors describe what happens at a single time may be called *synchronic*. They compress an input process that may actually be distributed over time into one initial instant. They assign inputs and outputs to different liner spaces, thus blocking their linear superposition and tacitly positing a superselection rule.

Theories whose ψ vectors are histories of what happens in all of space–time may be called *diachronic*. The Schwinger–Feynman quantum action principles are essentially diachronic theories. In the Schwinger (1970) source theory, a source is an element of a single linear space which I will call S, and describes both input and output (i/o) processes over the entire experimental space–time region, allowing their superposition and lifting the superselection law between them. Sources are called 'superlocal' because sources separated by timelike intervals, just like those separated by spacelike intervals, represent independent choices of the experimenter and are not related by dynamical equations. The distinction between input and output is made within S purely on the basis of the sign of the frequency.

In ordinary practice, S is an algebra with the addition operation $+$ for quantum superposition; a constant i that (like 1) stands for the vacuum as an element of S and for a quantum phase shift of a quarter-period as a multiplier on S; and a product $\vee$ (symmetric for bosons, antisymmetric for fermions) for the joint action of sources. If the algebra S is to be free of superselection laws, then its underlying ring of coefficients must be commutative; we provisionally assume the complex numbers $\mathbb{C}$ as is usual.

Each vector of the dual space S^D assigns a transition amplitude to each source. It therefore expresses a dynamical law or force law, and is called a field for short. We write F for an algebra of fields dually isomorphic to S, such that F and S are included in each other's duals. Each field determines a system of propagators, and one field $\langle \mathrm{vac}|$ in particular determines the vacuum propagators. This is called the vacuum field or the law of nature. It stores the forms of all the phenomenological interactions, and the charges, masses, and other coupling constants of the experimental quanta.

Feynman's path amplitude represents the dynamical law and is therefore an element of F, not S, defining the vacuum field.

Since histories and variables assign numerical values to each other, they are categorically dual. The algebra of dynamically allowed histories (DH) is therefore a sub-algebra, the dual concept to a quotient algebra, of the algebra of kinematically allowed histories (KH); namely the sub-algebra that annuls DE:

$$\mathrm{DH} = \mathrm{KH} \backslash \mathrm{DE}.$$

DH is just the phase space of the theory.[1]

3 The Marriage of *h* and *c*

The chasm between h and c is the key problem of physics today. I once took it for granted that the way to bridge it is to quantize gravity, and assigned quantum logics and relativity to different stories of the tower of physics, reversing their order (compare the right-hand side of Fig. 1 with the left-hand side of Fig. 2) but maintaining their separation (Finkelstein 1968). Now there seem to be no floors that divide them (Fig. 2, right-hand side).

There may be no further need to unify quantum theory and gravity if gravity and inertia are already macroscopic quantum effects. This suggestion evolves from a growing list of parallels between the stories of relativity and quantum theory which I have used in teaching them in recent decades:

- Each of these two theories has its own new fundamental constant (h and c) and a correspondence principle recovering the old physics in the transition to a singular limit ($h \rightarrow 0$, $c \rightarrow \infty$).
- Each is fundamentally epistemological, in that it sets a universal limitation upon the i/o processes that link us with our experimental

[1] John Barrett, personal communication.

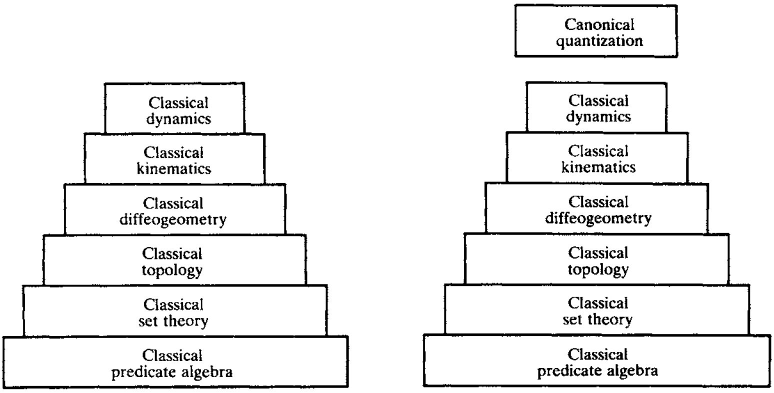


FIG. 1 Architectures of classical and quantum field theory.

Canonical quantization is added to an essentially finished classical field theory (left-hand tower) to make quantum field theory (right-hand tower).

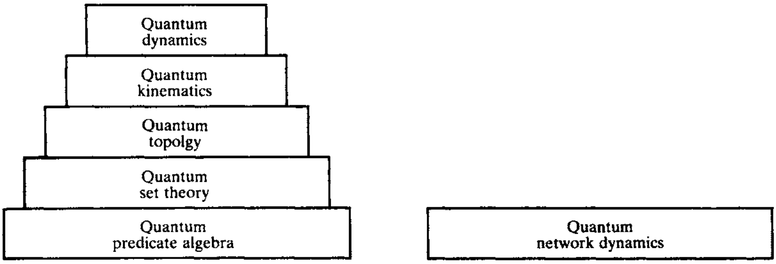


FIG. 2 Architectures of QND.

The tower on the left represents the plan to transfer the principle of superposition from the top to the bottom of the tower of Fig. 1. But because many previously separated concepts are united in the successor operator ι, all stories of the left-hand tower telescope into one on the right.

systems (h limiting the determinacy of these processes, c the signal speed).

- In each singular limit, basically non-commutative processes become commutative (p and x determinations for $h \to 0$, boosts for

$c\rightarrow\infty$). The novel non-commutativity is expressed by a pair paradox (the two slits, the two twins).
- Each constant links time t to another fundamental physical variable (energy E for h, space x for c), so that the new theory is conceptually more unified than the old in an unanticipated way (clocks then defining energies and distances respectively).
- Each theory extends the principle of relativity to a wider class of transformations and a richer class of experimenters (Dirac's quantum transformation theory, Einstein's special relativity).
- Each is expressible completely as the theory of a transfer relation for i/o experiments (the allowed transition of quantum theory, the causal relation of special relativity).
- The signals that one uses operationally to define the space–time points and causal relations of relativity are actually ensembles of quanta.

(It is not difficult to extend this parallel from h and c to Boltzmann's constant k, with quantum theory, space–time theory, and thermodynamics appearing as successively coarser statistical descriptions of the same processes.) This suggests that space–time vectors and ψ vectors belong not to widely separate levels, as in Fig. 1, but to the same level as in Fig. 2. QND is the simplest theory of that kind I could make.

3.1 *Incoherent and Coherent Quantization*

This presents a challenge: how do we extract space–time vectors from quantum ψ vectors? The two have different interpretations. The space–time vector can be the result of an observation; the ψ vector cannot, but describes the observation itself. Space–time vectors have classical (commuting) properties, while ψ vectors support quantum (non-commuting) ones. Space–time vectors (or suitable scalar products thereof) are themselves observable in single experiments, while ψ vectors are merely probability amplitudes.

