# Contents

# 1

# Fashion

## 1.1. MATHEMATICAL ELEGANCE AS A DRIVING FORCE

As mentioned in the preface, the issues discussed in this book were developed from three lectures given, by invitation of the Princeton University Press, at Princeton University in October 2003. My nervousness, with these lectures, in addressing such a knowledgeable audience as the Princeton scientific community, was perhaps at its greatest when it came to the topic of fashion, because the illustrative area that I had elected to discuss, namely string theory and some of its various descendants, had been developed to its heights in Princeton probably more than anywhere else in the world. Moreover, that subject is a distinctly technical one, and I cannot claim competence over many of its important ingredients, my familiarity with these technicalities being somewhat limited, particularly in view of my status as an outsider. Yet, it seemed to me, I should not allow myself to be too daunted by this shortcoming, for if only the insiders are considered competent to make critical comments about the subject, then the criticisms are likely to be limited to relatively technical issues, some of the broader aspects of criticism being, no doubt, significantly neglected.

Since these lectures were given, there have been three highly critical accounts of string theory: *Not Even Wrong* by Peter Woit, *The Trouble with Physics* by Lee Smolin, and *Farewell to Reality: How Fairytale Physics Betrays the Search for Scientific Truth* by Jim Baggott. Certainly, Woit and Smolin have had more direct experience than I have of the string-theory community and its over-fashionable status. My own criticisms of string theory in *The Road to Reality*, in chapter 31 and parts of chapter 34, have also appeared in the meantime (predating these three works), but my own critical remarks were perhaps somewhat more favourably disposed towards a physical role for string theory than were these others. Most of my comments will indeed be of a general nature, and are relatively insensitive to issues of great technicality.

Let me first make what surely ought to be a general (and perhaps obvious) point. We take note of the fact that the hugely impressive progress that physical theory has indeed made over several centuries has depended upon extremely precise and sophisticated mathematical schemes. It is evident, therefore, that any further significant progress must again depend crucially upon some distinctive mathematical framework. In order that any proposed new physical theory can improve upon what has been achieved up until now, making precise and unambiguous predictions that go beyond what had been possible before, it must also be based on some clear-cut mathematical scheme. Moreover, one would think, to be a proper mathematical theory it surely ought to make mathematical sense – which means, in effect, that it ought to be *mathematically consistent*. From a self-*in*consistent scheme, one could, in principle, deduce any answer one pleased.

Yet, self-consistency is actually a rather strong criterion and it turns out that not many proposals for physical theories – even among the very successful ones of the past – are in fact fully self-consistent. Often some strong elements of physical judgement must be invoked in order that the theory can be appropriately applied in an unambiguous way. Experiments are, of course, also central to physical theory, and the testing of a theory by experiment is very different from checking it for logical consistency. Both are important, but in practice one often finds that physicists do not care so much about achieving full mathematical self-consistency if the theory appears to fit the physical facts. This has been the case, to some considerable degree, even with the extraordinarily successful theory of quantum mechanics, as we shall be seeing in chapter 2 (and §1.3). The very first work in that subject, namely Max Planck's epoch-making proposal to explain the frequency spectrum of electromagnetic radiation in equilibrium with matter at a fixed temperature (the black-body spectrum; see §§2.2 and 2.11) required something of a hybrid picture which was not really fully self-consistent [Pais 2005]. Nor can it be said that the old quantum theory of the atom, as brilliantly proposed by Niels Bohr in 1913, was a fully self-consistent scheme. In the subsequent developments of quantum theory, a mathematical edifice of great sophistication has been constructed, in which a desire for mathematical consistency had been a powerful driving force. Yet, there remain issues of consistency that are still not properly addressed in current theory, as we shall see later, particularly in §2.13. But it is the *experimental* support, over a vast range of different kinds of physical phenomena, which is quantum theory's bedrock. Physicists tend not to be over-worried by detailed matters of mathematical or ontological inconsistency if the theory, when applied with appropriate judgement and careful calculation, continues to provide answers that are in excellent agreement with the results of observation – often with extraordinary precision – through delicate and precise experiment.

