International Journal of Modern Physics A
Vol. 19, Supplement (2004) 45–74

World Scientific
www.worldscientific.com

The Dirac Equation

Frank Wilczek
Center for Theoretical Physics
Massachusetts Institute of Technology
Cambridge, MA 02139-4307

> One cannot escape the feeling that these mathematical formulae have an independent existence and an intelligence of their own, that they are wiser than we are, wiser even than their discoverers, that we get more out of them than was originally put into them.
>
> — *H. Hertz, on Maxwell's equations for electromagnetism*

> A great deal of my work is just playing with equations and seeing what they give.
>
> —*P.A.M. Dirac*

> It gave just the properties one needed for an electron. That was really an unexpected bonus for me, completely unexpected.
>
> —*P.A.M. Dirac, on the Dirac equation*

Of all the equations of physics, perhaps the most "magical" is the Dirac equation. It is the most freely invented, the least conditioned by experiment, the one with the strangest and most startling consequences.

In early 1928 (the receipt date on the original paper is January 2), Paul Adrien Maurice Dirac (1902–1984), a 25-year-old recent convert from electrical engineering to theoretical physics, produced a remarkable equation, forever to be known as the Dirac equation. Dirac's goal was quite concrete, and quite topical. He wanted to produce an equation that would describe the behavior of electrons more accurately than previous equations. Those equations incorporated either special relativity or quantum mechanics, but not both. Several other more prominent and experienced physicists were

working on the same problem.

Unlike these other physicists, and unlike the great classics of physics, Newton and Maxwell, Dirac did not proceed from a minute study of experimental facts. Instead he guided his search using a few basic facts and perceived theoretical imperatives, some of which we now know to be wrong. Dirac sought to embody these principles in an economical, mathematically consistent scheme. By "playing with equations," as he put it, he hit upon a uniquely simple, elegant solution. This is, of course, the equation we now call the Dirac equation.

Some consequences of Dirac's equation could be compared with existing experimental observations. They worked quite well, and explained results that were otherwise quite mysterious. Specifically, as I'll describe below, Dirac's equation successfully predicts that electrons are always spinning and that they act as little bar magnets, and the rate of the spin and the strength of the magnetism. But other consequences appeared utterly inconsistent with obvious facts. Notably, Dirac's equation contains solutions that appear to describe a way for ordinary atoms to wink out into bursts of light, spontaneously, in a fraction of a second.

For several years Dirac and other physicists struggled with an extraordinary paradox. How can an equation be "obviously right" since it accounts accurately for many precise experimental results, and achingly beautiful to boot – and yet manifestly, catastrophically wrong?

The Dirac equation became the fulcrum on which fundamental physics pivoted. While keeping faith in its mathematical form, physicists were forced to reexamine the meaning of the symbols it contains. It was in this confused, intellectually painful re-examination – during which Werner Heisenberg wrote to his friend Wolfgang Pauli, "The saddest chapter of modern physics is and remains the Dirac theory" and "In order not to be irritated with Dirac I have decided to do something else for a change···" – that truly modern physics began.

A spectacular result was the prediction of *antimatter* – more precisely, that there should be a new particle with the same mass as the electron, but the opposite electric charge, and capable of annihilating an electron into pure energy. Particles of just this type were promptly identified, through painstaking scrutiny of cosmic ray tracks, by Carl Anderson in 1932.

The more profound, encompassing result was a complete reworking of the foundations of our description of matter. In this new physics, particles are mere ephemera. They are freely created and destroyed; indeed, their fleeting existence and exchange is the source of all interactions. The truly

fundamental objects are universal, transformative ethers: quantum fields. These are the concepts that underlie our modern, wonderfully successful Theory of Matter (usually called, quite inadequately, the Standard Model). And the Dirac equation itself, drastically reinterpreted and vastly generalized, but never abandoned, remains a central pillar in our understanding of Nature.

7. Dirac's Problem and the Unity of Nature

The immediate occasion for Dirac's discovery, and the way he himself thought about it, was the need to reconcile two successful, advanced theories of physics that had gotten slightly out of synch. By 1928 Einstein's special theory of relativity was already over two decades old, well digested, and fully established. (The general theory, which describes gravitation, is not part of our story here. Gravity is negligibly weak on atomic scales.) On the other hand, the new quantum mechanics of Heisenberg and Schrödinger, although quite a young theory, had already provided brilliant insight into the structure of atoms, and successfully explained a host of previously mysterious phenomena. Clearly, it captured essential features of the dynamics of electrons in atoms. The difficulty was that the equations developed by Heisenberg and Schrödinger did not take off from Einstein's relativistic mechanics, but from the old mechanics of Newton. Newtonian mechanics can be an excellent approximation for systems in which all velocities are much smaller than the speed of light, and this includes many cases of interest in atomic physics and chemistry. But the experimental data on atomic spectra, which one could address with the new quantum theory, was so accurate that small deviations from the Heisenberg-Schrödinger predictions could be observed. So there was a strong "practical" motivation to search for a more accurate electron equation, based on relativistic mechanics. Not only young Dirac, but also several other major physicists, were after such an equation.

In hindsight we can discern that much more ancient and fundamental dichotomies were in play: light versus matter; continuous versus discrete. These dichotomies present tremendous barriers to the goal of achieving a unified description of Nature. Of the theories Dirac and his contemporaries sought to reconcile, relativity was the child of light and the continuum, and quantum theory the child of matter and the discrete. After Dirac's revolution had run its course, all were reconciled, in the mind-stretching conceptual amalgam we call a quantum field.

The dichotomies light/matter and continuous/discrete go deep. They were experienced by the earliest sentient proto-humans. They were articulated clearly, and debated inconclusively, by the ancient Greeks. Specifically, Aristotle distinguished Fire and Earth as primary elements - light versus matter. And he argued, against the Atomists, in favor of a fundamental plenum ("Nature abhors a vacuum") - upholding the continuous, against the discrete.

These dichotomies were not relieved by the triumphs of classical physics; indeed, they were sharpened.

Newton's mechanics is best adapted to describing the motion of rigid bodies through empty space. While Newton himself in various places speculated on the possible primacy of either side of both dichotomies, Newton's followers emphasized his "hard, massy, impenetrable" atoms as the fundamental building-blocks of Nature. Even light was modeled in terms of particles.

Early in the nineteenth century a very different picture of light, according to which it consists of waves, scored brilliant successes. Physicists accepted that there must be a continuous, space-filling ether to support these waves. The discoveries of Faraday and Maxwell, assimilating light to the play of electric and magnetic fields, which are themselves continuous entities filling all space, refined and reinforced this idea.

Yet Maxwell himself, and Ludwig Boltzmann, succeeded in showing that the observed properties of gases, including many surprising details, could be explained if the gases were composed of many small, discrete, well-separated atoms moving through otherwise empty space. Furthermore J.J. Thomson experimentally, and Hendrik Lorentz theoretically, established the existence of electrons as building-blocks of matter. Electrons appear to be indestructible particles, of the sort that Newton would have appreciated.

Thus as the twentieth century opened, physics featured two quite different sorts of theories, living together in uneasy peace. Maxwell's electrodynamics is a continuum theory of electric and magnetic fields, and of light, that makes no mention of mass. Newton's mechanics is a theory of discrete particles, whose *only* mandatory properties are mass and electric charge[a].