From our experience with $h\rightarrow 0$, there are two ways to extract macroscopic classical variables from ψ vectors as $\tau\rightarrow 0$: incoherent and coherent. In both we use large quantum ensembles to produce macroscopic behaviour, but these ensembles may be incoherent (as in the Thomas–Fermi model of the atom, where we add probabilities) or coherent (as in superfluidity, where we add amplitudes). In one case what emerges as the classical variable is a probability distribution ρ, in the other a probability amplitude distribution ψ. We may call the respective inverse processes (going from macroscopic theory to quantum) coherent and incoherent quantization, according as they start from experiences with or without quantum phase data.

Canonical quantization of space–time structure (like that of a harmonic oscillator or a hydrogen atom) would yield a system having the usual space–time (or oscillator or atom) as a high-quantum-number excitation. In the classical limit of many excitations we lose all quantum phase information. Canonical quantization thus begins from a theory without quantum phases; I call it an incoherent quantization.

On the other hand, for example, we would make a correct theory of Josephson potentials not by canonically quantizing them but by coherently quantizing them, because they are themselves quantum phases.

Nambu has long ago pointed out that it is likely that the physical vacuum is actually a quantum condensation, a low-quantum-number limit, and this is now rather widely accepted. I propose that space–time structure too is a macroscopic quantum effect, and that gravity (specifically, the spin metric $\sigma_{s\Sigma\Sigma'}$ of Infeld and van der Waerden) is a quantum phase, an order parameter of the condensate. Then the quantum theory must relax to our usual space–time, not be excited to one, and one should not quantize $\sigma_{s\Sigma\Sigma'}$ canonically, any more than we do the velocity of liquid helium II, but should seek a quantum dynamical system whose macroscopic quantum ψ vectors define $\sigma_{s\Sigma\Sigma'}$. Since we start from ψ already carrying quantum phase information, this may be called a coherent quantization.

4. The Idea of the Network

I give now the simplest classical concept of space–time structure I can imagine on which we may practise coherent quantization.

4.1 Principle of Maximal Information

I am guided in my choice by the superposition principle. Since fundamental descriptions must admit superposition, they must give maximal information. I call this heuristic rule the principle of maximal information.

To put it more paradoxically, the only theories that may succeed at the atomic level are those that run the maximum risk of failure. For example, the Schrödinger equation of quantum theory concerns atomic statements of maximal information, not maximal generality: statements of maximum risk, therefore, not minimum.

To put it more formally, since equations give more information than relations of inequality, I insist on an equation-based theory, not a relation-based one.

For example, the lattice logic of Von Neumann is not much more suitable than thermodynamics as a language for any fundamental theory, no matter how useful it is for some didactic purposes, since its elements too have lost their quantum phases. But the Copenhagen quantum theory is relatively free from this criticism, in that it respects all the quantum phases of the endosystem and ignores only those of the exosystem.

I apply this principle in the next section.

4.2 Classical Causal Network

In the most familiar formulation of a causal structure, the elementary entities are point events α, β, . . . and they support a causal relation $\alpha C \beta$. At once, locality compels us to reject C as fundamental variable in favour of its germ, the local connection relation $\alpha c \beta$, 'α connects to β' (it being understood that this connection is meant to be immediate and to define the sense of time from past to future). When $\alpha c \beta$ we say that α is an input to β, and β is an output from α.

Then maximum risk compels us to reject the causal *relation* $\alpha c \beta$, which says little about β, in favour of the dynamical *equation*: $\beta = \iota(\alpha_1 \vee \ldots \vee \alpha_N)$, which we read as saying that β is the successor of events $\alpha_1, \ldots, \alpha_N$, and which defines β maximally. This change in expression makes no difference in the classical theory but affects the quantum theory to come. The operator ι generalizes Peano's successor operation of Section 2.1.

In a dynamical theory it seems necessary to divide the events α into dynamical processes δ (analogous to timelike future vectors) and space–time events α. In the language of graphs, these are oriented edges and vertices. In the simplest and provisional 'one-particle' theory, one of each enters into the dynamical equation:

$$\beta = \iota(\delta \vee \alpha). \tag{1}$$

An equation of the form (1) giving the N successors α_Σ, $\Sigma=1,\ldots,N$, of an event α, is called an N-ary node, or N-ode. A binary node, for example, is represented by a graph such as

$$<$$

Any collection of nodes defines a classical causal *network*. The classical network is invariant under exchange of the inputs to any event, and each N-ary node has the group S_N, the symmetric group on its N final events.

Mathematical topology today still describes timeless undirected spatial connections. Its module is a simplex, a set of points with all possible two-way connections. This is because Euler could cross the bridges of Königsburg in both directions. But our paths in space–time are one-way streets. It would be atavistic to think of the world topology as simplicial; physical topology needs to be built on asymmetric physical connections, with nodes replacing simplices.

4.3 Classical Sets

I express this prequantum network theory in a set algebra SET adapted to subsequent quantization. The sets of SET are ancestrally finite; this is the only kind of set we admit. Let 1 be the null set; 0, the undefined set; $\vee$, the disjoint union of classical sets, giving 0 when the sets are not disjoint; ι, the monad operator $\iota\colon \alpha\mapsto\iota\alpha=\{\alpha\}$, forming the monad (= unit set) of α.

When we read the symbol α as a set, $\iota\alpha$ is the unit set or monad of α, $\iota\iota\alpha$ is the monad of the monad of α, and so forth. When we read α as a class or predicate, as we are entitled to do by the principle of intensionality, $\iota\alpha$ is an identity predicate, the predicate of being α, and $\iota\iota\alpha$ is the predicate of being the predicate of being α, and so forth. We omit an intersection operation because it gives us no new sets; we omit complementation because the complement of a set in SET is not finite, hence not in SET. This language is rich enough to write our sets, but not to say much about them; we enlarge our vocabulary for dynamics in Section 4.4.

The concepts of 0 and $\vee$ are C. S. Peirce's (Peirce 1931–5), but the symbols are Grassmann's (Pierce uses ∞ and $+$; 0 is called θ by Eilenberg 1976); '1' means 'nothing' but '0' means nothing. ι is Peano's (1988). Set theory, primarily concerned since Cantor with the infinite, is often called the theory of the $\in$ relation. The set theory that we actually need is the theory of Peano's ι operation, which is more informative; for $a \in b$ does not define b, but the stronger statement $\iota a = b$ does. The relational view of set theory has been the greatest obstacle in the way of the present network theory, and the superposition principle and the maximum information principle of Section 4.1 seem to have at last cleared this obstacle.

Each event is now the set of its inputs. Each set represents a classical dynamical network. The successors of an N-ode form a set of N elements or N-ad. SET is the algebra of kinematically allowed histories of the classical network.

