The Physical Significance of the Factor 2

Peter Rowlands

IQ Group and Science Communication Unit, Department of Physics, University of Liverpool, Oliver Lodge Laboratory, Oxford Street, P.O. Box 147, Liverpool, L69 7ZE, UK. e-mail: p.rowlands@liverpool.ac.uk; prowl@hep.ph.liv.ac.uk; and prowl@csc.liv.ac.uk

Abstract. A numerical factor 2 or ½ occurs in many physics equations, classical, relativistic and quantum, and also in some aspects of mathematics. Analysis of this factor suggests that, in all significant examples, it has a common origin, a fact which has profound implications for the philosophical foundations of both physics and mathematics, and the relationship between them. It may be tracked down, ultimately, to the principle of duality, in which both physics and mathematics structure themselves by trying to avoid creating 'something' from nothing.

GEOMETRY AND KINEMATICS

We probably first come across the factor 2 in the formula for the area of a triangle, ½ × length of base × perpendicular height. In the right-angled triangle, it is created by bisecting a rectangle along a diagonal. If this is now taken as representing a straight-line graph, of, say, velocity (v) against time (t), under uniform acceleration, the area under the graph becomes the distance travelled, ½ vt. By comparison, for an object travelling at steady speed v throughout the same time interval t, the distance travelled is the area of the rectangle under the horizontal straight line representing steady v, that is, vt. The factor 2, in this case, distinguishes between steady conditions and steadily *changing* conditions.

Starting from an initial velocity u, and supposing the same uniform acceleration, we obtain the 'mean speed theorem', in which the total distance travelled under uniform acceleration equals the product of the mean speed and the time: $s = ½ (u + v) t$. If we additionally define uniform acceleration as $a = (v – u) / t$, we obtain the well-known equation for uniformly accelerated motion: $v^2 = u^2 + 2as$, which becomes $v^2 = 2as$, when $u = 0$. If we now apply this to a body of mass m, acted on by a uniform force $F = ma$, we find the work done over distance s is equal to the kinetic energy gained

$$Fs = mas = \frac{mv^2}{2} - \frac{mu^2}{2} \, ,$$

which reduces to ½ mv^2 if we start at zero speed. Using $p = mv$ to represent momentum, it is convenient also to express this formula in the form $p^2 / 2m$. It is easy, of course, to show that this formula applies additionally to the case of nonuniformly accelerated motion, using a simple integration of force (dp / dt) over displacement:

$$\int \frac{dp}{dt} \, \mathrm{ds} = \int mv \, dv \; = \frac{mv^2}{2} \; .$$

In principle, however, we see that a steady increase of velocity from 0 to v requires an averaging out which halves the values of significant dynamical quantities obtained under steady-state conditions.

KINETIC AND POTENTIAL ENERGY

The same factor makes its appearance, in precisely the same way, in molecular thermodynamics, quantum theory and relativity. Its significance here is that it relates the continuous aspect of physics to the discrete, and, since these aspects are required in the description of any physical system, the factor acquires a universal relevance. An obvious classical manifestation is the fact that *two* types of conservation of energy equation are commonly used in physics. Potential energy equations represent steady-state conditions, when there is no overall change in the energy distribution; kinetic energy equations apply when there is a redistribution of energy within a system though the energy remains conserved overall. Typically, we apply the potential energy equation to the case of a planet in a regular gravitational orbit. So, the force equation

$$\frac{mv^2}{r} = -\frac{GMm}{r^2}$$

leads to an equivalent potential energy relation

$$mv^2 = -\frac{GMm}{r} .$$

However, the changing conditions involved in the escape of a body of mass m from a gravitational field require a kinetic energy equation of the form

$$\frac{mv^2}{2} = \frac{GMm}{r} .$$

Numerically, in such cases, the potential energy term is twice the value of the kinetic, and we recognise that this is a special case of the virial theorem, according to which, in a conservative system governed by force terms inversely proportional to power n of the distance, or potential energy terms inversely proportional to power n – 1, the time-averaged kinetic and potential energies, $\overline{T}$ and $\overline{V}$, are related by the formula:

$$\overline{T} = \frac{(1-n)}{2}\overline{V} .$$

For the two special cases, of constant force and inverse-square-law force, $\overline{V}$ is numerically equal to $2\overline{T}$. Such forces, in fact, are overwhelmingly predominant in nature, because they are a natural consequence of three-dimensional space, and this may well be related to the geometric origin of the factor 2 in such formulae as that for the area of a triangle.

KINETIC THEORY OF GASES

Though the potential and kinetic energy equations may, at first sight, appear to be contradictory expressions of the general principle of the conservation of energy, they can be easily reconciled if we consider the kinetic energy relation to be concerned with the action side of Newton's third law, while the potential energy relation concerns both action and reaction. We can give many physical illustrations. To give a characteristic example, an old proof of Newton's of the mv^2 / r law for centripetal force, and hence of the formula mv^2 for orbital potential energy, has the satellite object being 'reflected' off the circle of the orbit, in a polygon with an increasing number of sides, which, in the limiting case, becomes a circle. The imagined physical reflection, by doubling the momentum through action and reaction, then produces the potential energy, rather than kinetic, energy formula.

Precisely the same principle applies in the derivation of Boyle's law, from what we often call the 'kinetic theory of gases'. Here, a real reflection of the ideal gas molecules off the walls of the container produces the momentum doubling, which indicates steady-state conditions, though it is immediately removed by the fact that we have to calculate the average time between collisions ($t = 2a / v$) as the time taken to travel *twice* the length of the container (a). The average force then becomes the momentum change / time = $2\ mv / t = mv^2 / a$, and the pressure due to one molecule in a cubical container of side a becomes mv^2 / a^3, or mv^2 / V (volume), leading for n molecules to the direct pressure-density relationship, which we call Boyle's law. The kinetic behaviour of the ideal gas molecules is actually irrelevant to the derivation, since the system describes a steady-state dynamics with positions of molecules constant on a time-average. Taking into account the three dimensions between which the velocity is distributed, the ratio of pressure and density (P / ρ) is derived from the *potential energy* term mv^2 for each molecule and is equal to one third of the average of the squared velocity, or $\overline{c}^2 / 3$. So, the relationship could have been derived (as was done by Newton) using a mathematical model in which the molecule positions remained fixed.

The kinetic behaviour only becomes significant when we introduce temperature as a measure of the average kinetic energy of the molecules of the gas. There is, however, no 'derivation' involved here, because temperature is not defined independently of this kinetic energy, and we are obliged to provide this definition by an *explicit* use of the virial theorem, to find the otherwise *unknown* average kinetic energy from the *known* potential energy. Assuming that the potential energy of each ideal gas molecule is kT for each degree of freedom, and, in total, $3kT$, and taking the pressure law as equivalent to a dynamical system involving a constant force, we apply the virial theorem to obtain the kinetic energy expression ($3kT / 2$) for each of these molecules.