Early quantum theory developed along two main branches, following the fork of our dichotomies, but with hints of convergence.

[a]That is, to predict the motion of a particle you need to know its charge and its mass: no more, no less. The value of the charge can be zero; then the particle will have only gravitational interactions.

One branch, beginning with Planck's work on radiation theory, and reaching a climax in Einstein's theory of photons, dealt with light. Its central result is that light comes in indivisible minimal units, photons, with energy and momentum proportional to the frequency of the light. This, of course, established a particle-like aspect of light.

The second branch, beginning with Bohr's atomic theory and reaching a climax in Schrödinger's wave equation, dealt with electrons. It established that the stable configurations of electrons around atomic nuclei were associated with regular patterns of wave vibrations. This established a wave-like property of matter.

Thus the fundamental dichotomies softened. Light is a bit like particles, and electrons are a bit like waves. But sharp contrasts remained. Two differences, in particular, appeared to distinguish light from matter sharply.

First, if light is to be made of particles, then they must be very peculiar particles, with internal structure, for light can be polarized. To do justice to this property of light, its particles must have some corresponding property. There can't be an adequate description of a light beam specifying only that it is composed of so-and-so many photons with such-and-such energies; those facts will tell us how bright the beam is, and what colors it contains, but not how it is polarized. To get a complete description, one must also be able to say which way the beam is polarized, and this means that its photons must somehow carry around arrows that allow them to keep a record of the light's polarity. This would seem to take us away from the traditional ideal of elementary particles. If there's an arrow, what's *it* made of? – and why can't it be separated from the particle?

Second, and more profound, photons are evanescent. Light can be radiated, as when you turn on a flashlight, or absorbed, as when you cover it with your hand. Therefore particles of light can be created or destroyed. This basic, familiar property of light and photons takes us far away from the traditional ideal of elementary particles. The stability of matter would seem to require indestructible building-blocks, with properties fundamentally different from evanescent photons.

The Dirac equation, and the crisis it provoked, forced physicists, finally, to transcend all these dichotomies. The consequence is a unified concept of substance, that is surely one of mankind's greatest intellectual achievements.

8. The Early Payoff: Spin

Dirac was working to reconcile the quantum mechanics of electrons with special relativity. He thought – mistakenly, we now know – that quantum theory required equations of a particularly simple kind, the kind mathematicians call first-order. Never mind why he thought so, or precisely what first-order means; the point is that he wanted an equation that is, in a certain very precise sense, of the simplest possible kind. Tension arises because it is not easy to find an equation that is both simple in this sense and also consistent with the requirements of special relativity. To construct such an equation, Dirac had to expand the terms of the discussion. He found he could not get by with a single first-order equation – he needed a system of four intricately related ones, and it is actually this system we refer to as "the" Dirac equation.

Two equations were quite welcome. Four, initially, were a big problem.

First, the good news.

Although the Bohr theory gave a good rough account of atomic spectra, there were many discrepant details. Some of the discrepancies concerned the number of electrons that could occupy each orbit, others involved the response of atoms to magnetic fields, as manifested in the movement of their spectral lines. Wolfgang Pauli had shown, through detailed analysis of the experimental evidence, that Bohr's model could only work, even roughly, for complex atoms if there were a tight restriction on how many electrons could occupy any given orbit. This is the origin of the famous Pauli exclusion principle. Today we learn this principle in the form "only one electron can occupy a given state". But Pauli's original proposal was not so neat; it came with some disturbing fine print. For the number of electrons that could occupy a given Bohr orbital was not one, but two. Pauli spoke obscurely of a "classically non-describable duplexity", but – needless to say – did not describe any reason for it.

In 1925 two Dutch graduate students, Samuel Goudsmit and George Uhlenbeck, devised a possible explanation of the magnetic response problems. If electrons were actually tiny magnets, they showed, the discrepancies would disappear. Their model's success required that all electrons must have the same magnetic strength, which they could calculate. They went on to propose a mechanism for the electron's magnetism. Electrons, of course, are electrically charged particles. Electric charge in circular motion generates magnetic fields. Thus, if for some reason electrons were always rotating about their own axis, their magnetism might be explained. This

intrinsic *spin* of electrons would have an additional virtue. If the rate of spin were the minimum allowed by quantum mechanics[b], then Pauli's "duplexity" would be explained. For the spin would have no possibility to vary in magnitude, but only the possibility to point either up or down. Many eminent physicists were quite skeptical of Goudsmit and Uhlenbeck. Pauli himself tried to dissuade them from publishing their work. For one thing, their model seemed to require the electron to rotate at an extraordinarily rapid rate, at its surface probably faster than the speed of light. For another, they gave no account of what holds an electron together. If it is an extended distribution of electric charge, all of the same sign, it will want to fly apart – and rotation, by introducing centrifugal forces, only makes the problem worse. Finally, there was a quantitative mismatch between their requirements for the strength of the electron's magnetism and the amount of its spin. The ratio of these two quantities is governed by a factor called the gyromagnetic ratio, written g. Classical mechanics predicts $g = 1$, whereas to fit the data Goudsmit and Uhlenbeck postulated $g = 2$. But despite these quite reasonable objections, their model stubbornly continued to agree with experimental results!

Enter Dirac. His system of equations allowed a class of solutions, for small velocities, in which only two of the four functions appearing in his equations are appreciable. This was duplexity, but with a difference. Here it fell out automatically as a consequence of implementing general principles, and most definitely did not have to be introduced *ad hoc*. Better yet, using his equation Dirac could calculate the magnetism of electrons, also without further assumptions. He got $g = 2$. Dirac's great paper of 1928 wastes no words. Upon demonstrating this result, he says simply

> The magnetic moment is just that assumed in the spinning electron model.

And a few pages later, after working out the consequences, he concludes laconically

> The present theory will thus, in the first approximation, lead to the same energy levels as those obtained by [C.G.] Darwin, which are in agreement with experiment.

His results spoke loudly for themselves, with no need for amplification. From

[b]In quantum mechanics, only certain values of the discrete spin are allowed. This is closely related to the restriction on allowed Bohr orbitals.

then on, there was no escaping Dirac's equation. Whatever difficulties arose – and there were some big and obvious ones – they would be occasions for struggle, not desertion. Such gleaming jewels of insight would be defended at all costs.

Although his intellectual starting point, as I mentioned, was quite different and more abstract, Dirac begins his paper by referring to Goudsmit, Uhlenbeck, and the experimental success of their model. Only in the second paragraph does he reveal his hand. What he says is quite pertinent to the themes I emphasized above.

> The question remains as to why Nature should have chosen this particular model for the electron instead of being satisfied with a point-charge. One would like to find some incompleteness in the previous methods of applying quantum mechanics to the point-charge such that, when removed, the whole of the duplexity phenomena follow without arbitrary assumptions.

Thus Dirac is not offering a new model of electrons, as such. Rather, he is defining a new *irreducible* property of matter, inherent in the nature of things, specifically in the consistent implementation of relativity and quantum theory, that arises even in the simplest possible case of structureless point particles. Electrons happen to be embodiments of this simplest possible form of matter. The valuable properties of Goudsmit and Uhlenbeck's "spin", specifically its fixed magnitude and its magnetic action, which aid in the description of observed realities, were retained, now based on a much deeper foundation. The arbitrary and unsatisfactory features of their model are bypassed.