Since set theory is a universal language for mathematics, it at first sounds trivial to say that the world may be described in it. But usually the same set symbols are given many different interpretations in physics, according to need; just consider, say, the variety of physical quantities that are represented by numbers: masses, charges, coordinates, and so on. The actual language is then not set theory, but a richer one with many logically independent concepts adjoined to set theory, such as distinguished sets, or proper individuals. Here we give each of the three set symbols 1, $\vee$, ι a single uniform physical interpretation (and 0 none at all). This is a genuine unification. To make this possible, we have had to adapt the set theory slightly. Our physical set theory does not have exactly the same primitives $\cup$, $\in$ as the usual mathematical set theory. The usual $\varnothing$ survives as 1, but $\cup$ and $\in$ have been replaced by $\vee$ and ι to meet the needs of QND.

4.4 Quantum Sets

In the quantum theory the classical sets of Section 4.3 are regarded as ψ vectors (coherent states) of the quantum set. The unit sets anticommute. The most general ψ vector or ket for a quantum set representing a quantum network is called a *qet*. Qets form a Grassmann algebra QET with zero 0, unity 1, product $\vee$, and linear operator ι interpreted like the same symbols of SET, and with one new operation, $+$, interpreted as quantum superposition. QET is generated from 1 by ι, $\vee$, and $+$. At first I also assume a complex imaginary i among the coefficients of the quantum theory; this is expected to be replaced at a later stage by a linear operator i with an approximate superselection law. I write $\iota\alpha$ also as the qet bracket $\langle\alpha|$.

The qet bracket combines the ket bracket $|\alpha\rangle$ of Dirac, the set bracket of Cantor, and the extensions of Grassmann. They nest like sets, add like kets, and multiply like extensions (the product $\vee$ combining the Grassmann product of fermionic quantum kinematics and the disjoint union of set theory).

I take QET to be the algebra KH of kinematically allowed histories of the quantum network.

Duality symmetry demands an operator ι^D to go with ι, obeying the same laws for dual qets that ι does for qets. Let QET_N be the subspace of QET consisting of grade-N qets, so that $\iota : \text{QET} \rightarrow \text{QET}_1$ is bijective; we define ι^D as the inverse of ι on QET_1 and 0 on QET_N for $N \neq 1$. With ι^D we can make the most general dual vector from 1; by combining these with the vectors of QET we can make the most general finite linear operator on QET. This is enough for QND.

The quantum kinematics is expressed by $+$ and $\vee$, and local causality by $\vee$ and ι. In the usual quantum theory these principles are assigned to different stories of the tower of physics of Fig. 1 ($+$ coming last, hence on top, as in archaeology). The fusion of relativity and quantum kinematics in QET permits the free intermixing of these principles.

QET is the algebra of both classes and sets. That is, I stipulate that the quantum properties of any quantum system with Fermi–Dirac statistics described in a linear space H constitute not merely the projective geometry of H but the Grassmann algebra over H. Grassmann had no interpretation for non-product elements of his algebra of higher grade, calling them 'imaginary'. Quantum superposition provides the interpretation of these elements that Grassmann lacked.

This fusion is crucial for QND and has the following curious consequences:

Second Fermi–Dirac quantization maps a linear space H into the Grassmann algebra $\mathbf{V}H$ over H, and the vectors ψ of H into generators of $\mathbf{V}H$. Since ι does these, we may call it a quantization operator. (Quantification would be a better name than quantization, since ι leads from a theory of 'True or false?' to one of 'How many?'.) Quantization is ordinarily not usable as a dynamical operator, since it enlarges the Hilbert space, but is carried out only by the theorist in setting up the kinematics, perhaps once or twice or (nowadays) thrice during the invention of a theory. For networks ι is a dynamical operator because $\mathbf{V}\text{QET} = \text{QET}$. Now ι becomes something that can go on in nature. Quantization occurs at least 10^{26} times per second in the vacuum network, according to the estimate of Section 5.6.

The quantum network may be regarded as a relativistic quantum generalization of a neural network. A node consisting of an event with its N successors corresponds to a neuron, but to one that exists only long enough for a single excitation; otherwise it would define a rest-frame.

4.5 *Local Relativity*

This is a form of the principle of equivalence: special relativity (specifically, Lorentz invariance) works in the tangent space. In the quantum network theory, events are subject to the same kind of quantum kinematics ordinarily used for fermions in quantum mechanics, except that events, being entirely transient entities, obey a non-unitary quantum theory. I call the following postulate the quantum equivalence principle, where now equivalence of two quantum entities means that no physical variable is changed by their permutation:

> The inputs to an event are equivalent quantum entities
> with Fermi–Dirac statistics.

This principle is incorporated in our interpretation of the causal node. In the quantum theory the group of a binary node is no longer S_2 but SL_2, identified with relativistic SL_2. The quantum equivalence principle gives rise in the following condensation to the Einstein equivalence principle, that SL_2 is a valid symmetry in the tangent space of the manifold. But the binary node does not support time reversal T or parity P; it is made of two-component irreversible spinor entities, not reversible vector ones. There is no need to break P or T in QND; the problem is to create them, as by the following condensation.

5 The Vacuum Condensate

I now form a trial vacuum for QND.

5.1 *The Basic Module*

The first thing to account for is the group SL_2 of the infinitesimal macroscopic causal relation of the tangent space. I infer that each event in the normal network has two successors which transform as a spinor. That is, in the actual world, events are related to their successors not by infinitesimal vectors but by finite spinors. I write two basic spinors as $\langle\Sigma| = \langle\uparrow|, \langle\downarrow|$. The two successors $\langle\psi_\Sigma|$ of the event ψ are given by

$$\psi_\Sigma = \langle\langle\Sigma| \vee \psi|.$$

Call these finite spinorial analogues of tangent vectors 'chronons' for short. After Z chronons, a path from an arbitrary event ψ may arrive at any of the events $\psi_{\Sigma, \ldots \Sigma'}$ with Z binary indices. I write (Σ) for a sequence of Z indices of the type Σ. Then a basis for the algebra of kinematically allowed paths is defined inductively by

$$\psi_{\Sigma'(\Sigma)} = \langle\langle\Sigma'| \vee \psi_{(\Sigma)}|.$$

Each of the polyspinors $\psi_{(\Sigma)}$ describes not merely an endpoint but an entire path, which may be developed from the endpoint by unbracketing it.