PHOTONS AND RADIATION PRESSURE

Photons, unlike material particles, are relativistic objects, so we might expect that the expressions for photon gases would be different in some respect from those for material gases. In fact, they are almost identical, as the radiation pressure of a photon gas within a fixed enclosure is one third of the energy density of radiation, that is:

$$P = \frac{1}{3} \rho c^2 \ .$$

The photon, as a 'gas' component, thus behaves in exactly the same way as a material particle, and, because the system is in steady state, the energy term mc^2 behaves as *potential*, not kinetic, energy, exactly as its form would suggest, with no mysterious 'relativistic factor' at work, as proposed by some authors. The photons are reflected off the walls of the container in the same way as the material gas molecules, although this time we can also consider the process as involving absorption and re-emission.

The whole reason for Einstein's introduction of $E = mc^2$ to represent the total energy of both photons and material particles was to preserve the *classical* laws of conservation of mass and conservation of energy. The total energy equation, unlike the *change of energy* formula $\Delta E = \Delta mc^2$, cannot be derived, by deductive means, from the postulates of relativity; it depends entirely on the choice of an integration constant in the relativistic expression for rate of energy change:

$$\frac{dT}{dt} = \mathbf{F.v} \ .$$

No problem arises if we recognise that mc^2 has a classical, *as well as relativistic*, meaning. Like many other significant results (the Schwarzschild radius, the equations for the expanding universe, the gravitational redshift, the spin of the electron), the expression does not arise from the theory of relativity itself but is a more fundamental truth which that theory has uncovered.

It would be extraordinary, in fact, if relativistic conditions should somehow conspire exactly to halve or double significant classical quantities. Relativistic factors are typically of the form $\gamma = (1 - v^2 / c^2)^{-1/2}$, implying some gradual change when $v \to c$, and it makes no physical sense to suppose that the transition involves discrete integers. $\Delta E = \Delta mc^2$ is a relativistic equation because it incorporates the γ factor in the Δm term, but $E = mc^2$ is not, and, for photons at least, the effects which depend only on $E = mc^2$ and not specifically on the 4-vector combination of space and time can be derived by classical approaches entirely independent of any concept of relativity. Examples of this can be found in calculations based on the classical corpuscular theory of light dating back to the seventeenth, eighteenth and nineteenth centuries, and are still used for practical purposes at the present day. We may mention, for example, Newton's calculation of atmospheric refraction in 1694, and his application of the formula for the velocity of waves in a medium to an optical aether in Query 21 of the *Opticks*. Essentially, Newton's formula,

$$c = (E \,/\, \rho)^{1/2}$$

where E is elasticity or pressure and ρ density, is an expression of the fact that the potential energy of the system of light corpuscles, or the aether that acts upon them, is equal to the work done at constant pressure as a product of pressure and volume. Newton's elasticity of the aether is essentially the same as the modern energy density of radiation (ρc^2), which is related by Maxwell's classical formula of 1873 to the radiation pressure. Light necessarily gives a 'correct' result for such a calculation when travelling through a vacuum, because there is no source of dissipation, and the virial relation takes on its ideal form.

PHOTONS IN A GRAVITATIONAL FIELD

There are many further relations between photon and material particle dynamics, which will be significant to us. Although light in free space has velocity c, and, therefore, no rest mass or kinetic energy, *as soon as you apply a gravitational field*, the light 'slows down', and, at least *behaves* as though it can be treated as a particle with kinetic energy *in the field*. The same, of course, applies to photons in a plasma, a system which has often been used as an analogy to the Higgs mechanism for acquiring mass in particle physics. An example directly applicable to photons is the use of the standard Newtonian escape velocity (or kinetic energy) equation

$$\frac{mv^2}{2} = \frac{GMm}{r}$$

to derive the Schwarzschild limit for a black hole, by purely classical means, as was done more than once in the eighteenth century. Assuming $v \to c$, we derive

$$r = \frac{2GM}{c^2}$$

with *no transition to a 'relativistic' value*.

A classic case of applying such a kinetic energy-type equation to light, is the derivation of the double gravitational bending, an effect often thought to be derivable only from the general relativistic field equations. We have been assured repeatedly, since Eddington's measurement of 1919, that the double bending is a relativistic effect, and that 'Newtonian' calculations, using the principle of equivalence, yield only half the correct value, although several authors have shown that the effect can be derived also from special relativity.

The 'Newtonian' calculation, we are told, originating with Soldner in 1801, starts from the potential energy equation (modified for a hyperbolic orbit), according to the expression:

$$mc^2 = \frac{GMm\,(e-1)}{r} \ ,$$

with e taken as the eccentricity of the hyperbolic orbit. Since 1 « e, the half-angle deflection becomes

$$\frac{1}{e} = \frac{GM}{c^2 r} \ ,$$

and the full angle deflection (in and out of the gravitational field)

$$\frac{2}{e} = \frac{2GM}{c^2 r} \ .$$

This is only half the general relativistic value. However, Soldner actually used the kinetic energy equation,

$$\frac{mc^2}{2} = \frac{GMm\,(e-1)}{r} \ ,$$

on the basis of Laplace's prior use of it for calculating the black hole radius, and he would have obtained the 'correct' total deflection if he had used the double angle in calculating his integral! Soldner's procedure was surely the correct one, for the case he was examining was that of an orbit in the process of formation (the reverse, in effect, of Laplace's escape velocity), and not of an orbit in steady state.

That a purely classical calculation of the light-bending is possible should not surprise us. Energy, in relativity, is, after all, *defined* to be consistent with its classical value in the case of a particle with no material component; and so relativity theory should not produce different energy equations to classical physics for light photons; it merely corrects our naïve understanding of what are steady-state and what are changing conditions. Of course, in the case of photons, we never see a material kinetic energy directly; the total energy balance means that it must be possible to treat it as though it does exist when the particle is 'slowed down' by a field.

There are many cases where a 'relativistic' correction (either special or general) is presumed to 'cause' the doubling of a physical effect, but such examples, are not illustrations of the fact that the calculation of the doubling has to be done relativistically, but that relativity provides one way of incorporating the effect of changing conditions if we begin with the potential, rather than the kinetic, energy equation. In the case of gravitational bending, the potential energy equation typically produces the effect of gravitational redshift, or time dilation, while relativity adds the corresponding length contraction. So authors have variously argued for the redshift being 'Newtonian' while the length-contraction or 'space-warping' is relativistic, or for the length contraction being Newtonian while the redshift is relativistic. Claims have also been made that the 'Newtonian' effect has to be added to that produced by the Einstein calculation of 1911, based on the equivalence principle (which also obtained only half of the correct value), or that the two effects are the same, and have to be supplemented by a 'true' relativistic effect, like the Thomas precession. All of these arguments are correct, but none is fundamental. The true reason is the choice of classical energy equation. If the potential energy equation is used where the kinetic energy equation is appropriate, then correct physical reasons can be found for almost *any* additional term which doubles the effect predicted. Even special relativity is only an alternative approach to a calculation that must also be valid classically, and the same applies to the even more famous case of the planetary perihelion precession.