We were looking for an arrow that would be a necessary and inseparable part of elementary bits of matter, like polarization for photons. Well, there it is!

The spin of the electron has many practical consequences. It is responsible for the phenomenon of ferromagnetism, and the enhancement of magnetic fields in the core of electric coils, which forms the heart of modern power technology (motors and dynamos). Active manipulation of electron spins allows us to store and retrieve a great deal of information in a very small volume (magnetic tape, disk drives). Even the much smaller and more inaccessible spin of atomic nuclei plays a big role in modern technology. Manipulating such spins with radio and magnetic fields, and sensing their response, is the basis of the magnetic resonance imaging (MRI) so useful in medicine. This application, among many others, would be inconceivable

(literally!) without the exquisite control of matter that only fundamental understanding can bring.

Spin in general, and Dirac's prediction for the magnetic moment in particular, has also played a seminal role in the subsequent development of fundamental physics. Small deviations from Dirac's $g = 2$ were discovered by Polykarp Kusch and collaborators in the 1940s. They provided some of the first quantitative evidence for the effects of virtual particles, a deep and characteristic property of quantum field theory. Very large deviations from $g = 2$ were observed for protons and neutrons in the 1930s. This was an early indication that protons and neutrons are not fundamental particles in the same sense that electrons are. But I'm getting ahead of the story.···

9. The Dramatic Surprise: Antimatter

Now for the 'bad' news.

Dirac's equation consists of four components. That is, it contains four separate wave functions to describe electrons. Two components have an attractive and immediately successful interpretation, as we just discussed, describing the two possible directions of an electron's spin. The extra doubling, by contrast, appeared at first to be quite problematic.

In fact, the extra equations contain solutions with *negative* energy (and either direction of spin). In classical (non-quantum) physics the existence of extra solutions would be embarrassing, but not necessarily catastrophic. For in classical physics, you can simply choose not to use these solutions. Of course that begs the question why *Nature* chooses not to use them, but it is a logically consistent procedure. In quantum mechanics, even this option is not available. In quantum physics, generally "that which is not forbidden is mandatory". In the specific case at hand, we can be quite specific and precise about this. All solutions of the electron's wave equation represent possible behaviors of the electron, that will arise in the right circumstances. Assuming Dirac's equation, if you start with an electron in one of the positive-energy solutions, you can calculate the rate for it to emit a photon and transition into one of the negative-energy solutions. Energy must be conserved overall, but that is not a problem here – it just means that the energy of the emitted photon would be *more* than that of the electron which emitted it! Anyway, the rate turns out to be ridiculously fast, a small fraction of a second. So you can't ignore the negative-energy solutions for long. And since an electron has never been observed to do something so peculiar as radiating more energy than it starts with, there

was, on the face of it, a terrible problem with the quantum mechanics of Dirac's equation.

Dirac was well aware of this problem. In his original paper, he simply acknowledged

> For this second class of solutions W [the energy] has a negative value. One gets over the difficulty on the classical theory by arbitrarily excluding those solutions that have a negative W. One cannot do this on the quantum theory, since in general a perturbation will cause transitions from states with W positive to states with W negative.··· The resulting theory is therefore still only an approximation, but it appears to be good enough to account for all the duplexity phenomena without arbitrary assumptions.

and left it at that. This was the situation that provoked Heisenberg's outbursts to Pauli, quoted earlier.

By the end of 1929 – not quite two years later – Dirac made a proposal to address the problem. It exploited the Pauli exclusion principle, according to which no two electrons obey the same solution of the wave equation. What Dirac proposed was a radically new conception of empty space. He proposed that what we consider 'empty' space is in reality chock-a-block with negative-energy electrons. In fact, according to Dirac, *'empty' space actually contains electrons obeying all the negative energy solutions.* The great virtue of this proposal is that it explains away the troublesome transitions from positive to negative solutions. A positive-energy electron can't go to a negative-energy solution, because there's always another electron already there, and the Pauli exclusion principle won't allow a second one to join it.

It sounds outrageous, on first hearing, to be told that what we perceive as empty space is actually quite full of stuff. But, on reflection, why not? We have been sculpted by evolution to perceive aspects of the world that are somehow useful for our survival and reproductive success. Since unchanging aspects of the world, upon which we can have little influence, are not useful in this way, it should not seem terribly peculiar that they would escape our untutored perception. In any case, we have no warrant to expect that naive intuitions about what is weird or unlikely provide reliable guidance for constructing models of fundamental structure in the microworld, because these intuitions derive from an entirely different realm of phenomena. We must take it as it comes. The validity of a model must be judged according to the fruitfulness and accuracy of its consequences.

So Dirac was quite fearless about outraging common sense. He focused, quite properly, on the observable consequences of his proposal.

Since we are considering the idea that the ordinary state of "empty" space is far from empty, it is helpful to have a different, more non-committal word for it. The one physicists like to use is "vacuum".

In Dirac's proposal, the vacuum is full of negative-energy electrons. This makes the vacuum a medium, with dynamical properties of its own. For example, photons can interact with the vacuum. One thing that can happen is that if you shine light on the vacuum, providing photons with enough energy, then a negative-energy electron can absorb one of these photons, and go into a positive-energy solution. The positive-energy solution would be observed as an ordinary electron, of course. But in the final state there is also a *hole* in the vacuum, because the solution originally occupied by the negative-energy electron is no longer occupied.

The idea of holes was, in the context of a dynamical vacuum, startlingly original, but it was not quite unprecedented. Dirac drew on an analogy with the theory of heavy atoms, which contain many electrons. Within such atoms, some of the electrons correspond to solutions of the wave equation that reside nearby the highly charged nucleus, and are very tightly bound. It takes a lot of energy to break such electrons free, and so under normal conditions they present an unchanging aspect of the atom. But if one of these electrons absorbs a high-energy photon (an X-ray) and is ejected from the atom, the change in the normal aspect of the atom is marked by its *absence*. The absence of an electron, which would have supplied negative charge, by contrast looks like a positive charge. The positive effective charge follows the orbit of the missing electron, so it has the properties of a positively charged particle.

Based on this analogy and other hand-waving arguments – the paper is quite short, and practically devoid of equations – Dirac proposed that holes in the vacuum are positively charged particles. The process where a photon excites a negative-energy electron in the vacuum to a positive energy is then interpreted as the photon creating an electron and a positively charged particle (the hole). Conversely, if there is a preexisting hole, then a positive-energy electron can emit a photon and occupy the vacant negative-energy solution. This is interpreted as the annihilation of an electron and a hole into pure energy. I referred to a photon being emitted, but this is only one possibility. Several photons might be emitted, or any other form of radiation that carries away the liberated energy.

Dirac's first hole theory paper was entitled "A Theory of Electrons and Protons". At the time protons were the only known positively charged particles. It was therefore natural to try to identify the hypothetical holes as protons. But severe difficulties with this identification were soon evident. Specifically, the two sorts of process we just discussed – production of electron-proton pairs, and annihilation of electron-proton pairs – have never been observed. The second is especially problematic, because it predicts that hydrogen atoms spontaneously self-destruct in microseconds – which, thankfully, they do not.

There was also a logical difficulty with the identification of holes with protons. Based on the symmetry of the equations, one could demonstrate that the holes must have the same mass as the electrons. But a proton has, of course, a much larger mass than an electron.