The situation with string theory is completely different from this. Here there appear to be *no* results whatever that provide it with experimental support. It is often argued that this is not surprising, since string theory, as it is now formulated as largely a *quantum gravity* theory, is fundamentally concerned with what is called the *Planck scale* of very tiny distances (or at least close to such distances), some $10^{-15}$ or $10^{-16}$ times smaller ($10^{-16}$ meaning, of course, down by a factor of a tenth of a thousandth of a millionth of a millionth) and hence with energies some $10^{15}$ or $10^{16}$ times larger than those that are accessible to current experimentation. (It should be noted that, according to basic principles of relativity, a small distance is essentially equivalent to a small time, via the speed of light, and, according to basic principles of quantum mechanics, a small time is essentially equivalent to a large energy, via Planck's constant; see §§2.2 and 2.11.) One must certainly face the evident fact that, powerful as our present-day particle accelerators may be, their currently foreseeable achievable energies fall enormously short of those that appear to have direct relevance to theories such as modern string theory that attempt to apply the principles of quantum mechanics to gravitational phenomena. Yet this situation can hardly be regarded as satisfactory for a physical theory, as experimental support is the ultimate criterion whereby it stands or falls.

Of course, it might be the case that we are entering a new phase of basic research into fundamental physics, where requirements of mathematical consistency become paramount, and in those situations where such requirements (together with a coherence with previously established principles) prove insufficient, additional criteria of *mathematical elegance* and simplicity must be invoked. While it may seem unscientific to appeal to such aesthetic desiderata in a fully objective search for the physical principles underlying the workings of the universe, it is remarkable how fruitful – indeed essential – such aesthetic judgements seem to have frequently proved to be. We have come across many examples in physics where beautiful mathematical ideas have turned out to underlie fundamental advances in understanding. The great theoretical physicist Paul Dirac [1963] was very explicit about the importance of aesthetic judgement in his discovery of the equation for the electron, and also in his prediction of anti-particles. Certainly, the Dirac equation has turned out to be absolutely fundamental to basic physics, and the aesthetic appeal of this equation is very widely appreciated. This is also the case with the idea of anti-particles, which resulted from Dirac's deep analysis of his own equation for the electron.

However, this role of aesthetic judgement is a very difficult issue to be objective about. It is often the case that some physicist might think that a particular scheme is very beautiful whereas another might emphatically *not* share that view! Elements of fashion can often assume unreasonable proportions when it comes to aesthetic judgements – in the world of theoretical physics, just as in the case of art or the design of clothing.

It should be made clear that the question of aesthetic judgment in physics is more subtle than just what is often referred to as *Occam's razor* – the removal of unnecessary complication. Indeed, a judgement as to which of two opposing theories is actually the "simpler", and perhaps therefore more elegant, need by no means be a straightforward matter. For example, is Einstein's general relativity a simple theory or not? Is it simpler or more complicated than Newton's theory of gravity? Or is Einstein's theory simpler or more complicated than a theory, put forward in 1894 by Aspeth Hall (some 21 years before Einstein proposed his general theory of relativity), which is just like Newton's but where the inverse square law of gravitation is replaced by one in which the gravitational force between a mass $M$ and a mass $m$ is $GmMr^{-2.00000016}$, rather than Newton's $GmMr^{-2}$. Hall's theory was proposed in order to explain the observed slight deviation from the predictions of Newton's theory with regard to the advance of the perihelion of the planet Mercury that had been known since about 1843. (The perihelion is the closest point to the Sun that a planet reaches while tracing its orbit [Roseveare 1982].) This theory also gave a very slightly better agreement with Venus's motion than did Newton's. In a certain sense, Hall's theory is only marginally more complicated than Newton's, although it depends on how much additional "complication" one considers to be involved in replacing the nice simple number "2" by "2.00000016". Undoubtedly, there is a loss of mathematical elegance in this replacement, but as noted above, a strong element of subjectivity comes into such judgements. Perhaps more to the point is that there are certain elegant mathematical properties that follow from the inverse square law (basically, expressing a conservation of "flux lines" of gravitational force, which would not be exactly true in Hall's theory). But again, one might consider this an aesthetic matter whose physical significance should not be overrated.