THE GYROMAGNETIC RATIO OF THE ELECTRON

It has also been assumed that relativity is needed to explain the anomalous magnetic moment or, equivalently, the gyromagnetic ratio of a Bohr electron acquiring energy in a magnetic field. 'Classical' reasoning, we assured, would show the energy acquired by an electron changing its angular frequency from ω_0 to ω in a magnetic field **B** to be of the form

$$m\,(\omega^2 - \omega_0^2) = e\omega_0 rB \ \ ,$$

leading, after factorization of $(\omega^2 - \omega_0^2)$, to an angular frequency change

$$\Delta\omega = \frac{eB}{2mr} \ .$$

However, a relativistic effect (the Thomas precession, again) ensures that the classical $e\omega_0 rB$ is replaced by $2e\omega_0 rB$, leading to

$$\Delta\omega = \frac{eB}{mr} \ .$$

But, once again, relativistic and classical treatments coincide when, as with the light-bending example, the *kinetic* energy equation is recognised as the one applied to changing conditions, *at the instant we 'switch on' the field.* Then, we automatically write

$$\frac{1}{2}\,m(\omega^2 - \omega_0^2) = e\omega_0 rB \ \ ,$$

which is no more, in principle, than the equation of motion for uniform acceleration

$$v^2 - u^2 = 2as \ \ .$$

So, the Thomas precession or 'relativistic' correction is needed if we begin with the potential energy equation applicable to a steady state, but not if we apply the kinetic energy used for changing conditions.

ELECTRON SPIN

The gyromagnetic ratio is, of course, produced ultimately by the electron spin, which is one of the most famous cases of the factor ½. Traditionally, this is derived from the relativistic Dirac equation by consideration of the commutator

$$[\hat{\sigma}, \mathcal{H}] = [\hat{\sigma}, i\gamma_0\boldsymbol{\gamma}.\mathbf{p} + \gamma_0 m] \ .$$

Purely formal reasoning shows that this reduces to $2\gamma_0\, \boldsymbol{\gamma} \times \mathbf{p}$, or in our multivariate vector terminology (equivalent to Pauli matrices), to $2\boldsymbol{ij}\ \mathbf{1} \times \mathbf{p}$. The significant thing here is that the factor 2 emerges from the anticommuting properties of the vector operators in an equation such as

$$[\hat{\sigma}, \mathcal{H}] = 2\boldsymbol{j}\ (\mathbf{ij}p_2 + \mathbf{ik}p_3 + \mathbf{ji}p_1 + \mathbf{jk}p_3 + \mathbf{ki}p_1 + \mathbf{kj}p_2) \ .$$

Ultimately, this leads to

$$[\mathbf{L} + \hat{\sigma} / 2, \mathcal{H}] = 0 \ .$$

where **L** is the orbital angular momentum, from which we find that ($\mathbf{L} + \hat{\sigma} / 2$) is a constant of the motion.

There is, however, a way of deriving the same result (at least in its manifestation in the presence of a magnetic field) from the Schrödinger equation, which can easily be shown to be a nonrelativistic limit to the bispinor form of the Dirac equation. In principle, this should mean that the spin ½ term that arises from the Dirac equation has nothing to do with the fact that the equation is relativistic, but is a result of the fundamentally multivariate nature of its use of the momentum operator, as the formal derivation from the Dirac equation would suggest.

It is significant that the standard derivation of the Schrödinger equation begins with the classical expression for kinetic energy, $p^2 / 2m = mv^2 / 2$.

$$T = (E - V) = \frac{p^2}{2m} \ ,$$

followed by substitution of the quantum operators $E = i\ \partial / \partial t$ and $\mathbf{p} = -\,i\ \nabla$, acting on the wavefunction ψ, for the corresponding classical terms, to give:

$$(E - V)\ \psi = -\frac{1}{2m}\ \nabla^2\psi$$

or

$$i\,\frac{\partial\psi}{\partial t} - V\psi = -\frac{1}{2m}\ \nabla^2\psi \ ,$$

in the time-varying case. Various authors [e.g. Gough, 1990] have shown that, using a multivariate operator, $\mathbf{p} = -i\nabla + e\mathbf{A}$, in the absence of scalar potential V, we derive:

$$2mE\psi = (-i\boldsymbol{\nabla} + e\mathbf{A})\ (-i\boldsymbol{\nabla} + e\mathbf{A})\ \psi$$

leading ultimately, by a relatively easy derivation, to

$$2mE\psi = (-i\boldsymbol{\nabla} + e\mathbf{A}).(-i\boldsymbol{\nabla} + e\mathbf{A})\ \psi + 2m\ \boldsymbol{\mu}.\mathbf{B}\ ,$$

which is the conventional form of the Schrödinger equation in a magnetic field for spin up, supplied by the usual *ad hoc* addition of Pauli matrices, and a similar expression may be derived from the spin down state. The 2 in the expression $2m\ \boldsymbol{\mu}.\mathbf{B}$ arises from the anticommuting properties of the multivariate vectors. In effect, spin is purely a property of the multivariate nature of the **p** term, and has nothing to do with whether the equation used is relativistic or not. (This is equivalent to stating the well-known fact that the 4π rotation involved in spin is purely a property of the rotation group.) We also see that the factor 2 is both introduced with the transition in the Schrödinger equation from the classical kinetic energy term, and, at the same time, produced by the anticommuting nature of the momentum operator. It is precisely because the Schrödinger equation is derived via a kinetic energy term that this factor enters into the expression for the spin, and this process is essentially the same as the process which, through the anticommuting quantities of the Dirac equation, makes ($\mathbf{L} + \hat{\boldsymbol{\sigma}} / 2$) a constant of the motion.

THE HARMONIC OSCILLATOR AND HEISENBERG UNCERTAINTY

The Schrödinger equation also allows an easy calculation of the eigenvalues of the quantum harmonic oscillator, in which a *varying* potential energy term, ½ $m\omega^2x^2$, taken from the classical kinetic energy term ½ mv^2, is added to the Hamiltonian. A formal derivation is hardly necessary to show that the ½ in the expression for the ground state or 'zero-point' energy,

$$E_0 = \frac{\hbar\omega}{2}\ ,$$

carries over directly from this original introduction.

Anticommuting operators also introduce the factor 2 in the Heisenberg uncertainty relation for the same reason as they do in the treatment of electron spin, and the Heisenberg term also relates directly to the zero-point energy derived from the kinetic energy of the harmonic oscillator. The formal derivation of the uncertainty principle assumes a state represented by a state vector ψ which is an eigenvector of the operator P. If Q is an operator which anticommutes with P, we derive

$$(\Delta p)\ (\Delta q) \geq (1/2)\ [P,Q] \geq \hbar\ / 2$$

where the factor 2 in comes from the noncommutation of the *p* operator.