In 1931 Dirac withdrew his earlier identification of holes with protons, and accepted the logical outcome of his own equation and the dynamical vacuum it required:

> A hole, if there was one, would be a new kind of elementary particle, unknown to experimental physics, having the same mass and opposite charge of the electron.

On August 2, 1932, Carl Anderson, an American experimentalist studying photographs of the tracks left by cosmic rays in a cloud chamber, noticed some tracks that lost energy as expected for electrons, but were bent in the opposite direction by the magnetic field. He interpreted this as indicating the existence of a new particle, now known as the antielectron or positron, with the same mass as the electron but the opposite electric charge. Ironically, Anderson was completely unaware of Dirac's prediction.

Thousands of miles away from his rooms at Saint John's, Dirac's holes – the product of his theoretical vision and revision – had been found, descending from the skies of Pasadena. So in the long run the "bad" news turned out to be "even better" news. Negative-energy frogs became positronic princes.

Today positrons are no longer a marvel, but a tool. A notable use is to take pictures of the brain in action – PET scans, for positron-electron tomography. How do positrons get into your head? They are snuck in by injecting molecules containing atoms whose nuclei are radioactive, and decay with positrons as one of their decay products. These positrons do not go very far before they annihilate against some nearby electron, usually producing two photons, which escape your skull, and can be detected. Then you can reconstruct where the original molecule went, to map out metabolism,

and you can also study the energy loss of the photons on the way out, to get a density profile, and ultimately an image, of the brain tissue.

Another notable application is to fundamental physics. You can accelerate positrons to high energy, as you can of course electrons, and bring the beams together. Then the positrons and electrons will annihilate, producing a highly concentrated form of "pure energy". Much of the progress in fundamental physics over the past half century has been based on studies of this type, at a series of great accelerators all over the world, the latest and greatest being the LEP (large electron-positron) collider at CERN, outside Geneva. I'll be discussing a stunning highlight of this physics a little later.

The physical ideas of Dirac's hole theory, which as I mentioned had some of its roots in the earlier study of heavy atoms, fed back in a big way into solid state physics. In solids one has a reference or ground configuration of electrons, with the lowest possible energy, in which electrons occupy all the available states up to a certain level. This ground configuration is the analogue of the vacuum in hole theory. There are also configurations of higher energy, wherein some of the low-energy states are not used by any electron. In these configurations there are vacancies or "holes" – that's what they're called, technically – where an electron would ordinarily be. Such holes behave in many respects like positively charged particles. Solid-state diodes and transistors are based on clever manipulation of holes and electron densities at junctions between different materials. One also has the beautiful possibility to direct electrons and holes to a place where they can combine (annihilate). This allows you to design a source of photons that you can control quite precisely, and leads to such mainstays of modern technology as LEDs (light-emitting diodes) and solid-state lasers.

In the years since 1932 many additional examples of anti-particles have been observed. In fact, for every particle that has ever been discovered, a corresponding anti-particle has also been found. There are antineutrons, antiprotons, antimuons (the muon itself is a particle very similar to the electron, but heavier), antiquarks of various sorts, even antineutrinos, and anti-π mesons, anti-K mesons,[c]. Many of these particles do not obey the Dirac equation, and some of them do not even obey the Pauli exclusion principle. So the physical reason for the existence of antimatter must be very general – much more general than the arguments that first led Dirac to predict the existence of positrons.

[c]An interesting case is the photon, which is its own antiparticle. This is not possible for a charged particle, but the photon is electrically neutral.

In fact, there is a very general argument that if you implement both quantum mechanics and special relativity, every particle must have a corresponding antiparticle. A proper presentation of the argument requires either a sophisticated mathematical background or a lot of patience. Here I'll be content with a rough version, which shows why antimatter is a plausible consequence of implementing both relativity and quantum mechanics, but doesn't quite nail the case.

Consider a particle, let's say a shmoo, to give it a name (while emphasizing that it could be *anything*), moving east at very nearly the speed of light. According to quantum mechanics, there is actually some uncertainty in its position. So there's some probability, if you measure it, that you will find that the shmoo is slightly west of its expected mean position at an initial time, and slightly east of its expected mean position at a later time. So it has traveled further than you might have expected during this interval – which means it was traveling more quickly. But since the expected velocity was essentially the speed of light, the faster speeds required to accommodate uncertainty threaten to violate special relativity, which requires that particles cannot move faster than the speed of light. It's a paradox.

With antiparticles, you can escape the paradox. It requires orchestrating a symphony of weird ideas, but it's the only way people have figured out how to do it, and it seems to be Nature's way. The central idea is that, yes, uncertainty does mean that you can find a shmoo where special relativity tells you your shmoo can't be - but the shmoo you observe is not necessarily the same as the one you were looking for! For it's also possible that at the later time there are two shmoos, the original one and a new one. To make this consistent there must also be an anti-shmoo, to balance the charge, and to cancel out any other conserved quantities that might be associated with the additional shmoo. What about the energy balance - aren't we getting out more than we put in? Here, as often in quantum theory, to avoid contradictions you must be specific and concrete in thinking about what it means to measure something. One way to measure the shmoo's position would be to shine light on it. But to measure the position of a fast-moving shmoo accurately we have to use high-energy photons, and there's also then the possibility such a photon will create a shmoo-anti-shmoo pair. And in that case – closing the circle – when you report the result of your position measurement, you might be talking about the wrong shmoo!

10. The Deepest Meanings: Quantum Field Theory

Dirac's hole theory is brilliantly clever, but Nature goes deeper. Although hole theory is internally consistent, and can cover a wide range of applications, there are several important considerations that force us to go beyond it.

First, there are particles that do not have spin, and do not obey the Dirac equation, and yet have antiparticles. This is no accident: the existence of antiparticles is a general consequence of combining quantum mechanics and special relativity, as I just discussed. Specifically, for example, positively charged π^+ mesons (discovered in 1947) or W^+ bosons (discovered in 1983) are quite important players in elementary particle physics, and they do have antiparticles π^- and W^-. But we can't use Dirac's hole theory to make sense of these antiparticles, because π^+ and W^+ particles don't obey the Pauli exclusion principle. So there is no possibility of interpreting their antiparticles as holes in a filled sea of negative-energy solutions. If there are negative-energy solutions, whatever equation they satisfy[d], occupying them with one particle will not prevent another particle from entering the same state. Thus catastrophic transitions into negative-energy states, which Dirac's hole theory prevents for electrons, must be banished in a different way.

Second, there are processes in which the number of electrons minus the number of positrons changes. An example is the decay of a neutron into a proton, an electron, and an antineutrino. In hole theory the excitation of a negative-energy electron into a positive-energy state is interpreted as creation of a positron-electron pair, and de-excitation of a positive-energy electron into an unoccupied negative-energy state is interpreted as annihilation of an electron-positron pair. In neither case does the difference between the number of electrons and the number of positrons change. Hole theory cannot accommodate changes in this difference. So there are definitely important processes in Nature, even ones specifically involving electrons, that do not fit easily into Dirac's hole theory.

The third and final reason harks back to our initial discussion. We were looking to break down the great dichotomies light/matter and continuous/discrete. Relativity and quantum mechanics, separately, brought us close to success, and the Dirac equation, with its implication of spin,

[d]In fact these particles obey wave equations that do have negative-energy solutions.

brought us closer still. But so far we haven't quite got there. Photons are evanescent, electrons ... well, they're evanescent too, as a matter of experimental fact, as I just mentioned, but we haven't yet adequately fit that feature into our theoretical discussion. In hole theory electrons can come and go, but only as positrons go and come.