But what about Einstein's general relativity? There is certainly an enormous increase in the difficulty of applying Einstein's theory to specific physical systems, beyond the difficulty of applying Newton's theory (or even Hall's), when it comes to examining the implications of this theory in detail. The equations, when written out explicitly, are immensely more complicated in Einstein's theory, and they are difficult even to write down in full detail. Moreover, they are immensely harder to solve, and there are many nonlinearities in Einstein's theory which do not appear in Newton's (these tending to invalidate the simple flux-law arguments that must already be abandoned in Hall's theory). (See §§A.4 and A.11 for the meaning of *linearity*, and for its special role in quantum mechanics see §2.4.) Even more serious is the fact that the physical interpretation of Einstein's theory depends upon eliminating spurious coordinate effects that arise from the making of particular choices of coordinates, such choices being supposed to have no physical relevance in Einstein's theory. In practical terms, there is no doubt that Einstein's theory is usually immensely more difficult to handle than is Newton's (or even Hall's) gravitational theory.

Yet, there is still an important sense in which Einstein's theory is actually a very simple one – even possibly simpler (or more "natural") than Newton's. Einstein's theory depends upon the mathematical theory of Riemannian (or, more strictly, as we shall be seeing in §1.7, *pseudo*-Riemannian) geometry, of arbitrarily curved 4-manifolds (see also §A.5). This is not an altogether easy body of mathematical technique to master, for we need to understand what a tensor is and what the purpose of such quantities is, and how to construct the particular tensor object **R**, called the *Riemann curvature tensor*, from the *metric tensor* **g** which defines the geometry. Then by means of a contraction and a trace-reversal we find how to construct the *Einstein tensor* **G**. Nevertheless, the general geometrical ideas behind the formalism are reasonably simple to grasp, and once the ingredients of this type of curved geometry are indeed understood, one finds that there is a very restricted family of possible (or plausible) equations that can be written down, which are consistent with the proposed general physical and geometrical requirements. Among these possibilities, the very simplest gives us Einstein's famous field equation $\mathbf{G} = 8\pi\gamma\mathbf{T}$ of general relativity (where **T** is the mass–energy tensor of matter and $\gamma$ is Newton's gravitational constant – given according to Newton's particular definition, so that even the "$8\pi$" is not really a complication, but merely a matter of how we wish to define $\gamma$).

There is just one minor, and still very simple, modification of the Einstein field equation that can be made, which leaves the essential requirements of the scheme intact, namely the inclusion of a constant number $\Lambda$, referred to as the *cosmological constant* (which Einstein introduced in 1917 for reasons that he later discarded) so that Einstein's equations with $\Lambda$ now become $\mathbf{G} = 8\pi\gamma\mathbf{T} + \Lambda\mathbf{g}$. The quantity $\Lambda$ is now frequently referred to as *dark energy*, presumably to allow for a possibility of generalizing Einstein's theory so that $\Lambda$ might vary. There are, however, strong mathematical constraints obstructing such considerations, and in §§3.1, 3.7, 3.8, and 4.3, where $\Lambda$ will be playing a significant role for us, I shall restrict attention to situations where $\Lambda$ is indeed non-varying. The cosmological constant will have considerable relevance in chapter 3 (and also §1.15). Indeed, relatively recent observations point strongly to the actual physical presence of $\Lambda$ having a tiny (apparently constant) positive value. This evidence for $\Lambda > 0$ – or possibly for some more general form of "dark energy" – is now very impressive, and has been growing since the initial observations of Perlmutter et al. [1999], Riess et al. [1998], and their collaborators, leading to the award of the 2011 Nobel Prize in physics to Saul Perlmutter, Brian P. Schmidt, and Adam G. Riess. This $\Lambda > 0$ has immediate relevance only to the very distant cosmological scales, and observations concerning celestial motions at a more local scale can be adequately treated according to Einstein's original and simpler $\mathbf{G} = 8\pi\gamma\mathbf{T}$. This equation is now found to have an unprecedented precision in modelling the behaviour, under gravity, of celestial bodies, the observed $\Lambda$ value having no significant impact on such local dynamics.