The next thing to understand is the dominance of the vector representation $D(\frac{1}{2}, \frac{1}{2})$ in local space–time structure. The space–time manifold (actually, its tangent bundle) seems to be made of vectors, not spinors. I infer that two fermionic chronons in sequence form a pair described by a qet with the vectorial transformation law of $\langle\Sigma * \Sigma|$ (chronon and complex conjugate chronon). Hence it follows that the operator ι is antilinear. Then spinors and 'antispinors' (complex conjugated spinors) alternate in sequence in $\psi_{(\Sigma)}$. I write σ for the pair index $\Sigma * \Sigma$. The module is described by a pair spinor $\langle\sigma|$ and the path with even Z has the form $\psi_{(\sigma)}$. Thus the basic module of our trial vacuum has one initial and four final events, but less symmetry than the 'pentacle' considered earlier (cf. Finkelstein and Rodriguez 1984, 1986).

The actual vacuum may well be made of superpositions of several different modules, with complex amplitudes to be determined.

5.2 *The Topology of the Vacuum*

Finally, we must account for the macroscopic observable nature of space–time vectors. They are not quantum ψ vectors, or states, describing quantum channels, but serve as parameters, or even (for one-particle theory) as observables themselves. To make such concrete vectors we must 'condense' the ψ vector. I infer a bosonic condensation of a large number of modules into a four-dimensional quantum lattice. In superconductivity, electrons pair with rather remote P transforms; in the vacuum, chronons pair with neighbouring C transforms. The algebra of dynamically allowed paths DH in the vacuum is then the symmetric subspace of the tensors $\psi_{(\sigma)}$. I write Σ_+ for the symmetrizing operator, normalized to be idempotent; $\{\sigma\}$ for a symmetric sequence of σ, the collective index of a symmetric tensor; and $\psi_{\{\sigma\}}$ for the symmetrized tensor $\Sigma_+\psi_{(\sigma)}$. The paths in the condensate have the form $\psi_{\{\sigma\}}$, with a symmetric sequence of $Z/2$ σ-indices. It means that two paths which differ only in the order of their elements or chronons lead to the same event. This specifies the connection of the basic nodes of the network in the normal phase.

One may identify the coordinate operator x^σ on this reduced algebra with the Bose–Einstein creation operator $\Sigma_+(\sigma|\Sigma_+$, and the momentum operator ∂_σ with the dual destruction operator, so that their commutation relations follow not from the differential calculus but from simple algebra. (For $N=2$ there is also the possibility that x and ∂ are linear combinations of the creation and destruction operators.)

A $\psi_{\{\sigma\}}$ is defined by four integers giving the number of indices of four independent kinds—four occupation numbers, in other words. These basic events thus form a four-dimensional hypercubical network which may be regarded as a synthesis and extension of Feynman's two-dimensional draughts-board[2] and Penrose's (1971) two-dimensional spin network. A simple candidate for the Dirac dynamics of a spin $\frac{1}{2}$ quantum is given elsewhere along the lines of Feynman's draughts-game (National Research Council 1986*b*). What moves on the network are not separate draughts-pieces, however, but topological defects in the network itself. Such a network theory is more unified than the most unified field theories.

Such a quantum condensation may account for the following features of standard physics, which would otherwise be incomprehensible in network theory.

5.3 *Supermobility*

The Law of Inertia and momentum conservation are problems in any discrete space–time, since the network is not invariant under

[2] Feynman and Hibbs (1965; ch. 2); see also the Nobel Address of R. P. Feynman (1972: esp. 168–9).

translation. The corresponding momentum transfer in a crystal is an Umklapp process. Since we see no Umklapp in the vacuum up to enormous energies E, we may be inclined to place an extremely low upper bound $1/E$ on the fundamental cell size or chronon of the network. But this would be incorrect if inertia is a macroscopic quantum effect. Since Newton's First Law states that the mobility of a particle, usually defined as $(\partial[\text{force}]/\partial[\text{velocity}])^{-1}$ at zero velocity, is infinite in zero-temperature vacuum, we may call the Law of Inertia *supermobility*, to emphasize that it belongs to the same family as the other more recently discovered macroscopic quantum phenomena, superfluidity and superconductivity. Since the event-pairing occurs between neighbours in space–time rather than momentum space, the vacuum in thermal equilibrium is presumably not a two-fluid system like liquid He II.

5.4 Macroscopic Vectors

Momenta and gauge vector fields are macroscopic dynamical variables and yet have to be made out of microscopic quantum ψ vectors like the spinors ψ^{Σ} of each cell, which are not: Such 'condensation' of Ψ vectors into macroscopic variables occurs in the other two main superflows. Here spinors ψ^{Σ} and antispinors ψ^{Σ^*} must pair and condense to vectors $\psi^{\Sigma\Sigma^*}$. All physical time–space vectors v^s are regarded as macroscopic ψ vectors $v^{\Sigma\Sigma^*}$ of condensed aggregates of Σ–Σ^* pairs, present only in the low-temperature phase, vacuum I. The usual Minkowski coordinate and derivative operators may be modelled in network theory by the creation and annihilation operators for such pairs. There must be a fundamental constant with the dimensions of time to convert our pure numerical operators to physical time units. I designate this ultimately small quantum of time or chronon by ת (tav), so that it is as far in the alphabet as it is in magnitude from Cantor's alephs. In the limit as a fundamental time scale ת→0, the algebra of these creation and annihilation operators approaches that of the space–time coordinates and energy momenta.

5.5 Reality

Spinors and pair spinors are complex, yet time–space and gauge vectors are real. If this is a spontaneous breaking of gauge invariance as in superconductivity, then it is necessary to Hermitian-symmetrize the module $\langle\sigma|$ before forming the sequence $\psi_{\{\sigma\}}$; I leave this step for later.

5.6 Time Asymmetry and Parity Violation

The chronons $\langle\Sigma|$ do not support time reversal and parity. In the ordered phase, Vacuum I, we may use the predecessors of an event to define the T transform of its successors, and we may use the next-nearest successors to define the P transform of the successors. In the disordered phase, Vacuum II, these symmetries disappear. We identify this disappearance tentatively with the weak P and T violation. Then on dimensional grounds (which are not reliable in a theory with so many large dimensionless numbers available) the critical temperature T_C and the chronon ɲ should be closer to $M_W \sim 10^2$ GeV, the mass of the W particle, than to the Planck mass $M_P \sim 10^{19}$ GeV. This suggests that balls of Vacuum II are produced in experiments today as well as in the creation of the universe. The experimental implications of this are not yet known.

5.7 Internal Particle Symmetries

To convey the philosophy of this approach to nature, let me sketch a tentative programme for understanding the inertial particle symmetries in network dynamics. Fortunately, these all seem to be gauge symmetries, and may therefore act on defects in the vacuum network, in the way that the Burgers vector does on defects in a crystal. The path from gauge fields to topology is already rather well marked in a simplicial theory of space–time:

- Every gauge field is the commutator of a momentum 4-vector with itself.
- Momentum is represented by the exterior derivative of differential geometry.
- The exterior derivative is a limit of the coboundary operator of cohomology.
- The coboundary operator and its commutation relations express the topology of the complex (Finkelstein and Rodriguez 1984, 1986).

If we can chase through this line of connections in the present network theory of space–time, we will arrive at a topological theory of the gauge fields.