These are not so much contradictions as indications of missed opportunity. They indicate that there ought to be some alternative to hole theory that covers all forms of matter, and that treats the creation and destruction of particles as a primary phenomenon.

Ironically, Dirac himself had earlier constructed the prototype of such a theory. In 1927, he applied the principles of the new quantum mechanics to Maxwell's equations of classical electrodynamics. He showed that Einstein's revolutionary postulate that light comes in particles – photons – was a consequence of the logical application of these principles, and that the properties of photons were correctly accounted for. Few observations are so common as that light can be created from non-light, say by a flashlight, or aborbed and annihilated, say by a black cat. But translated into the language of photons, this means that the quantum theory of Maxwell's equations is a theory of the creation and destruction of particles (photons). Indeed, the electromagnetic field appears, in Dirac's quantum theory of electromagnetism, primarily as an agent of creation and destruction. Photons arise as excitations of this field, which is the primary object. Photons come and go, but the field abides. The full significance of this development seems to have escaped Dirac and all of his contemporaries for some time, perhaps precisely because of the apparent specialness of light (dichotomy!). But it is a general construction, which can be applied to the object that appears in Dirac's equation – the electron field – as well.

The result of a logical application of the principles of quantum mechanics to Dirac's equation is an object similar to what he found for Maxwell's equations. It is an object that destroys electrons, and creates positrons[e]. Both are examples of *quantum fields.* When the object that appears in Dirac's equation is interpreted as a quantum field, the negative-energy solutions take on a completely different meaning, with no problematic aspects. The positive-energy solutions multiply electron destruction operators, while the negative-energy solutions multiply positron creation operators. In this framework, the difference between the two kinds of solution is that negative

[e]There is also a closely related object, the Hermitean conjugate, that creates electrons and destroys positrons.

energy represents the energy you need to borrow to make a positron, while positive energy is what you gain by destroying an electron. The possibility of negative numbers is no more paradoxical here than in your bank balance.

With the development of quantum field theory, the opportunities that Dirac's equation and hole theory made evident, but did not quite fulfill, were finally met. The description of light and matter was put, at last, on a common footing. Dirac said, with understandable satisfaction, that with the emergence of quantum electrodynamics physicists had attained foundational equations adequate to describe "all of chemistry, and most of physics".

In 1932 Enrico Fermi constructed a successful theory of radioactive decays (beta decays), including the neutron decay I mentioned before, by exporting the concepts of quantum field theory far from their origin. Since these processes involve the creation and destruction of protons – the epitome of 'stable' matter – the old dichotomies had finally been transcended. Both particles and light are epiphenomena, surface manifestations of the deeper and abiding realities, quantum fields. These fields fill all of space, and in this sense they are continuous. But the excitations they create, whether we recognize them as particles of matter or as particles of light, are discrete.

In hole theory we had a picture of the vacuum as filled with a sea of negative-energy electrons. In quantum field theory, the picture is quite different from this. But there is no returning to innocence. The new picture of the vacuum differs even more radically from naive "empty space". Quantum uncertainty, combined with the possibility of processes of creation and destruction, implies a vacuum teeming with activity. Pairs of particles and antiparticles fleetingly come to be and pass away. I once wrote a sonnet about virtual particles, and here it comes:

Beware of thinking nothing's there –
Remove what you can; despite your care
Behind remains a restless seething
Of mindless clones beyond conceiving.

They come in a wink, and dance about;
Whatever they touch is seized by doubt:
What am I doing here? What should I weigh?
Such thoughts often lead to rapid decay.

Fear not! The terminology's misleading;
Decay is virtual particle breeding

And seething, though mindless, can serve noble ends,
The clone-stuff, exchanged, makes a bond between friends.

To be or not? The choice seems clear enough,
But Hamlet oscillated. So does this stuff.

11. Aftermaths

With the genesis of quantum field theory, we reach a natural intellectual boundary for our discussion of the Dirac equation. By the mid-1930s the immediate paradoxes this equation raised had been resolved, and its initial promise had been amply fulfilled. Dirac received the Nobel Prize in 1933, Anderson in 1935.

In later years the understanding of quantum field theory deepened, and its applications broadened. Using it, physicists have constructed (and established with an astonishing degree of rigor and beyond all reasonable doubt) what will stand for the foreseeable future – perhaps for all time – as the working Theory of Matter. How this happened, and the nature of the theory, is an epic story involving many other ideas, in which the Dirac equation as such plays a distinguished but not a dominant role. But some later developments are so closely linked to our main themes, and so pretty in themselves, that they deserve mention here.

There is another sense in which the genesis of quantum field theory marks a natural boundary. It is the limit beyond which Dirac himself did not progress. Like Einstein, in his later years Dirac took a separate path. He paid no attention to most of the work of other physicists, and dissented from the rest. In the marvelous developments that his work commenced, Dirac's own participation was peripheral.

11.1. *QED and Magnetic Moments*

Interaction with the ever-present dynamical vacuum of quantum field theory modifies the observed properties of particles. We do not see the hypothetical properties of the "bare" particles, but rather the physical particles, "dressed" by their interaction with the quantum fluctuations in the dynamical vacuum.

In particular, the physical electron is not the bare electron, and it does not quite satisfy Dirac's $g = 2$. When Polykarp Kusch made very accurate measurements, in 1947, he found that g is larger than 2 by a factor 1.00119. Now this is not a very large correction, quantitatively, but it was a great

stimulus to theoretical physics, because it provided a very concrete challenge. At that time there were so many loose ends in fundamental physics – a plethora of unexpected, newly discovered particles including muons, π mesons, and others, no satisfactory theory explaining what force holds atomic nuclei together, fragmentary and undigested results about radioactive decays, anomalies in high-energy cosmic rays – that it was hard to know where to focus. In fact, there was a basic philosophical conflict about strategy.

Most of the older generation, the founders of quantum theory, including Einstein, Schrödinger, Bohr, Heisenberg, and Pauli, were prepared for another revolution. They thought it was fruitless to spend time trying to carry out more accurate calculations in quantum electrodynamics, since this theory was surely incomplete and probably just wrong. It did not help that the calculations required to get more accurate results are very difficult, and that they seemed to give senseless (infinite) answers. So the old masters were searching for a different kind of theory, unfortunately with no clear direction.

Ironically, it was a younger generation of theorists – Schwinger, Feynman, Dyson, and Tomonaga in Japan – who played a conservative role[f]. They found a way to perform the more accurate calculations, and get meaningful finite results, without changing the underlying theory. The theory they used, in fact, was just the one Dirac had constructed in the 20s and 30s. The result of an epochal calculation by Schwinger, including the effects of the dynamic vacuum, was a small correction to Dirac's $g = 2$. It too was reported in 1947, and it agreed spectacularly well with Kusch's contemporary measurements. Many other triumphs followed. Kusch received the Nobel Prize in 1955; Schwinger, Feynman, and Tomonaga jointly in 1965 (the delay is hard to understand!).

Strangely enough, Dirac did not accept the new procedures. Caution was perhaps justified in the early days, when the mathematical methods being used were unfamiliar and not entirely well defined and involved a certain amount of inspired guesswork. But the technical difficulties were cleaned up in due course. [g]

[f]Seminal contributions were also made by the slightly older theorists Kramers and Bethe, and by the theorist-turned-experimentalist Lamb.