Historically of most importance, in this regard, is the double-neutron-star system PSR1913+16, one component of which is a *pulsar*, sending very precisely timed electromagnetic signals that are received at the Earth. The motion of each star about the other, being very cleanly a purely gravitational effect, is modelled by general relativity to an extraordinary precision that can be argued to be of about one part in $10^{14}$ overall, accumulated over a period of about 40 years. The period 40 years is roughly $10^9$ seconds, so a precision of one in $10^{14}$ means an agreement between observation and theory to about $10^{-5}$ (one hundred thousandth) of a second over that period – which is, very remarkably, indeed just what is found. More recently, other systems [Kramer et al. 2006] involving one or even a pair of pulsars, have the potential to increase this precision considerably, when the systems have been observed for a comparable length of time as has PSR19+16.

To call this figure of $10^{14}$ a measure of the observed precision of general relativity is open to some question, however. Indeed, the particular masses and orbital parameters have to be calculated from the observed motions, rather than being numbers coming from theory or independent observation. Moreover, much of this extraordinary precision is already in Newton's gravitational theory.

Yet, we are concerned here with gravitational theory overall, and Einstein's theory incorporates the deductions from Newton's theory (giving Kepler's elliptical orbits, etc.) as a first approximation, but provides various corrections to the Keplerian orbits (including the perihelion advance), and finally a loss of energy from the system which is precisely in accord with a remarkable prediction from general relativity: that such a massive system in accelerated motion should lose energy through the emission of gravitational waves – ripples in space-time which are the gravitational analogues of electromagnetic waves (i.e. light) that electrically charged bodies emit when they are involved in accelerated motion. As a striking further confirmation of the existence and precise form of such gravitational radiation is the announcement [Abbott et al. 2016] of their direct detection by the LIGO gravitational wave detector, which also provides excellent direct evidence of another of the predictions of general relativity: the existence of black holes, which we shall be coming to in §3.2, and discussed also in later parts of chapter 3, and in §4.3.

It should be emphasized that this precision goes enormously beyond – by an additional factor of about $10^8$ (i.e. one hundred million) or more – that which was observationally available to Einstein when he first formulated his gravitational theory. The observed precision in Newton's gravitational theory could itself be argued to be around one part in $10^7$. Accordingly, the "1 part in $10^{14}$" precision of general relativity was already "out there" in nature, before Einstein formulated his own theory. Yet that additional precision (by a factor of around one hundred million), being unknown to Einstein, can have played no role whatever in his formulating his theory. Thus this new mathematical model of nature was not a man-made construction invented merely in an attempt to find the best theory to fit the facts; the mathematical scheme was, in a clear sense, already there in the works of nature herself. This mathematical simplicity, or elegance, or however one should describe it, is a genuine part of nature's ways, and it is not simply that our minds are attuned to being impressed by such mathematical beauty.

On the other hand, when we try deliberately to use the criterion of mathematical beauty in formulating our theories, we are easily led astray. General relativity is certainly a very beautiful theory, but how does one judge the elegance of physical theories generally? Different people have very different aesthetic judgements. It is not necessarily obvious that one person's view as to what is elegant will be the same as somebody else's, or whether one person's aesthetic judgement will be superior or inferior to another's, in formulating a successful physical theory. Moreover, the inherent beauty in a theory is often not obvious at first, and may be revealed only later when the depths of its mathematical structure become apparent through later technical developments. Newtonian dynamics is a case in point. Much of the undoubted beauty in Newton's framework was revealed only much later, through the magnificent works of such great mathematicians as Euler, Lagrange, Laplace, and Hamilton (as the terms *Euler–Lagrange equations*, the *Laplacian operator*, *Lagrangians*, and *Hamiltonians* – which are key ingredients of modern physical theory – bear witness). The role of Newton's Third Law, for example, which asserts that every action has an equal and opposite reaction, finds a central place in the Lagrangian formulation of modern physics. It would not surprise me to find that the beauty that is frequently asserted to be present in successful modern theories is often to some extent *post hoc*. The very success of a physical theory, both observational and mathematical, may contribute significantly to the aesthetic qualities that it is later perceived to possess. It would follow from all this that judgements of the merits of some proposed physical theory through its claimed aesthetic qualities are likely to be problematic or at least ambiguous. It is unquestionably more reliable to form one's judgements of a new theory on the basis of its agreement with current observation and on its predictive power.