Since we already have the network correspondent to spin, and none to charge or the other coupling constants, the first step will likely be a theory of the transport of spin in the vacuum network, describing which defects in the network produce torsion and curvature in space–time. That would constitute a quantum theory of gravity.

The next part of the problem will be to extend the gauge theory of nets from gravity to the other interactions. This will require us to assign defect transformations to the known internal symmetry generators.

In the most immediate model of colour SU_3 symmetry within network theory, each event supports not only the two 'external' chronons already described, which link up into long fibres of the macroscopic vacuum, but also three additional microscopic 'internal' chronons, which do not. This vacuum resembles a fur-covered draughts-board. Although the global structure of the network is four-dimensional, as if the nodes were binary, each node is actually a 5-ode. The colour group then mixes the internal chronons. This discrete quantum version of Kaluza–Klein theory puts a heavy responsibility on the network dynamics, which must bind just three internal chronons to every two external ones, but QND suggests that such a structure actually exists. More generally, colour ought to label a natural trio of distinguishable defects in the vacuum network which are isomorphic but not mixed by SL_2; I have not found such defects yet, except for the loops already described.

In the vacuum network every event has two local SL_2 groups. One is the symmetry of the binary node already discussed; the second is a T-image of the first, mixing two predecessors. Its existence depends on the condensate structure. In the vacuum network one combination of these two SL_2 generators survives as an exact global relativistic symmetry; perhaps this leaves another combination for an approximate internal particle symmetry.

5.8 Concepts of Dynamics

Once we have fixed a kinematics, like QET, the problem is a dynamical law. I do not trust the concept of dynamics of Section 2.5 very far. In the diachronic quantum descriptions, the dynamical law is merely an element of the space F of fields, and the concept of an eternal constant dynamical law or vacuum seems atavistic. In this section I speculate on alternative concepts of dynamics.

It is tempting to seek principles influencing, if not determining, the law within the one surviving story of the physics building of Fig. 2, rather than adding a new story. The only natural alternative to an eternal law that I can imagine at present is an evolving one, along the lines suggested by C. S. Peirce in his theory of the First Flash (his term was significantly more accurate than 'Big Bang', since light is older as well as faster than sound; I shall use it for the first quantum event of the Big Bang), and in other writings on his evolutionary cosmology. One first step toward an evolutionary dynamics is to move the law from its classical locus outside the dynamical theory to within the realm of evolving entities; this step is already taken in the diachronic quantum theories, where the ambient vacuum is the law. A next step, according to one speculation, is to organize the vacuum out of quasilocal genetic elements capable of reproduction and selection. Schrödinger taught us to think of organic structure as encoded in an 'aperiodic crystal'; perhaps we must learn to think of vacuum structure in the same way. The module constructed in Section 5.1 is the simplest vector structure; the actual modules may carry much more information.

A variable dynamics is already built into ordinary physics in the concept of external field. For example, Newton's law of motion for the earth is contingent on the absence of strong gravitational waves; if we include the gravitational field in the system, Newton's law is no longer even a candidate for the dynamical law. Similarly, Dirac's equation governs a hydrogen atom only in the absence of external muon beams. In general, quantum theory insists on dividing the system into (I will say) exosystem and endosystem, but does not tell us where to put the quantum partition. Each experimenter within the same exosystem is represented in quantum theory by a complete set of endosystem variables. The concept of dynamical law is relative to the quantum partition; by shifting this interface we may convert an external field, with no evolutionary equations, to an internal one, governed by a dynamical law, although a most complex one, since it governs what was the experimenter before the shift. We know that external fields can change the charges and masses in the endosystem. It seems most likely that the entire dynamical law of the endosystem is a surrogate description of the exosystem.

It requires an extension of relativity to deal with such shifts.

- First relativities, including Einstein's, fix endosystem and exosystem, and merely permute the descriptors among each other. There is really only one Experimenter in classical mechanics, who measures everything.
- Second relativities, like the Dirac quantum transformation theory, are more general; they fix the endosystem but change the experiment: one may determine the x component of spin, another the y. There is a genuine plurality of experimenters in quantum theory; but they all look at the same endosystem.
- Third relativity shifts the quantum partition between exosystem and endosystem: one endosystem is the nucleus of the atom, another includes the electrons as well. The third relativity principle is: Descriptions of the system with different partitions are mutually consistent. To my knowledge, this was first used by Von Neumann in his theory of measurement, when he postulates that the determinations by an experimenter I of an experimentee III do not depend on whether the apparatus II is lumped with I or III. Since third relativity is properly stronger than the relativity principles of general relativity, it may conceivably provide a stronger restriction of the dynamical principle.

6 Summary

In diachronic theories the vacuum carries all the information of the dynamical law. QND is a diachronic theory that fuses quantum and relativity principles in a fundamental and locally finite way. What is usually called unification keeps the vertical structure of physics more or less intact, and seeks a horizontal merger on the top levels. In QND particles are defects in the vacuum network and their unification is a consequence of a vertical merger of all the levels of the usual quantum field theory (quantum predicate algebra, quantum kinematics, quantum topology, and quantum chronometry) into the one of QND. Minkowski space–time is a macroscopic quantum effect, with an underlying periodic condensation. Defects in this periodic structure give rise naturally to gauge fields. The hypercubical lattice of qets constructed here is proposed for this periodic structure.

Network theory requires us to adapt quantum logic to classical set theory in order to fuse them. The quantum logic of classes I use here fulfils principles of intensionality and extensionality that Von Neumann's lacks. The set theory is based on disjoint union and the ι operator instead of union and the $\in$ relation, and is isomorphic to the class theory.

Networks provide a fundamental theory of quantum spin: a spin transformation permutes the two causal inputs to an event among themselves. The 'quantum two-valuedness' of Pauli counts the two successors of each event.

References

Eilenberg, S. (1976), *Automata, Languages and Categories*, vol. B, Academic Press, New York.

Einstein, A. (1936), 'Physics and Reality', *J. Franklin Inst.* 221: 313–47.

Feynman, R. P. (1972), 'The Development of the Space–Time View of Quantum Electrodynamics', in Nobelstiftelse, *Nobel Lectures—Physics, 1976–1970*, Elsevier, Amsterdam.

—— and Hibbs, A. R. (1965), *Quantum Mechanics via Path Integrals*, McGraw-Hill, New York.

Finkelstein, D. (1968), 'Matter, Space and Logic', in R. S. Cohen and M. W. Wartofsky (eds.), *Boston Studies in the Philosophy of Science*, vol. 5, pp. 199–215, Reidel, Dordrecht, Holland.

—— (1989), 'Quantum Net Dynamics', *Int. J. Theor. Phys.* 28: 441–67.

—— and Rodriguez, E. (1984), 'Relativity of Topology and Dynamics', *Int. J. Theor. Phys.* 23: 1065–98.