[g]Although QED does have problems of principle, if it is regarded (unrealistically!) as a completely closed theory, they are problems at a different level than what troubled Dirac, and they are very plausibly solved by embedding QED into a larger, asymptotically free theory - see below. This has very little practical effect on most of its predictions.

Feynman called QED "the jewel of physics – our proudest possession." But in 1951 Dirac wrote

> Recent work by Lamb, Schwinger and Feynman and others has been very successful... but the resulting theory is an ugly and incomplete one.

And in his last paper, in 1984,

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

You might notice a certain contrast in tone between the young Dirac, who clung to his equation like a barnacle because it explained experimental results, and the older inhabitant of the same body.

Today the experimental determination of the magnetic moment of the electron is

$$(g/2)_{\text{expermient}} = 1.001\,159\,652\,188\,4\,(43)$$

while the theoretical prediction, firmly based on QED, calculated to high accuracy, is

$$(g/2)_{\text{theory}} = 1.001\,159\,652\,187\,9\,(43)$$

where the uncertainty in the last two digits is indicated. It is the toughest, most accurate confrontation between intricate – but precisely defined! – theoretical calculations and delicate – but precisely controlled! – experiments in all of science. That's what Feynman meant by "our proudest possession".

Ever more accurate determination of the magnetic moment of the electron, and of its kindred particle the muon, remains an important frontier of experimental physics. With the accuracies now achievable, the results will be sensitive to effects of quantum fluctuations due to hypothetical new heavy particles – in particular, those expected to be associated with supersymmetry.

11.2. *QCD and the Theory of Matter*

The magnetic moment of the proton does not satisfy Dirac's $g = 2$, but instead has $g \approx 5.6$. For neutrons it is worse. Neutrons are electrically

neutral, so the simple Dirac equation for neutrons predicts no magnetic moment at all. In fact the neutron has a magnetic moment about 2/3 as large as that of a proton, and with the opposite orientation relative to spin. That corresponds to an infinite value of g, since the neutron is electrically neutral. The discrepant values of these magnetic moments were the earliest definite indication that protons and neutrons are more complicated objects than electrons.

With further study, many more complications appeared. The forces among protons and neutrons were found to be very complicated. They depend not only on the distance between them, but also on their velocities, and spin orientations, and all combinations of these together, in a bewildering way. In fact, it soon appeared that they are not "forces" in the traditional sense at all. To have a force between protons, in the traditional sense, would mean that the motion of one proton can be affected by the presence of another, so that when you shoot one proton by another, it swerves. What you actually observe is that when one proton collides with another, typically many particles emerge, most of which are highly unstable. There are π mesons, K mesons, ρ mesons, Λ and Σ baryons, their antiparticles, and many more. All these particles interact very powerfully with each other. And so the problem of nuclear forces, a frontier of physics starting in the 1930s, became the problem of understanding a vast new world of particles and reactions, the most powerful in Nature. Even the terminology changed. Physicists no longer refer to nuclear forces, but to the strong interaction.

Now we know that all the complexities of the strong interaction can be described, at a fundamental level, by a theory called quantum chromodynamics, or QCD, a vast generalization of QED. The elementary building blocks of QCD are quarks and gluons. There are six different kinds, or 'flavors', of quarks: u, d, s, c, b, t (up, down, strange, charm, bottom, top). The quarks are very similar to one another, differing mainly in their mass. Only the lightest ones, u and d, are found in ordinary matter. Making an analogy to the building blocks of QED, quarks play roughly the role of electrons, and gluons play roughly the role of photons. The big difference is that whereas in QED there is just one type of charge, and one photon, in QCD there are three types of charge, called colors, and eight gluons. Some gluons respond to color charges, similarly to the way photons respond to electric charge. Others mediate transitions between one color and another. Thus (say) a u quark with blue charge can radiate a gluon and turn into a u quark with green charge. Since all the charges overall must be con-

served, this particular gluon must have blue charge +1, green charge −1. Since gluons themselves carry unbalanced color charge, in QCD there are elementary processes where gluons radiate other gluons. There is nothing like this in QED. Photons are electrically neutral, and to a very good approximation they do not interact with other photons. Much of the richness and complexity of QCD arises because of this new feature.

Described thus baldly and verbally, without grounding in concepts or phenomena, QCD might seem both arbitrary and fantastic. In fact QCD is a theory of compelling symmetry and mathematical beauty. Unfortunately, I won't be able to do justice to those aspects here. But some brief explications are in order .···

How did we arrive at such a theory? And how do we know it's right? In the case of QCD, these are two very different questions. The historical path to its discovery was tortuous, with many false trails and blind alleys. But in retrospect, it didn't have to be that way. If the right kind of ultra-high-energy accelerators had come on line earlier, QCD would have stared us in the face[h]. This *gedanken*-history brings together most of the ideas I've discussed in this article, and forms a fitting conclusion to its physical part.

When electrons and positrons are accelerated to ultrahigh energy and then made to collide, two kinds of events are observed. In one kind of event the particles in the final states are leptons and photons. For this class of events, usually the final state is just a lepton and its anti-lepton; but in about 1% of the events there is also a photon, and in about 0.01% of the events there are also two photons. The probability for these sorts of events, and for the various particles to come out at various angles with different energies, can all be computed using QED, and it all works out very nicely. Conversely, if you hadn't known about QED, you could have figured out the basic rules for the fundamental interaction of QED – that is, the emission of a photon by an electron – just by studying these events. The fundamental interaction of light with matter is laid out right before your eyes.

In the other kind of event, you see something rather different. Instead of just two or at most a handful of particles coming out, there are many. And they are different kinds of particles. The particles you see in this second class of events are things like π mesons, K mesons, protons, neutrons, and their antiparticles – all particles that, unlike photons and leptons, have strong interactions. The angular distribution of these particles is very structured. They do not come out independently, every which way. Rather, they emerge

[h]Up to a couple of profound but well-posed and solvable problems, as I'll shortly discuss.

in just a few directions, making narrow sprays or (as they're usually called) "jets". About 90% of the time there are just two jets, in opposite directions; roughly 10% of the time there are three jets, 1% four jets – you can guess the pattern.

Now if you squint a little, and don't resolve the individual particles, but just follow the flow of energy and momentum, then the two kinds of events – the QED 'particle' events, and the 'jetty' events with strongly interacting particles – look just the same!

So (in this imaginary history) it would have been hard to resist the temptation to treat the jets as if they are particles, and propose rules for the likelihood of different radiation patterns, with different numbers, angles, and energies of the jet-particles, in direct analogy to the procedures that work for QED. And this would work out very nicely, because rules quite similar to those for QED actually do describe the observations. Of course, the rules that work are precisely those of QCD, including the new processes where glue radiates glue. All these rules – the foundational elements of the entire theory – could have been derived directly from the data. "Quarks" and "gluons" would be words with direct and precise operational definitions, in terms of jets.

Still, there would have been two big conceptual puzzles. Why do the experiments show 'quarks' and 'gluons' instead of just quarks and gluons – that is, jets, instead of just particles? And how do you connect the theoretical concepts that directly and successfully describe the high-energy events to all the other phenomena of the strong interaction? The connection between the supposedly foundational theory and the mundane observations is, to say the least, not obvious. For example, you would like to construct protons out of the 'quarks' and 'gluons' that appear in the fundamental theory. But this looks hopeless, since the jets in terms of which 'quarks' and 'gluons' are operationally defined often contain, among other things, protons.