Yet, with regard to experimental support, often the crucial experiments are not available, such as with the utterly prohibitive high energies that single particles might have to attain – absurdly in excess of those available in current particle accelerators (see §1.10) – that are often argued to be required in any proper observational test of any quantum-gravity theory. More modest experimental proposals may also be unavailable, due perhaps to the expense of the experiments or their intrinsic difficulty. Even with very successful experiments, it is quite often the case that the experimenters collect enormous amounts of data, and the problem is of a quite different kind, namely the matter of digging out some key piece of information from that morass of data. This kind of thing is certainly true in particle physics, where powerful accelerators and particle colliders now produce masses of information, and it is now also becoming true in cosmology, where modern observations of the cosmic microwave background (CMB) produce very large amounts of data (see §§3.4, 3.9, and 4.3). Much of this data is considered not to be especially informative, as it simply confirms what is already known, as gleaned from earlier experiments. A great deal of statistical processing is needed in order to extract some tiny residual – which is the new feature that the experimentalists are looking for – which might confirm or refute some suggested theoretical proposal.

A point that should be made here is that this statistical processing is likely to be very specific to current theory, geared to finding out what slight additional effect that theory might predict. It is very possible that some radically different set of ideas, departing significantly from what is currently fashionable, may remain untested even though some definitive answer might actually lie hidden in the existing data, being unrevealed because the statistical procedures that physicists have adopted are too directly tuned to current theory. We shall be seeing what appears to be a striking example of this in §4.3. Even when it is clear how definitive information might be statistically extracted from an existing morass of reliable data, the inordinate amount of computer time that this can require may sometimes constitute a huge barrier to the actual carrying out of the analysis, particularly when more fashionable pursuits may be in direct competition.

Even more to the point is the fact that the experiments themselves are usually enormously expensive and their specific design is likely to be geared to the testing of theories which are within the framework of conventional ideas. Any theoretical scheme which departs too radically from the general consensus may find it hard for sufficient funds to be provided to enable it to be properly tested. A very expensive experimental apparatus, after all, requires many committees of established experts to approve its construction, and such experts are likely to be those who have played their parts in developing the current perspectives.

In relation to this issue we may consider the Large Hadron Collider (LHC) in Geneva, Switzerland, whose construction was completed in 2008. It has a 27 km (17 mile) tunnel running under two countries (France and Switzerland), initially coming into action in 2010. It is now credited with finding the hitherto elusive Higgs particle, of great importance in particle physics, particularly in relation to its role in assigning mass to weakly interacting particles. The 2013 Nobel Prize in physics was awarded to Peter Higgs and François Englert for their part in the ground-breaking work of predicting the existence and properties of this particle.

This is undoubtedly a magnificent achievement, and I have no wish to underrate its undoubted importance. Nevertheless, the LHC appears to provide a case in point. The way in which the very high-energy encounters between particles are analysed requires the presence of extremely expensive detectors, which have been geared to glean information in relation to prevailing particle-physics theory. It may not be at all easy to obtain information of relevance to unconventional ideas concerning the underlying nature of fundamental particles and their interactions. In general, proposals which depart drastically from a prevailing perspective may well find it much harder to have a chance of being adequately funded, and also may find great difficulty in being tested at all, by definitive experiments.

A further important factor is that graduate students, when in search of a problem to work on for a doctorate degree, tend to be highly constrained with regard to appropriate topics of research. Research students working in unfashionable areas, even if leading to successful doctoral degrees, may well find extreme difficulty in obtaining academic jobs afterwards, no matter how talented, knowledgeable, or original they may be. Jobs are limited and research funding hard to come by. Research supervisors are, more likely than not, interested mainly in developing ideas that they themselves had been involved in promoting, and these are likely to be in areas that are already fashionable. Moreover, a supervisor interested in developing an idea that is outside the mainstream may well be reluctant to encourage a potential student to work in such an area, owing to the disadvantage that it may be to the student when it comes to competing, subsequently, in a highly competitive job market where those with an expertise in fashionable areas will have a distinct advantage.