—— (1986), 'Algebras and Manifolds', *Physica* 18D: 197–208.

National Research Council (1985*a*), *Physics through the 1990s: Elementary-Particle Physics*, National Academy Press, Washington, DC.

—— (1985*b*), *Physics through the 1990s: Gravitation, Cosmology, and Cosmic Ray Physics*, National Academy Press, Washington, DC.

Peano, G. (1888), *Calcolo geometrico secundo l'Ausdehnungslehre di H. Grassmann, proceduto dalle operazioni della logica deduttiva*, Bocca, Turin.

Peirce, C. S. (1931–5), 'On an Improvement in Boole's Calculus of Logic', *Proc. Am. Acad. Arts and Sciences*, 7: 250–61; 'Upon the Logics of Mathematics', *Proc. Am. Acad. Arts and Sciences*, 7: 402–12. Reprinted in C. Hartshorne and P. Weiss (eds.), *The Collected Works of Charles Sanders Peirce*, vols. 1–IV, Cambridge, Mass.

Penrose, R. (1971), 'Angular Momentum: An Approach to Combinatorial Space–Time', in T. Bastion (ed.), *Quantum Theory and Beyond*, Cambridge University Press.

Schwinger, J. (1970), *Particles, Sources and Felds*, Addison-Wesley, Reading, Mass.

11

Topology of the Vacuum

MICHAEL ATIYAH

1 Introduction

The fundamental concepts in physics concerning the nature of matter and force, and their relationship to space–time, have always found their most precise and productive form in a mathematical framework. Radical new developments in physics have had a major impact on mathematics and vice versa. The most striking example of this interaction in the twentieth century is certainly Einstein's General Theory of Relativity, in which gravitational force is interpreted as the curvature or distortion of space–time. This has provided a paradigm for the interpretation of all fundamental forces and has been the mechanism leading to 'unified field theories'.

In a field theory, dealing with forces such as gravitation or electromagnetism, empty space—the vacuum—alters its nature in the presence of the relevant force. This alteration takes the form of a geometrical distortion which will change the trajectory of any material object that enters the appropriate position of space. This is what we might call the *classical theory* of the vacuum, classical in the sense of pre-quantum theory.

Quantum theory requires us to modify our physical ideas in profound ways, and we must inevitably start by re-examining the classical vacuum. Quantum field theory attempts to deal with the classical force-fields in a quantum-mechanical way, and the *quantum vacuum* that emerges from this theory is a complex and mysterious structure which stretches mathematics to its utmost limits. Quantum fields fluctuate and convert themselves into particles in a bewildering manner, indicating in particular the fact that the conventional separation between force and matter cannot be maintained.

A genuine quantum vacuum should therefore be devoid of both matter and fields of force. There should be no particles and no geometric distortion of space–time. This ultimate vacuum might appear so empty of features as to be mathematically trivial and uninteresting. It is a further surprise of quantum theory that this is a misleading conclusion. The quantum vacuum does indeed have interesting geometrical features, but these relate not to the traditional geometry of Euclid, Riemann, etc., involving measurement, but to that modern branch of the subject known as *topology*, which is concerned with qualitative properties of space. Unlike measurement, which can be conducted on a small local scale, topological features are visible only on a 'global' scale.

This relation between topology and the quantum vacuum has been recognized only quite recently, and its full implications are just now being explored. It is still too early to predict how this will alter our understanding of the universe, but it is clear that we have reached a deeper level in the dialogue between mathematics and physics.

The purpose of the paper is to indicate briefly what topology is and how it relates to quantum theory. Much of the material described is due to Edward Witten of the Institute for Advanced Study in Princeton (Witten 1989).

2 The Bohm–Aharonov Effect

A fundamental experiment, first suggested by Bohm and Aharonov, consists in sending a beam of electrons round a solenoid carrying a magnetic flux (Aharonov 1986). The experiment shows that the electrons exhibit interference patterns, depending on the strength of the magnetic flux. Thus, the electrons are physically affected by the magnetic field even though the field lies entirely inside the solenoid and the electrons travel in the exterior region (Fig. 1).

This Bohm–Aharonov effect therefore shows that, even in a force-free region, there are physical effects. These effects are quantum-mechanical, since they correspond to phase shifts in the wave function of the electron, and they have a topological origin since the force-free region has a cylindrical hole in it. This exhibits clearly the basic relation between quantum theory and topology, particularly in relation to notions of the vacuum.

Topology may be roughly defined as the study of 'holes' and related phenomena. The size of a hole and its exact location is irrelevant in

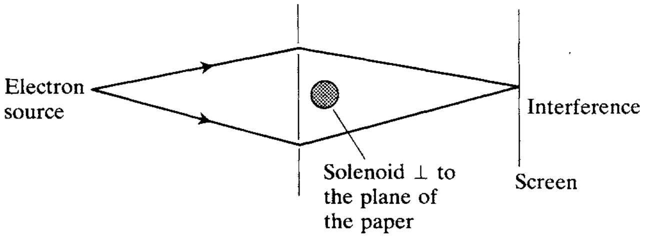


FIG. 1.

Source: H. Brown and H. R. Harré (eds.), *Philosophical Foundations of Quantum Field Theory*, Clarendon Press, Oxford, 1988, Fig. 3.15.

topology. In a different context, one might describe Christopher Columbus as an experimental topologist, demonstrating that the earth was spherical.

Going beyond the single cylindral hole or flux tube of the Bohm–Aharonov experiment, we can consider a complicated knotted configuration of tubes. The study and classification of such knots is a typical and difficult problem in topology. The more elaborate the knot, the more intricate is the structure of the external vacuum. In fact, over the past few years it has emerged that the topology of knots is intimately related to physics, and that the formal machinery of quantum field theory can be used to solve difficult topological problems in the theory of knots.

These developments strongly suggest that topological aspects of 3-dimensional space, as manifested by knots, should play some fundamental role in quantum physics. Moreover, this should involve the basic physical framework, prior to the introduction of matter or force, in other words 'the vacuum'.

3 Topological Quantum Field Theories

The relation between the topology of knots and quantum field theory which has just been alluded to has been very beautifully elaborated in a recent work of Witten (1988). Moreover, this is just the latest of a number of different theories due to Witten, all of which may generically be described as 'topological quantum field theories'. The essential characteristics of such theories are that they have the formal structure of a quantum theory (e.g. dealing with probabilities), but that the information they produce is purely topological (e.g. information about the nature of a knot). There is no measurement or genuine dynamics in such theories; they deal with nature at a more primordial level. At a later stage one may imagine superimposing a more conventional physical theory on to the topological background.