There is an elegant solution to these problems. It is the phenomenon of *asymptotic freedom* in QCD. According to asymptotic freedom, radiation events that involve large changes in the flow of energy and momentum are rare, while radiation events that involve only small changes in energy and momentum are very common. Asymptotic freedom is not a separate assumption, but a deep mathematical consequence of the structure of QCD.

Asymptotic freedom neatly explains why there are jets in electron-positron annihilations at high energies, in the class of events containing strongly interacting particles. Immediately after the electron and positron annihilate, you have a quark and an antiquark emerging. They are mov-

ing rapidly, in opposite directions. They quickly radiate gluons, and the gluons themselves radiate, and a complicated cascade develops, with many particles. But despite all this commotion the overall flow of energy and momentum is not significantly disturbed. Radiations that disturb the flow of energy and momentum are rare, according to asymptotic freedom. So there is a large multiplicity of particles all moving in the same direction, the direction originally staked out by the quark or antiquark. In a word, we've produced a jet. When one of those rare radiations that disturbs the flow of energy and momentum takes place, the radiated gluon starts a jet of its own. Then we have a three-jet event. And so forth.

Asymptotic freedom also indicates why the description of protons (and the other strongly interacting particles) that we actually observe as individual stable, or quasi-stable, entities are complicated objects. For such particles are, more or less by definition, configurations of quarks, antiquarks, and gluons that have a reasonable degree of stability. But since the quarks, antiquarks, and gluons all have a very high probability for radiating, no simple configuration will have this property. The only possibility for stability involves dynamic equilibrium, in which the emission of radiation in one part of the system is balanced by its absorption somewhere else.

As things actually happened, asymptotic freedom was discovered theoretically (by David Gross and me, and independently by David Politzer) and QCD was proposed as the theory of the strong interaction (by Gross and me) in 1973, based on much less direct evidence. The existence of jets was anticipated, and their properties were *predicted* theoretically, in considerable detail, before their experimental observation. Based on these experiments, and many others, today QCD is accepted as the fundamental theory of the strong interaction, on a par with QED as the description of the electromagnetic interaction.

There has also been enormous progress in using QCD to describe the properties of protons, neutrons, and the other strongly interacting particles. This involves very demanding numerical work, using the most powerful computers, but the results are worth it. One highlight is that we can calculate from first principles, with no important free parameters, the masses of protons and neutrons. As I explained, from a fundamental point of view these particles are quite complicated dynamical equilibria of quarks, antiquarks, and gluons. Most of their mass – and therefore most of the mass of matter, including human brains and bodies – arises from the pure energy of these objects, themselves essentially massless, in motion, according to $m = E/c^2$. At this level, at least, we are ethereal creatures.

Dirac said that QED described "most of physics, and all of chemistry". Indeed, it is the fundamental theory of the outer structure of atoms (and much more). In the same sense, QCD is the fundamental theory of atomic nuclei (and much more). Together, they constitute a remarkably complete, well tested, fruitful and economical Theory of Matter.

12. The Fertility of Reason

I've now discussed in some detail how "playing with equations" led Dirac to an equation laden with consequences that he did not anticipate, and that in many ways he resisted, but that proved to be true and enormously fruitful. How could such a thing happen? Can mathematics be truly creative? Is it really possible, by logical processing or calculation, to arrive at essentially new insights – to get out more than you put in?

This question is especially timely today, since it lies at the heart of debates regarding the nature of machine intelligence – whether it may develop into a species of mind on a par with human intelligence, or even its eventual superior.

At first sight, the arguments against appear compelling.

Most powerful, at least psychologically, is the argument from introspection. Reflecting on our own thought processes, we can hardly avoid an unshakeable intuition that they do not consist exclusively, or even primarily, of rule-based symbol manipulation. It just doesn't feel that way. We normally think in images and emotions, not just symbols. And our streams of thought are constantly stimulated and redirected by interactions with the external world, and by internal drives, in ways that don't seem to resemble at all the unfolding of mathematical algorithms.

Another argument derives from our experience with modern digital computers. For these are, in a sense, ideal mathematicians. They follow precise rules (axioms) with a relentlessness, speed, and freedom from error that far surpasses what is possible for humans. And in many specialized, essentially mathematical tasks, such as arranging airline flight or oil delivery schedules to maximize profits, they far surpass human performance. Yet by common, reasonable standards even the most powerful modern computers remain fragile, limited, and just plain dopey. A trivial programming mistake, a few lines of virus code, or a memory flaw can bring a powerful machine to a halt, or send it into an orgy of self-destruction. Communication can take place only in a rigidly controlled format, supporting none of the richness of

natural language. Absurd output can, and often does, emerge uncensored and unremarked.

Upon closer scrutiny, however, these arguments raise questions and doubts. Although the nature of the map from patterns of electrical signals in nerve cells to processes of human thought remains deeply mysterious in many respects, quite a bit is known, especially about the early stages of sensory processing. Nothing that has been discovered so far suggests that anything more exotic than electric and chemical signalling, following well-established physical laws, is involved. The vast majority of scientists accept as a working hypothesis that a map from patterns of electric signals to thought must and does exist. The pattern of photons impinging on our retina is broken up and parsed out into elementary units, fed into a bewildering series of different channels, processed, and (somehow) reassembled to give us the deceptively simple "picture of the world", organized into objects in space, that we easily take for granted. The fact is we do not have the slightest idea how we accomplish most of what we do, even – perhaps especially – our most basic mental feats. People who've attempted to construct machines that can recognize objects appearing in pictures, or that can walk around and explore the world like a toddler, have had a very frustrating time, even though they can do these things very easily themselves. They can't teach others how they do these things because they don't know themselves. Thus it seems clear that introspection is an unreliable guide to the deep structure of thought, both as regards what is known and what is unknown.

Turning to experience with computers, any negative verdict is surely premature, since they are evolving rapidly. One recent benchmark is the victory of Deep Blue over the great world chess champion Garry Kasparov in a brief match. No one competent to judge would deny that play at this level would be judged a profoundly creative accomplishment, if it were performed by a human. Yet such success in a limited domain only sharpens the question: What is missing, that prevents the emergence of creativity from pure calculation over a broad front? In thinking about this tremendous question, I believe case studies can be of considerable value.

In modern physics, and perhaps in the whole of intellectual history, no episode better illustrates the profoundly creative nature of mathematical reasoning than the history of the Dirac equation. In hindsight, we know that what Dirac was trying to do is strictly impossible. The rules of quantum mechanics, as they were understood in 1928, cannot be made consistent

with special relativity. Yet from inconsistent assumptions Dirac was led to an equation that remains a cornerstone of physics to this day.

So here we are presented with a specific, significant, well-documented example of how mathematical reasoning about the physical world, culminating in a specific equation, led to results that came as a complete surprise to the thinker himself. Seemingly in defiance of some law of conservation, he got out much more than he put in. How was such a leap possible? Why did Dirac, in particular, achieve it? What drove Dirac and his contemporaries to persist in clinging to his equation, when it led them out to sea?[i]

Insights emerge from two of Dirac's own remarks. In his characteristically terse essay "My Life as a Physicist" he pays extended tribute to the value of his training as an engineer, including:

> The engineering course influenced me very strongly.··· I've learned that, in the description of nature, one has to tolerate approximations, and that even work with approximations can be interesting and can sometimes be beautiful.