BOSONS AND FERMIONS

The origin of the factor 2 in the spin states of fermions and bosons, is, once again, the virial relation between kinetic and potential energies. In our reformulation of the Dirac equation for fermions, we take the classical relativistic energy-momentum conservation equation:

$$E^2 - p^2c^2 - m_o{}^2c^4 = 0 \ ,$$

where m_o is rest mass, and factorize using our quaternion-multivariate-4-vector operators to give:

$$(\pm \boldsymbol{k}E \pm i\boldsymbol{i}\,\mathbf{p} + i\boldsymbol{j}\,m_o)\,(\pm \boldsymbol{k}E \pm i\boldsymbol{i}\,\mathbf{p} + i\boldsymbol{j}\,m_o) = 0 \ .$$

We then apply the Correspondence Principle to obtain

$$\left(\pm i\boldsymbol{k}\frac{\partial}{\partial t} \pm \boldsymbol{i}\,\nabla + i\boldsymbol{j}m_o\right)\psi \ = \ 0 \ ,$$

where

$$\psi = (\pm \boldsymbol{k}E \pm i\boldsymbol{i}\,\mathbf{p} + i\boldsymbol{j}\,m_o)\,e^{-i(Et\,-\,\mathbf{p.r})} \ .$$

for a fermion. We can proceed to show that a spin 1 boson wavefunction (incorporating fermion-antifermion combination) is the sum of

$$\begin{array}{l}
(\boldsymbol{k}E + i\boldsymbol{i}\,\mathbf{p} + i\boldsymbol{j}\,m_o)\,(-\boldsymbol{k}E + i\boldsymbol{i}\,\mathbf{p} + i\boldsymbol{j}\,m_o)\\
(\boldsymbol{k}E - i\boldsymbol{i}\,\mathbf{p} + i\boldsymbol{j}\,m_o)\,(-\boldsymbol{k}E - i\boldsymbol{i}\,\mathbf{p} + i\boldsymbol{j}\,m_o)\\
(-\boldsymbol{k}E + i\boldsymbol{i}\,\mathbf{p} + i\boldsymbol{j}\,m_o)\,(\boldsymbol{k}E + i\boldsymbol{i}\,\mathbf{p} + i\boldsymbol{j}\,m_o)\\
(-\boldsymbol{k}E - i\boldsymbol{i}\,\mathbf{p} + i\boldsymbol{j}\,m_o)\,(\boldsymbol{k}E - i\boldsymbol{i}\,\mathbf{p} + i\boldsymbol{j}\,m_o)
\end{array}$$

while a spin 0 boson reverses the signs of **p** in the second column. The fermion wavefunction is effectively a nilpotent or square root of 0, and the boson wavefunction a product of two nilpotents (each not nilpotent to the other).

From both Dirac and Schrödinger equations, effectively describing kinetic energy states, we see that fermions have half-integral spins. The Klein-Gordon equation, which applies to bosons, however, is the potential energy equation, based on $E = mc^2$, where *m* is now the ‘relativistic’, rather than the rest mass, and bosons derive their integral spin values from the fact that the energy term in this equation contains unit values of the mass *m*. Here, we quantize the classical relativistic energy-momentum equation directly, to obtain

$$\frac{\partial^2\psi}{\partial t^2} - \nabla^2\psi = m^2\psi \ ,$$

in units where $\hbar = c = 1$.

The difference between the fermion and boson cases is that we use the kinetic energy relation when we consider a particle as an object in itself, described by a rest mass m_0, undergoing a continuous change; and the potential energy relation when we consider a particle within its 'environment', with 'relativistic mass', in an equilibrium state requiring a discrete transition for any change. Kinetic energy is associated with rest mass, because it cannot be defined without it – light 'slowing down' in a gravitational field or condensed matter is effectively equivalent to adopting a rest mass. Potential energy is associated with 'relativistic' mass because the latter is *defined* through a potential energy-type term ($E = mc^2$), light in free space being the extreme case, with no kinetic energy or rest mass, and 100 per cent potential energy or relativistic mass.

The description is also related to the halving process that occurs, for a material particle, when we expand its relativistic mass-energy term (mc^2) to find its kinetic energy (½ mv^2). So we can take the relativistic energy conservation equation

$$E - mc^2 = E^2 - p^2c^2 - m_o{}^2c^4 = 0 \ \ .$$

as a 'relativistic' mass or potential energy equation, treating at one go the particle interacting with its environment, and proceed to quantize to a Klein-Gordon equation, with integral spin. Or, we can separate out the kinetic energy term using the rest mass m_0, and take the square root of

$$E^2 = m_o{}^2c^4 \left(1 - \frac{v^2}{c^2}\right)^{-1} ,$$

to obtain

$$E = m_oc^2 + \frac{m_ov^2}{2} + \dots \ .$$

from which, as we have seen, we derive the Schrödinger equation, and spin ½.. The ½ is, indeed, a statement of the act of square-rooting, which is precisely what happens when we split 0 into two nilpotents in the Dirac equation; the ½ in the Schrödinger approximation is a manifestation of this which we can trace through the ½ in the relativistic binomial approximation. The origin of the same factor in the derivation of spin from the Dirac equation, is seen in the behaviour of the anticommuting terms which result from the process of taking the nilpotent: the anticommuting and binomial factors have precisely the same origin.

ZERO POINT ENERGY AND RADIATION REACTION

The significance of the factor 2 in all our examples lies in the fact that it relates together two parallel but almost independent streams of physics: the continuous and the discontinuous. Expressions involving half units of $\hbar$ represent an average or integrated increase from 0 to $\hbar$. The half-values are characteristic of the continuous option in physics, the integral ones of the discontinuous option. Schrödinger and Heisenberg are examples of these options; and yet another completely continuous theory, stochastic electrodynamics, based on the existence of zero-point energy of value $\hbar\omega / 2$, has developed as a rival to the purely discrete theory of the quantum with energy $\hbar\omega$.

The discrete and continuous options are not only possible, but actually *required* within a system. Each type of system has to incorporate the alternative option in some way. Schrödinger, for instance, has a continuous system based on ½ $\hbar$, but incorporates discreteness (based on $\hbar$) in the process of measurement – the so-called collapse of the wavefunction. Heisenberg, on the other hand, has a discrete system, based on $\hbar$, but incorporates continuity (and ½ $\hbar$) in the process of measurement – via the uncertainty principle and zero-point energy. Nature, it would seem, always manages to provide a route by which ½ $\hbar\omega$ in one context becomes $\hbar\omega$ in another. A characteristic example is black-body radiation, where the spontaneous emission of energy of value $\hbar\omega$ is produced by the combined effect of the ½ $\hbar\omega$ units of energy provided by both oscillators and zero-point field.

The ½ $h\nu$ or ½ $\hbar\omega$ for black body radiation appears in both the theories of Planck, of 1911, and of Einstein and Stern, of 1913. In quantum mechanics, as we have seen, the zero-point energy term is derived from the harmonic oscillator solution of the Schrödinger equation, while, in the Heisenberg formulation, it appears as a result of the ½ $\hbar$ term involved in the uncertainty principle. While the derivation via Schrödinger shows the kinetic origins of the factor 2, the derivation from the uncertainty principle suggests an origin in continuum physics.