The first of these 'topological quantum field theories' to have been discovered was not the one related to knots in 3-dimensional space, but a more intricate one (Witten 1988) related to the deep mathematical results of Simon Donaldson (Oxford) on 4-dimensional geometry. The four dimensions are those of space–time, but the physical significance of Donaldson's work remains unclear. However, the emergence of 'topological quantum field theories' is likely to have a major impact on current thinking, and may possibly herald a conceptual and technical breakthrough. At the very least, the significance of topological ideas for the study of the quantum vacuum has been highlighted and has already had significant consequences of a purely mathematical character.

References

Aharonov, Y. (1986), 'Non-Local Phenomena and the Aharonov–Bohm Effect', in R. Penrose and C. J. Isham (eds.), *Quantum Concepts in Space and Time*, pp. 41–56, Clarendon Press, Oxford.

Witten, E. (1988), 'Topological Quantum Field Theory', *Comm. Math. Phys.* 117: 353–86.

—— (1989), 'Quantum Field Theory and the Jones Polynomial', *Comm. Math. Phys.* 121: 351–99.

12

How Empty is the Vacuum?

PETER J. BRAAM

1 Introduction

Classically, a portion of space is called a vacuum if there are no particles, in particular no air molecules, in it. This definition allows for the presence of an electromagnetic field in the vacuum.

Quantum-mechanically, this cannot be correct, since the electromagnetic field is built up of photons, small particles carrying the electromagnetic interaction. So a more up-to-date definition would define a vacuum to be 'empty space', without electromagnetic fields or particles present.

If, one day, we have a quantum theory of gravity, it may well be that the Riemannian metric on space, the abstract entity that allows us to measure lengths, is a particle in the quantum-gravitational sense. Then we would have to define the vacuum to be a 'pure space' in which metric properties no longer play a role. Bending and stretching will leave such a vacuum unaffected, and the structure of such spaces is governed by topology, rather than geometry. Conceivably, the modelling of space by so-called manifolds (see below) will have to be given up at some stage, so a yet more final definition may identify the vacuum with a ray in a Hilbert space.

Clearly, the classical notion of vacuum, allowing for space with metric properties and Maxwell fields in it, seems most interesting. We shall come back to this in Section 4. However, it may come as a surprise that the theory of what we called 'pure space' is very rich as well. Our story is about what can be learned about 'pure space' through considering 'fields' on the space.

I would like to thank David Austin for carefully reading the manuscript and suggesting many improvements.

2 Some Mathematical Preliminaries

First of all, I should point out that, to a mathematician, empty space does not necessarily mean good old 3-dimensional Euclidean space. A good class of spaces to work with are *manifolds*, spaces which locally look like everyday vector spaces. Examples of manifolds are the sphere and the doughnut. Both are not equivalent to a Euclidean space, as the doughnut has a loop, which cannot be found in a Euclidean space; similarly, the sphere has a hole of dimension 3, bounded by the 2-dimensional surface. These loops and holes are not going to disappear if we stretch or bend the space, so they are non-trivial in a topological sense.

Manifolds exist in all dimensions. The 2-dimensional ones are easy to understand. If they are oriented they look like one in the sequence drawn in Fig. 1.

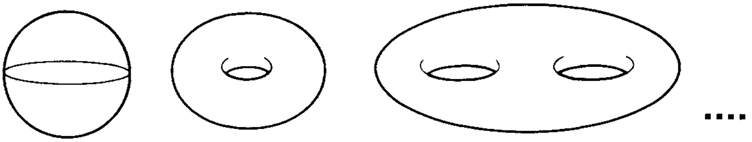


FIG. 1 Two-dimensional manifolds.

In dimension 3 things are much harder. For instance, here is a class of 3-manifolds which is still only partly understood. Remove a knotted circle from 3-dimensional Euclidean space. We are left with a 3-dimensional space in which the circle is absent. It can be shown that two such spaces are the same, up to bending and stretching, if and only if the knots we took out were the same. Some experimenting with a rope will quickly show that many inequivalent knots exist.

Also, the topological structure of 4-dimensional manifolds is very complicated and only partly understood. Unfortunately, it is a little harder to indicate the complications encountered in 4-dimensional topology.

Surprisingly enough, the theory is less complicated if we go over to dimensions 5 or higher. It is a remarkable and poorly understood fact that topology is complicated exactly in the dimensions of space and space–time. In the same spirit, it is not so surprising that precisely in these dimensions methods of physics are currently being exploited very successfully to answer topological questions.

3 The Jones–Witten Theory of Knots

Starting in the early 1980s, V. Jones found many new invariants of knots (Jones 1987). Some outstanding problems in knot theory have been solved using Jones's theory; yet, from a purely mathematical point of view, the theory is not quite satisfactory.

Essentially, Jones first draws the knot in the plane as the closure of a braid (cf. Fig. 2). To compute the invariants he applies a very complicated formula, which depends on the precise form of the braid. The result is an invariant for knots. The key point is as follows. A knot is the closure of many different braids. Jones's formulae depend on choosing an arbitrary braid to represent the knot. Thus the formulae are really formulae for braids, which happen to give the same answer on all braids representing the same knot. Therefore it becomes very difficult to see what properties of the knot itself are really expressed by these formulae. It was an outstanding problem for several years to find an intrinsic, geometric definition of the Jones polynomials, without referring to braids.

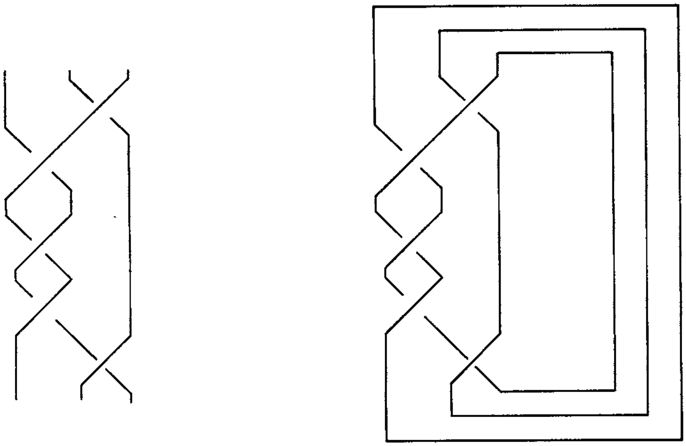


FIG. 2 Braid and closure.

This problem has recently been resolved by Ed Witten. Witten describes a quantum field theory for a 3-manifold with knots in it. The theory has no unnecessary ingredients; only the ambient 3-manifold and an oriented collection of knots is given. Witten then produces a large amount of invariants, all resulting from a path integral over fields. Conformal field theory allows for a careful analysis of the properties of such integrals, and Witten shows that the invariants should coincide with the Jones polynomials.

Because Witten's theory is intrinsic, it lends itself to generalizations. Probably the most profound one is that he can define invariants for knots in arbitrary 3-manifolds, where the concept of braids is not present.