Along this line, one source of Dirac's (and others') early faith in his equation, which allowed him to overlook its apparent flaws, was simply that he could find approximate solutions of it that agreed brilliantly with experimental data on the spectrum of hydrogen. In his earliest papers he was content to mention, without claiming to solve, the difficulty that there were other solutions, apparently equally valid mathematically, that had no reasonable physical interpretation.

Along what might superficially seem to be a very different line, Dirac often paid tribute to the heuristic power of mathematical beauty:

> The research worker, in his efforts to express the fundamental laws of Nature in mathematical form, should strive mainly for mathematical beauty.

This was another source of early faith in Dirac's equation. It was (and is) extraordinarily beautiful.

Unfortunately, it is difficult to make precise, and all but impossible to convey to a lay reader, the nature of mathematical beauty. But we can draw some analogies with other sorts of beauty. One feature that can make

[i]Much later, in the 1960s, Heisenberg recalled "Up till that time [1928] I had the impression that, in quantum theory, we had come into the harbor, into the port. Dirac's paper threw us out into the sea again."

a piece of music, a novel, or a play beautiful is the accumulation of tension between important, well-developed themes, which is then resolved in a surprising and convincing way. One feature that can make a work of architecture or sculpture beautiful is symmetry – balance of proportions, intricacy toward a purpose. The Dirac equation possesses both these features to the highest degree.

Recall that Dirac was working to reconcile the quantum mechanics of electrons with special relativity. It is quite beautiful to see how the tension between conflicting demands of simplicity and relativity can be harmonized, and to find that there is essentially only one way to do it. That is one aspect of the mathematical beauty of the Dirac equation. Another aspect, its symmetry and balance, is almost sensual. Space and time, energy and momentum, appear on an equal footing. The different terms in the system of equations must be choreographed to the music of relativity, and the pattern of 0s and 1s (and i s) dances before your eyes.

The lines converge when the needs of physics lead to mathematical beauty, or – in rare and magical moments – when the requirements of mathematics lead to physical truth. Dirac searched for a mathematical equation satisfying physically motivated hypotheses. He found that to do so he actually needed a system of equations, with four components. This was a surprise. Two components were most welcome, as they clearly represented the two possible directions of an electron's spin. But the extra doubling at first had no convincing physical interpretation. Indeed, it undermined the assumed meaning of the equation. Yet the equation had taken on a life of its own, transcending the ideas that gave birth to it, and before very long the two extra components were recognized to portend the spinning positron, as we saw.

With this convergence, I think, we reach the heart of Dirac's method in reaching the Dirac equation, which was likewise Maxwell's in reaching the Maxwell equations, and Einstein's in reaching both the special and the general theories of relativity. They proceed by *experimental logic*. That concept is an oxymoron only on the surface. In experimental logic, one formulates hypotheses in equations, and experiments with those equations. That is, one tries to improve the equations from the point of view of beauty and consistency, and then checks whether the "improved" equations elucidate some feature of Nature. Mathematicians recognize the technique of "proof by contradiction": To prove A, you assume the opposite of A, and reach a contradiction. Experimental logic is "validation by fruitfulness": To validate A, assume it, and show that it leads to fruitful consequences. Relative

to routine deductive logic, experimental logic abides by the Jesuit credo "It is more blessed to ask forgiveness than permission." Indeed, as we have seen, experimental logic does not regard inconsistency as an irremediable catastrophe. If a line of investigation has some success, and is fruitful, it should not be abandoned on account of its inconsistency, or its approximate nature. Rather, we should look for a way to make it true.

With all this in mind, let us return to the question of the creativity of mathematical reasoning. I said before that modern digital computers are, in a sense, ideal mathematicians. Within any reasonable, precisely axiomatized domain of mathematics, we know how to program a computer so it will systematically prove all the valid theorems[j]. A modern machine of this sort could churn through its program, and output valid theorems, much faster and more reliably than any human mathematician could. But running such a program to do advanced mathematics would be no better than setting the proverbial horde of monkeys to typing, hoping to reproduce Shakespeare. You'd get a lot of true theorems, but essentially all of them would be trivial, with the gems hopelessly buried amidst the rubbish. In practice, if you peruse journals of mathematics or mathematical physics, not to speak of literary magazines, you won't find much work submitted by computers. Attempts to teach computers to do "real" creative mathematics, like the attempts to teach them to recognize real objects or navigate the real world, have had very limited success. Now we begin to see that these are closely related problems. Creative mathematics and physics rely not on perfect logic, but rather on an experimental logic. Experimental logic involves noticing patterns, playing with them, making assumptions to explain them, and – especially – recognizing beauty. And creative physics requires more: abilities to sense and cherish patterns in the world, and to value not only logical consistency, but also (approximate!) fidelity to the world as observed.

So, returning to the central question: Can purely mathematical reasoning be creative? Undoubtedly, if it is used a la Dirac, in concert with the

[j]This is a consequence of Godel's completeness theorem for first-order predicate logic. Sophisticated readers may wonder how this result, that all valid theorems can be proved in mechanical fashion, can be consistent with Godel's famous incompleteness theorem. (It's not a misprint: Godel proved both completeness and incompleteness theorems.) To make a long story short, Godel's incompleteness theorem shows that in any rich mathematical system you will be able to formulate meaningful statements such that neither the statement nor its denial is a theorem. Such "incompleteness" does not contradict the possibility of systematically enumerating all the theorems.

abilities to tolerate approximations, to recognize beauty, and to learn by interacting with the real world. Each of these factors has played a role in all the great episodes of progress in physics. The question returns, as a challenge to ground those abilities in specific mechanisms.

Acknowledgments

My work is supported in part by funds provided by the U.S. Department of Energy (D.O.E.) under cooperative research agreement #DF-FC02-94ER40818. This presentation is adapted from my chapter "A Piece of Magic: The Dirac Equation" in the book *It Must Be Beautiful, The Great Equations of Modern Science*, ed. G. Farmelo (Granta Books, 2002).

References

1. For background material on atomic physics and quantum theory, including excerpts from important original sources, I highly recommend H. Boorse and L. Motz, *The World of the Atom* (Basic Books, 1966). Of course, some of its more "timely" parts appear somewhat dated today.
2. Dirac's classic is *The Principles of Quantum Mechanics* (Fourth Edition, Cambridge 1958).
3. A demanding but honest and beautiful treatment of the principles of quantum electrodynamics, with no mathematical prerequisites, is R.P. Feynman, *QED: The Strange Theory of Light and Matter* (Princeton. 1985).
4. For a brief account of QCD, easily accessible after Feynman's book, with no mathematical prerequisites, see F. Wilczek, "QCD Made Simple", *Physics Today*, **53N8** 22–28, (2000) . I'm at work on a full account, to be called simply *QCD* (Princeton).
5. For a conceptual review of quantum field theory, see my article "Quantum Field Theory" in the American Physical Society Centenary issue of *Rev. Mod. Phys.* **71**, S85–S95, (1999); this issue is also published as *More Things in Heaven and Earth – A Celebration of Physics at the Millennium*, B. Bederson, ed. (Springer-Verlag, New York), (1999) It contains several other reflective articles that touch on many of our themes.