The ½ $\hbar\omega \rightarrow \hbar\omega$ transition for black body radiation can also be seen in terms of radiation reaction. Rather surprisingly, perhaps, this is again connected with the distinction between the relativistic and rest masses of an object. When we define a rest mass we effectively define an isolated object, and we cannot define *kinetic* energy in terms of anything but this rest mass. If, however, we take a *relativistic* mass, we are already incorporating the effects of the environment. In the case of a photon, which has no rest mass, and only a relativistic mass, the energy mc^2 behaves exactly like a classical potential energy term, for example as a component of a photon gas producing the radiation pressure $\rho c^2 / 3$. Action and reaction occurs in this instance because the doubling of the value of the energy term comes from the doubling of the momentum produced by the rebound of the photons from the walls of the container, or absorption and re-emission. The same thing happens with radiation reaction, which produces an otherwise ‘mysterious’ doubling of energy from ½ $h\nu$ to $h\nu$. In a different context, Feynman and Wheeler produce a doubling of the contribution of the retarded wave in electromagnetic theory, at the expense of the advanced wave, by assuming that the vacuum behaves as a perfect absorber and reradiator of radiation.

It seems that incorporating radiation reaction means that we are also incorporating the effect of Newton's third law, as in the case of many other processes. However, many of the same results, as in the parallel case of the anomalous magnetic moment of the electron, are also explained by special relativity. C. K. Whitney [2000] has argued that the correct result for the electron is obtained, without relativity, by treating the transmission of light as a two-step process involving absorption and emission, which, in our terms, is equivalent to incorporating action and reaction, or the potential energy equation, and, as we have seen, the same result follows classically by defining the potential energy at the moment the field is switched on. If, however, we use kinetic energy, or a one-step process, we also need relativity, because, once we introduce rest mass, we can no longer use classical equations. ('Relativistic mass' is, of course, specifically *designed* to preserve classical energy conservation!) The two-step process is analogous to the use of radiation reaction, so it follows, in principle, that a radiation reaction is equivalent to adding a relativistic 'correction' (such as the Thomas precession).

Whitney argues further that the two-step process removes those special relativistic paradoxes which involve apparent reciprocity, which is interesting, because special relativity, by including only one side of the calculation, effectively removes reciprocity, and so leads to such things as asymmetric ageing in the twin paradox. Similar arguments also apply to the idea that the problem lies in attempting to define a one-way speed of light that cannot be measured, because a two-way speed measurement of the speed of light also requires a two-step process.

OBJECT PLUS ENVIRONMENT

We have already proposed that the factor 2 originates in the symmetry between the action of an object and the reaction of its environment. While a fermionic object on its own shows changing behaviour, requiring an integration which generates a factor ½ in the kinetic energy term, and a sign change when it rotates through 2π, a conservative 'system' of object plus environment shows unchanging behaviour, requiring a potential energy term, which is twice the kinetic energy.

Taking 'environment' to apply to either material or vacuum, we can makes sense, not only of the boson / fermion distinction and the spin 1 / ½ division in a fundamental way, but also understand such concepts as supersymmetry, vacuum polarization, pair production, renormalization, *zitterbewegung*, and so on, because the halving of energy in 'isolating' the fermion from its vacuum or material 'environment' is the same process as mathematically square-rooting the quantum operator via the Dirac equation. Integral spins may be automatically produced from half-integral spin electrons using the Berry phase, and, by generalizing this kind of result to all possible environments, we may extend the principle in the direction of supersymmetry. In principle, we propose that energy principles determine that *all* fermions, in whatever circumstances, may be regarded either as isolated spin ½ objects or as spin 1 objects in conjunction with some particular material or vacuum environment, or, indeed, the ‘rest of the universe’.

In this context, fermions with spin ½ become spin 1 particles when taken in conjunction with their environment, whatever that may be. The Jahn-Teller effect and Aharonov-Bohm effect are examples. Treated semi-classically, the Jahn-Teller effect couples the factors associated with the motions of the electronic and nuclear coordinates so that different parts of the total wave function change sign in a coordinated manner to preserve the single-valuedness of the total wave function.

In more general terms, we can consider a similar relationship existing between a fermion and ‘the rest of the universe’, the *total* wavefunction representing fermion plus ‘rest of the universe’ being necessarily single-valued, and automatically introducing the extra term known as the Berry phase. This duality occurs with the actual creation of the fermion state. Splitting away a fermion from a ‘system’ (or ‘the universe’), we have to introduce a coupling as a mathematical description of the splitting. The reverse effect must also exist, with bosons of spin 0 or 1 coupling to an ‘environment’ to produce fermion-like states. Perhaps the Higgs mechanism occurs in this way, but a more immediate possibility is the coupling of gluons to the quark-gluon plasma to deliver the total spin of ½ or 3/2 to a baryon.

Fermions and bosons, it would seem, always produce a ‘reaction’ within their environment, which couples them to the appropriate wavefunction-changing term, so that the potential / kinetic energy relation can be maintained at the same time as its opposite. The whole process of renormalization depends on an infinite chain of such couplings through the vacuum. The coupling of the vacuum to fermions generates ‘boson-images’ and vice versa.

RENORMALIZATION AND SUPERSYMMETRY

To understand the principle of renormalization, we need to use the nilpotent version of the Dirac wavefunction, which is, typically, $(\boldsymbol{k}E + i\boldsymbol{i}\mathbf{p} + i\boldsymbol{j}m)$ for a fermion and $(-\boldsymbol{k}E + i\boldsymbol{i}\mathbf{p} + i\boldsymbol{j}m)$ for an antifermion, these being abbreviated representations of 4-term bra and ket vectors, cycling through the full range of $\pm E$ and $\pm\mathbf{p}$ values. In terms of the ‘environment’ principle, a fermion generates an infinite series of interacting terms of the form:

$(\boldsymbol{k}E + i\boldsymbol{i}\mathbf{p} + i\boldsymbol{j}m)$
$(\boldsymbol{k}E + i\boldsymbol{i}\mathbf{p} + i\boldsymbol{j}m)\ (-\boldsymbol{k}E + i\boldsymbol{i}\mathbf{p} + i\boldsymbol{j}m)$
$(\boldsymbol{k}E + i\boldsymbol{k}\mathbf{p} + i\boldsymbol{j}m)\ (-\boldsymbol{k}E + i\boldsymbol{i}\mathbf{p} + i\boldsymbol{j}m)(\boldsymbol{k}E + i\boldsymbol{i}\mathbf{p} + i\boldsymbol{j}m)$
$(\boldsymbol{k}E + i\boldsymbol{i}\mathbf{p} + i\boldsymbol{j}m)\ (-\boldsymbol{k}E + i\boldsymbol{i}\mathbf{p} + i\boldsymbol{j}m)(\boldsymbol{k}E + i\boldsymbol{i}\mathbf{p} + i\boldsymbol{j}m)\ (-\boldsymbol{k}E + i\boldsymbol{i}\mathbf{p} + i\boldsymbol{j}m)$, etc.