It is fair to say that Witten's theory is one about pure space. His field theory does not rely on metric properties of the 3-manifold, and is therefore a topological quantum field theory. A special feature of such theories is that there are no dynamics: any path in the configuration space satisfies the Euler–Lagrange equation. Also, the theory is manifestly a topological invariant; that is, if two spaces with knots differ only by bending and stretching, the answers will be the same.

It will be a very challenging task to provide a solid mathematical foundation for Witten's theory.

4 Vacuum Solutions and the Donaldson–Witten Theory

To begin our discussion of 4-dimensional space, pick a 4-dimensional manifold X with a metric on it. Instead of looking at Maxwell fields, we look for gauge fields on X which satisfy the Yang–Mills equations. We do not include further particle fields, which is customary when gauge theory is used in particle physics. The Yang–Mills equations are the analogous equations to the Maxwell equations in case one studies gauge fields describing the weak or strong interaction. Such fields on X are called vacuum solutions, much in the spirit of the classical vacuum discussed above.

A vacuum solution has a quantized charge $k = 0, 1, \ldots$ If we fix the charge we find a complicated space of solutions called the Yang–Mills moduli space. This moduli space depends on X and the metric on X so it is not yet a topological invariant of X.

The moduli space has all kinds of loops and holes. Some holes in the moduli space will disappear if we vary the metric on X, others persist. Clearly, the ones that persist must be (differential) topological invariants of X. It was Donaldson's remarkable observation (Donaldson 1983) that such loops did exist and contained deep information about the 4–manifold. So Donaldson's theory is one using classical vacuum to say something about pure space. It is a very powerful theory. It shows, for example, that even on good old 4–dimensional Euclidean space there are infinitely many inequivalent ways of defining differentiability of functions; that is, the differential topology of 4–dimensional vector spaces is very non-trivial.

Also here Witten (1988) has developed a complicated supersymmetric quantum field theory, such that expectation values of certain observables give the relevant information about persisting loops. Atiyah, Donaldson, and Quillen established that the theory has the flavour of an algebraic topological theory in infinite dimensions, and it might be possible to give it a rigorous foundation without having rigorously to construct the much feared Feynmann path integrals on infinite-dimensional spaces.

Witten's quantum field theory in its original version depends on a metric on X just like Donaldson's theory. Witten shows that his expectation values do not change if we vary the metric, thereby showing that they are topological invariants—that is, invariants of pure space. However, his theory is full of all kinds of fields and superfields. So again, a non-vacuous theory gives information about the ultimate vacuum, pure space.

We shall end this paper with a more technical description of this Donaldson–Witten theory. In gauge theory the issue is to find anti-self-dual connections on some bundle over the 4-manifold X. These are precisely the connections for which the self-dual part of the curvature vanishes. Thus we are looking for zeroes of a function s defined on the space A of connections; the function assigns to a connection the self-dual part of its curvature, which is an element of an infinite-dimensional vector space V. Gauge invariance enters the picture in the following way. If we apply a gauge transformation to the connection, then the self-dual part of the curvature is conjugated by this gauge transformation. Thus the function s does not descend to a function on the quotient space A/G of connections modulo gauge transformations, as its values get twisted in V. Instead, it becomes a section of a vector bundle W over A/G. Looking for anti-self-dual connections modulo gauge transformations amounts to looking for the zeroes of a section of this bundle.

We shall now assume that there are only finitely many zeroes of this section. In practice this situation is not always met, but it does happen in interesting situations. Donaldson's invariant is now simply the number of zeroes, that is the number of anti-self-dual connections, counted with signs.

As indicated in Fig. 3, non-triviality of a bundle can force a section to have zeroes. Generally, the number of zeroes of a section of a bundle E over a compact manifold M is an invariant of E, called the Euler number. In particular, it is independent of the section. The Euler number can be computed in various ways, and this is precisely the distinction between Donaldson's theory and the Witten–Donaldson theory.

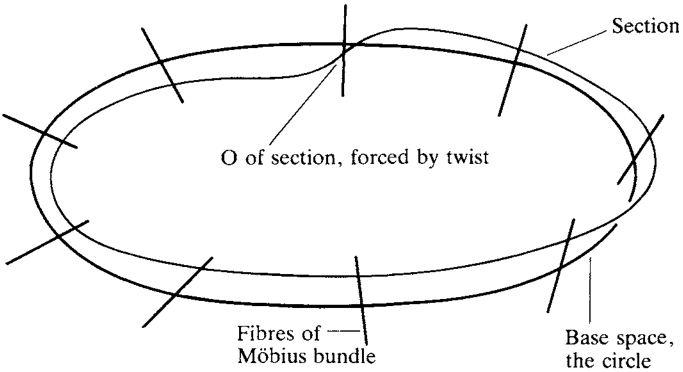


FIG. 3 Möbius bundle with section.

First of all, one can simply compute the number of zeroes of a section; in the case at hand, this amounts to solving the self-dual Yang–Mills equations. A more sophisticated approach is to integrate a differential form over the total space of E. The form is the square of the Thom class of the bundle. In the Yang–Mills case we are integrating the square of the Thom class over the infinite-dimensional bundle W over an infinite-dimensional space A/G, and such integrals are normally called Feynman path integrals. This is more or less what Witten does to compute Donaldson's polynomial invariants, but there is one further observation to be made.

An expression for the Thom class itself can be obtained by using G-equivariant cohomology of the spaces V and A; the equivariant cohomology of a G-space P is the ordinary cohomology of the space $(EG \times P)/G$, where EG is the universal G-bundle over BG, the classifying space of G. De Rham-type models for equivariant cohomology can be found and lead to a more computational version of equivariant cohomology; in particular, Mathai and Quillen (1986) derived completely explicit expressions for the Thom classes using such models. As established by Atiyah, Witten's Lagrangians are precisely Mathai *et al.*'s formulae for the gauge theory case.

The original formula used by Witten is based on the so-called Cartan model of equivariant cohomology. It depends on the existence of a connection on the bundle E. In the case of gauge theory, a connection on the bundle W over A/G can be found using a metric on the original 4-manifold X, thereby breaking the natural symmetry of the problem. Soon after Witten's paper appeared, various physicists found out that the general covariance of the Lagrangian can be made obvious if one uses the formula for the Thom class arising from the Weyl algebra model of equivariant cohomology. The expression for the Thom class of E no longer requires a connection on E. This formula had also been found by Mathai *et al.*

References

Donaldson, S. K. (1983), 'An Application of Gauge Theory to Four Dimensional Topology', *J. Diff. Geometry* 18: 279–315.

Jones, V. F. R. (1987), 'Hecke Algebra Representations of Braid Groups and Link Polynomials', *Ann. Math.* 126: 335–88.

Mathai, V. and Quillen, D. G. (1986), 'Superconnections, Thom Classes, and Equivariant Differential Forms', *Topology* 25: 85–110.

Witten, E. (1988), 'Topological Quantum Field Theory', *Comm. Math. Phys.* 117: 353–86.

Index

9 780198 244493