The ($\boldsymbol{k}E + \boldsymbol{i}i\mathbf{p} + \boldsymbol{i}\boldsymbol{j}m$) and ($-\boldsymbol{k}E + \boldsymbol{i}i\mathbf{p} + \boldsymbol{i}\boldsymbol{j}m$) vectors are an expression of the behaviour of the vacuum state, which acts like a 'mirror image' to the fermion. An expression such as

$$(\boldsymbol{k}E + i\boldsymbol{i}\mathbf{p} + i\boldsymbol{j}m)\ \boldsymbol{k}\ (\boldsymbol{k}E + i\boldsymbol{i}\mathbf{p} + i\boldsymbol{j}m)$$

is part of an infinite regression of images of the form

$$(\boldsymbol{k}E + i\boldsymbol{i}\mathbf{p} + i\boldsymbol{j}m)\ \boldsymbol{k}\ (\boldsymbol{k}E + i\boldsymbol{i}\mathbf{p} + i\boldsymbol{j}m)\ \boldsymbol{k}\ (\boldsymbol{k}E + i\boldsymbol{i}\mathbf{p} + i\boldsymbol{j}m)\ \boldsymbol{k}\ (\boldsymbol{k}E + i\boldsymbol{i}\mathbf{p} + i\boldsymbol{j}m)\ ...$$

where the vacuum state depends on the operator that acts upon it, the vacuum state of ($\boldsymbol{k}E + i\boldsymbol{i}\mathbf{p} + i\boldsymbol{j}m$), for example, becoming $\boldsymbol{k}$ ($\boldsymbol{k}E + i\boldsymbol{i}\mathbf{p} + i\boldsymbol{j}m$). In addition,

$$(\boldsymbol{k}E + i\boldsymbol{i}\mathbf{p} + i\boldsymbol{j}m)\ \boldsymbol{k}\ (\boldsymbol{k}E + i\boldsymbol{i}\mathbf{p} + i\boldsymbol{j}m)\ \boldsymbol{k}\ (\boldsymbol{k}E + i\boldsymbol{i}\mathbf{p} + i\boldsymbol{j}m)\ \boldsymbol{k}\ (\boldsymbol{k}E + i\boldsymbol{i}\mathbf{p} + i\boldsymbol{j}m)\ ...$$

is the same as

$$(\boldsymbol{k}E + i\boldsymbol{i}\mathbf{p} + i\boldsymbol{j}m)\ (-\boldsymbol{k}E + i\boldsymbol{i}\mathbf{p} + i\boldsymbol{j}m)\ (\boldsymbol{k}E + i\boldsymbol{i}\mathbf{p} + i\boldsymbol{j}m)\ (\ -\boldsymbol{k}E + i\boldsymbol{i}\mathbf{p} + i\boldsymbol{j}m)\ ....$$

It thus appears that the infinite series of creation acts by a fermion on vacuum is the mechanism for creating an infinite series of alternating boson and fermion states as required for supersymmetry and renormalization. The 'mirror imaging' process implies an infinite range of virtual E values in vacuum adding up to a single finite value, exactly as in renormalisation, with equal numbers of boson and fermion loops cancelling through their opposite signs.

This fundamental relation also leads to the significant fact that the nilpotent wavefunctions, in principle, produce a kind of supersymmetry, with the supersymmetric partners not being so much realisable particles, as the couplings of the fermions and bosons to vacuum states. The nilpotent operators defined for fermion wavefunctions are, in fact, also supersymmetry operators, which produce the supersymmetric partner in the particle itself. The Q generator for supersymmetry is simply the term ($\boldsymbol{k}E + i\boldsymbol{i}\mathbf{p} + i\boldsymbol{j}m$), and its Hermitian conjugate $Q\dagger$ is ($-\boldsymbol{k}E + i\boldsymbol{i}\mathbf{p} + i\boldsymbol{j}m$). Multiplying by ($\boldsymbol{k}E + i\boldsymbol{i}\mathbf{p} + i\boldsymbol{j}m$) converts bosons to fermions, or antifermions to bosons (the $\mathbf{p}$ can, of course, be + or –). Multiplying by ($-\boldsymbol{k}E + i\boldsymbol{i}\mathbf{p} + i\boldsymbol{j}m$) produces the reverse conversion of bosons to antifermions, or fermions to bosons.

In this context, while the spin ½ state is that of the isolated fermion, and due to kinetic energy, implying continuous variation, unit spin comes from the potential energy of a stable state, and represents either a boson with two nilpotents (which are not nilpotent to each other), or a bosonic-type state produced by a fermion interacting with its material environment or vacuum, and, as a consequence, manifesting Berry phase, Aharonov-Bohm or Jahn-Teller effect, Thomas precession, relativistic correction, radiation reaction, quantum Hall effect, Cooper pairing, *zitterbewegung*, or whatever else is needed to produce the ‘conjugate’ environmental spin state. The isolated fermion thus represents the action half of Newton’s third law, while in the case of the fermion interacting with its environment, it is the action and reaction pair. The existence of a ‘supersymmetric’ partner seemingly comes from the duality represented by the choice of fermion or fermion plus environment.

THE AHARONOV-BOHM EFFECT

A consideration of the Aharonov-Bohm effect suggests that it may lead to a more profound understanding of the meaning of the factor 2 in fundamental physics. In this effect, electron interference fringes, produced by a Young’s slit arrangement, are shifted by half a wavelength in the presence of a solenoid whose magnetic field, being internal, does not interact with the electron but whose vector potential does. The half-wavelength shift turns out to be a feature of the topology of the space surrounding the discrete flux-lines of the solenoid, which is not *simply-connected*, and cannot be deformed continuously down to a point. Effectively, the half-wavelength shift, or equivalent acquisition by the electron of a half-wavelength Berry phase, implies that an electron path between source and slit, round the solenoid, involves a *double-circuit* of the flux line (to achieve the same phase), and a path that goes round a circuit twice cannot be continuously deformed into a path which goes round once (as would be the case in a space without flux-lines).

The presence of the flux line is equivalent, as in the quantum Hall effect and fractional quantum Hall effect, to the extra fermionic ½-spin which is provided by the electron acting in step with the nucleus in the Jahn-Teller effect and makes the potential function single-valued, and the circuit for the complete system a single loop. It is particularly significant that the *U*(1) (electromagnetic) group responsible for the fact that the vacuum space is not simply connected is isomorphic to the integers under addition. In effect, the spin-½, ½-wavelength-inducing nature of the fermionic state (in the case of either the electron or the flux line) is a product of discreteness in both the fermion (and its charge) and the space in which it acts. In principle, the very act of creating a discrete particle requires a splitting of the continuum vacuum into *two* discrete halves (as with the bisecting of the rectangular figure with which we started, or, in another context, the Dedekind cut, which defines the relation between real and rational numbers), or (relating the concept of discreteness to that of dimensionality) two square roots of 0. Mathematically, the identification of 1 as separate from 0 also implies that 1 + 1 = 2, reflecting the fact that physics and mathematics have a common origin in the process of counting.

THE ORIGINS OF THE FACTOR 2

The numerical factor 2 has become an almost universal component of fundamental physics, playing a significant role in both quantum theory and relativity. Its origin and meaning can be explained in surprisingly simple terms, using relatively unsophisticated mathematics. We see it in terms of either action and reaction (A); commutation relations (C); absorption and emission (E); object and environment (O); relativity (R); the virial relation (V); or continuity and discontinuity (X). The overlap between many of these explanations in the case of individual phenomena demonstrates that they are really all part of the same overall process:

Kinematics						V	X
Gases	A					V	
Orbits	A					V	X
Radiation pressure	A		E			V	
Gravitational light deflection					R	V	
Fermion / boson spin		C		O	R	V	
Zero-point energy	A	C				V	X
Radiation reaction	A		E		R	V	
SR paradoxes	A		E				

For example, kinetic energy variation may be thought of as continuous, but starting from a discrete state; potential energy variation, on the other hand, is a discrete variation, starting from a continuous state. Each creates the opposite in its variation from itself. Kinetic energy and potential energy create each other, in the same way as they are related by a numerical relationship. But kinetic energy also relates to a changing state, while potential energy is usually related to a fixed one. We can further consider the kinetic energy relation to be concerned with the action side of Newton's third law, while the potential energy relation concerns both action and reaction. The factor 2 is also an expression of the discreteness of both material particles (or charges) and the spaces between them, as opposed to the continuity of the vacuum in terms of energy. The same discreteness also implies (though more subtly) the concept of dimensionality, which is responsible for the noncommutativity of the momentum operator, as well as the discreteness of the division of rectangles into triangles.

In more general terms, the factor 2 is an expression of a fundamental duality in the whole concept of 'nature', a duality that is the result of trying to create something from nothing – the Aharonov-Bohm effect is a classic case, as is also the nilpotent algebra used for the fermion wavefunction. Fundamentally, physics does this when it sets up a probe to investigate an intrinsically uncharacterizable nature. Nature responds with symmetrical opposites to the characterization assumed by the probe, which, in its simplest form, is constituted by a discrete point in space. It has been demonstrated previously that this generates a symmetrical group of fundamental parameters (space – the original probe – time, mass and charge – the combined response), which are defined by properties which split the parameters into three C_2 groupings, depending on whether they are conserved or nonconserved, real (or orderable) or imaginary (or nonorderable), continuous or discrete. Each of these divisions may be held responsible for a factor 2, for duality seems to be the necessary result of any attempt to create singularity.

space	real	nonconserved	countable (3-D)
time	imaginary	nonconserved	noncountable (1-D)
mass	real	conserved	noncountable (3-D)
charge	imaginary	conserved	countable (1-D)

Careful study of the factor 2 reveals that it is either the link between the continuous and discrete physical domains, or between the changing and the fixed, or the real and imaginary (orderable and nonorderable), the three dualities of the group, and, in every physical instance, between more than one of these.

While the continuous v. discrete duality is obvious from the distinction between potential and kinetic energies, this distinction also incorporates the duality between conserved and nonconserved quantities, or fixed and changing conditions. The duality may also be expressed in terms of the distinction between space-like and time-like theories (for example, those of Heisenberg and Schrödinger, or of quantum mechanics and stochastic electrodynamics), which are not only distinguished by being discrete and continuous, but also by being real and imaginary. Though a single duality separates such theories, it is open to more than one interpretation because each pair of parameters is always separated by two distinct dualities.

THE STRUCTURE OF DUALITY

At the most profound level, as we have said, the factor 2 is an expression of the fundamental nature of duality, in physics, mathematics, and even ontology and epistemology. In simple terms, we can’t define something without defining also what it is not, and we can’t even characterize ‘nature’ or ‘reality’, even to the extent of saying whether it has an independent existence (is ontological) or is a product of our perception (is epistemological). It is possible to explain this on the basis that physics and mathematics are attempts at creating something from nothing. A ‘theory of everything’ needs first to be a ‘theory of nothing’. We start from ‘nothing’ and we end with ‘nothing’, and duality is there to ensure that when we introduce ‘something’, we still end with nothing. But this does not mean that we cannot determine its structure. The fundamental duality operates in the most simple way possible.

We begin with the simplest possible symmetry group, C_2, which we can describe in mathematical terms, by the use of the elements 1 and –1, but which, physically, is just anything and its opposite. Thus, as soon as we imagine 1 as different from 0, we also invoke its automatic negation, or the thing we describe as –1. So, defining 1, at all, automatically creates a dual system, equivalent to requiring 1 + 1 = 2, and generating the Peano idea of 'successor'. This seemingly leads to the creation of a natural (binary) numbering system, while avoiding the Gödel problem through the zero totality. (Work on this, and on programming aspects, is in progress with B. M. Diaz.) We also have no option but to relate –1 to 1 in some way other than defining their totality as 0, and the identity –1 × –1 or $(-1)^2 = 1$ then becomes deeply significant in establishing that the relation between these elements is a *group* relationship, and that the 'multiplication' and 'squaring' of elements, in addition to identity and inversion, are operations which are fundamental to the principle of duality.

However, we require a dual, even for this C_2 group. We need to extend to four elements, to find an equivalent to $C_2 \times C_2$. The only way of extending the original group is if the two unknown elements acquire the characters that we describe by the symbols i and $-i$. Though the group of 1, –1, i, $-i$ is not, of course, $C_2 \times C_2$, but C_4, it contains the same *information* as $C_2 \times C_2$, for we can write this information in the form of the complex ordered pairs: 1, i; 1, $-i$; –1, i; –1, $-i$, which *is* of the form $C_2 \times C_2$, and is the natural mathematical expression of complex numbers.

To dual again, we need to imagine another complexification, involving, new terms, which we could describe as j and $-j$. However, we now have the complication of the product ij, which must be yet another new term. The result of this process is the definition of the necessarily *cyclic* and noncommutative operators $\boldsymbol{i}$, $-\boldsymbol{i}$, $\boldsymbol{j}$, $-\boldsymbol{j}$, $\boldsymbol{k}$, $-\boldsymbol{k}$, which we describe as quaternions. The definition of the quaternion group Q_8, with elements 1, –1, $\boldsymbol{i}$, $-\boldsymbol{i}$, $\boldsymbol{j}$, $-\boldsymbol{j}$, $\boldsymbol{k}$, $-\boldsymbol{k}$, is simply a statement of the fact that the complex C_4 group has been dualistically extended, and, we can, again, represent the same information by a C_2 multiplication, using a group of the form $C_2 \times C_2 \times C_2$.

Continuing the process further, we dual Q_8 by complexifying it to the complex quaternion or multivariate 'vector' group 1, –1, i, $-i$, $\boldsymbol{i}$, $-\boldsymbol{i}$, $\boldsymbol{j}$, $-\boldsymbol{j}$, $\boldsymbol{k}$, $-\boldsymbol{k}$, $i\boldsymbol{i}$, $-i\boldsymbol{i}$, $i\boldsymbol{j}$, $-i\boldsymbol{j}$, $i\boldsymbol{k}$, $-i\boldsymbol{k}$, of order 16, which has a related $C_2 \times C_2 \times C_2 \times C_2$ formulation, and which may also be written 1, –1, i, $-i$, $i\mathbf{i}$, $i\mathbf{i}$, $i\mathbf{j}$, $-i\mathbf{j}$, $i\mathbf{k}$, $-i\mathbf{k}$, $\mathbf{i}$, $-\mathbf{i}$, $\mathbf{j}$, $-\mathbf{j}$, $\mathbf{k}$, $-\mathbf{k}$, where a complex quaternion, such as $i\boldsymbol{i}$ becomes the equivalent of the multivariate vector $\mathbf{i}$ (see Appendix I). (The alternative dualling of quaternions to octonions, with sixteen components, fails the test of group structure, as octonions are nonassociative.) We then expand the complex terms to a three-dimensional status, to produce a double quaternion group, say 1, –1, $\boldsymbol{I}$, $-\boldsymbol{I}$, $\boldsymbol{J}$, $-\boldsymbol{J}$, $\boldsymbol{K}$, $-\boldsymbol{K}$, $\boldsymbol{i}$, $-\boldsymbol{i}$, $\boldsymbol{j}$, $-\boldsymbol{j}$, $\boldsymbol{k}$, $-\boldsymbol{k}$, of order 32, which has a related $C_2 \times C_2 \times C_2 \times C_2 \times C_2$ formulation. Then we complexify again, to produce a multivariate vector-quaternion group 1, –1, i, $-i$, $i\mathbf{i}$, $-i\mathbf{i}$, $i\mathbf{j}$, $-i\mathbf{j}$, $i\mathbf{k}$, $-i\mathbf{k}$, $\mathbf{i}$, $-\mathbf{i}$, $\mathbf{j}$, $-\mathbf{j}$, $\mathbf{k}$, $-\mathbf{k}$, $\boldsymbol{i}$, $-\boldsymbol{i}$, $\boldsymbol{j}$, $-\boldsymbol{j}$, $\boldsymbol{k}$, $-\boldsymbol{k}$, $i\boldsymbol{i}$, $-i\boldsymbol{i}$, $i\boldsymbol{j}$, $-i\boldsymbol{j}$, $i\boldsymbol{k}$, $-i\boldsymbol{k}$, and 36 real and complex combinations of vectors and quaternions, forming a group of 64, with a related $C_2 \times C_2 \times C_2 \times C_2 \times C_2 \times C_2$ formulation. This is the algebra of the Dirac gamma matrices. Though further dualling is possible on the same basis, it is clear that only three fundamental principles are required to continue the dualling to infinity – opposite signs (or equivalent), the distinction between real and imaginary components, and the introduction of cyclic dimensionality – and to establish every conceivable combination of these, that is to establish that every type of dualling is itself dualled, requires a group of 64 elements. Thus, the Dirac algebra includes a dualling of each of the dualling processes within itself, the eight groups of objects involved producing every possible combination of + / – × real / complex × nondimensional / dimensional:

C_2	C_2	± 1	
C_4	$C_2 \times C_2$	$\pm 1, \pm i$	complexify
Q_8	$C_2 \times C_2 \times C_2$	$\pm 1, \pm \boldsymbol{i}, \pm \boldsymbol{j}, \pm \boldsymbol{k}$	dimensionalize
V_{16}	$C_2 \times C_2 \times C_2 \times C_2$	$\pm 1, \pm i, \pm \boldsymbol{i}, \pm \boldsymbol{j}, \pm \boldsymbol{k}$	complexify
QQ_{32}	$C_2 \times C_2 \times C_2 \times C_2 \times C_2$	$\pm 1, \pm \boldsymbol{I}, \pm \boldsymbol{J}, \pm \boldsymbol{K}, \pm \boldsymbol{i}, \pm \boldsymbol{j}, \pm \boldsymbol{k}$	dimensionalize
VQ_{64}	$C_2 \times C_2 \times C_2 \times C_2 \times C_2 \times C_2$	$\pm 1, \pm i, \pm \boldsymbol{I}, \pm \boldsymbol{J}, \pm \boldsymbol{K}, \pm \boldsymbol{i}, \pm \boldsymbol{j}, \pm \boldsymbol{k}$	complexify

Though this appears to be a purely mathematical argument, in fact, it has a fundamental physical significance, in relation to the properties of the fundamental parameter group, for we can see now that two of the distinctions between the parameters, which we have derived inductively from observed physical characteristics (real / imaginary and noncountable / countable), are identical to the C_2 distinctions which extend the original C_2 duality into complexity and cyclic dimensionality. (Particularly significant, here, is the fact that countability or discreteness is a necessary requirement for cyclic multidimensionality, for unidimensionality is an obviously necessary property of a continuous or noncountable quantity – it can't have an origin. Multidimensionality is also a necessary property of discreteness, which has to have a reference or origin.) However, even the original C_2 duality (1 / –1) originated from the act of creating 'something from nothing' (1 from 0), the very definition of *nonconservation*, as is the concept of 'successor' which it implies. So, in principle, the group of space, time, mass and charge has all the elements required to extend physical duality to infinity.

And, when we express the parameters mass, time, space, and charge in terms of the respective scalar, pseudoscalar, vector and quaternion units (1, i, $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$, $\boldsymbol{i}$, $\boldsymbol{j}$, $\boldsymbol{k}$), which their combined properties require, it becomes evident that the combination of the four parameters in the Dirac equation produces the complete self-dualling which we require. In addition, the Dirac nilpotent is the perfect way of producing something from nothing; its structure effectively incorporates or generates all the discrete and continuous groups of interest in fundamental physics, from C_2 to E_8 [Rowlands, Cullerne and Koberlien, 2001]; while the infinite imaging of the fermion state in the vacuum and the infinite entanglement of all nilpotent fermion states extend the dualling to infinity.

From the construction of dualities in terms of successive C_2 applications, it is possible to see why, in general, the constant terms in alternative approaches to physical explanation produce effects which are 2 × the changing terms, the real produce ones which are 2 × the imaginary, and the discrete ones which are 2 × the continuous: the multiplication occurs in the direction which doubles the options. The first combines + and – cases where it remains constant; the second involves squaring imaginary parameters to produce real ones; and the third combines dimensionality and noncommutivity with discreteness, and so doubles the elements. Examples of the first include action + reaction, absorption + emission, radiation + reaction, potential v. kinetic energy, relativistic v. rest mass, uniform v. uniformly accelerated motion, and rectangles v. triangles. Examples of the second include bosons v. fermions, and space-like v. time-like systems. Examples of the third include fermion + 'environment' (Aharonov-Bohm, Berry phase, Jahn-Teller, etc.), space-like v. time-like systems, particles v. waves, Heisenberg v. Schrödinger / the harmonic oscillator, quantum mechanics v stochastic electrodynamics / zero point energy; 4π v. 2π rotation, and all cases in which physical dimensionality or noncommutativity is involved.

The very concept of duality also implies that the actual processes of counting and generating numbers are created at the same time as the concepts of discreteness, nonconservation, and orderability are separated from those of continuity, conservation, and nonorderability. The mathematical processes of addition and squaring are, in effect, 'created' at the same time as the physical quantities to which they apply, while all the other fundamental mathematical concepts and processes (e.g. the Dedekind cut) are, in some way, defined by dualling. The factor 2 thus expresses dualities which are fundamental to the creation of both mathematics and physics, and duality provides a philosophy on which both can be based.

REFERENCES

Gough, W. [1990], 'Mixing scalars and vectors – an elegant view of physics'. *Eur. J. Phys.*, **11**, 326-33.

Rowlands, P., Cullerne, J. P., and Koberlein, B. D. [2001], 'The group structure basis of a foundational approach to physics', arXiv:physics/0110091.

Whitney, C. K. [2000], 'How can paradox happen?', *Proceedings of Conference on Physical Interpretations of Relativity Theory VII*, British Society for Philosophy of Science, London, September 2000, 338-51.