The same issues arise when it comes to the funding of research projects. Proposals in fashionable areas are far more likely to receive approval (see also §1.12). Again, the proposals will be judged by acknowledged experts, and those are overwhelmingly likely to be working in areas that are already fashionable, and to which they themselves may well have been significant contributors. Projects that deviate too much from the currently accepted norms, even if well thought through and highly original, are very likely to be left without support. Moreover, this is not just a matter of limitations on the funds available, as the influence of fashion appears to be particularly relevant in the United States, where the availability of funds for scientific research remains relatively high.

It must be said, of course, that most unfashionable areas of research will be considerably less likely to develop into successful theories than any of those that are already fashionable. A radical new perspective will in the vast majority of cases have little chance of developing into a viable proposal. Needless to say, as with Einstein's general relativity, any such radical perspective must already agree with what has been previously experimentally established, and if not, then an expensive experimental test may well not be needed for the rejection of inappropriate ideas. But for theoretical proposals that are in agreement with all previously performed experiments, and where there is no current prospect for experimental confirmation or refutation – perhaps for reasons such as those just described – it seems that we must fall back on mathematical consistency, general applicability, and aesthetic criteria when we form our judgements of the plausibility and relevance of some proposed physical theory. It is in such circumstances that the role of fashion may begin to attain excessive proportions, so we must be very careful not to allow the fashionable nature of some particular theory to cloud our judgements as to its actual physical plausibility.

## 1.2. SOME FASHIONABLE PHYSICS OF THE PAST

This is particularly important for theories which purport to be probing the very foundations of physical reality, such as modern-day string theory, and we must be very wary of assigning too much plausibility to such a theory on account of

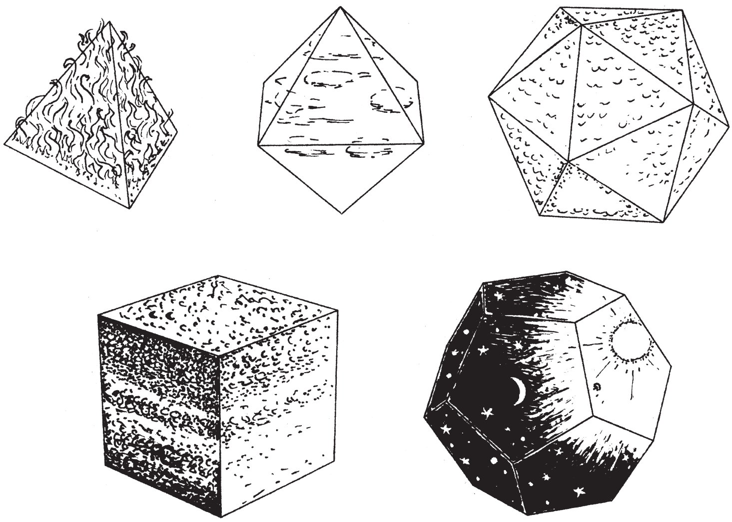

Figure 1-1: The five elements of Ancient Greece: fire (tetrahedron), air (octahedron), water (icosahedron), earth (cube), and aether (dodecahedron).

its fashionable status. Before addressing present-day physical ideas, however, it will be instructive to mention some fashionable scientific theories of the past that we do not take seriously today. There are large numbers of them and I am sure that a good many readers will have little knowledge of most, for the sufficient reason that if we do not now take these theories seriously, we are fairly unlikely to learn about them – unless, of course, we are good historians of science; but most physicists are not. At least let me mention a few of the better-known ones.

In particular, there is the ancient Greek theory that the Platonic solids are to be associated with what they regarded as the basic elements of material substance, as represented in figure 1-1. Here, *fire* is represented as the regular tetrahedron, *air* as the octahedron, *water* as the icosahedron, and *earth* as the cube, and where, in addition, the celestial *aether* (or *firmament*, or *quintessence*) was later introduced, of which it was supposed that celestial bodies were composed, and was taken to be represented as the regular dodecahedron. The ancient Greeks appear to have formulated this sort of view – or at least many did – and I suppose it could well have qualified as a fashionable theory of the time.

Initially, they just had the four elements of fire, air, water, and earth, and this collection of primitive entities seemed to accord well with the four perfectly regular polyhedral shapes that were known at the time, but when the dodecahedron was later discovered, the theory needed to be extended to accommodate this additional polyhedron! Accordingly, the celestial substance which composed such supposedly perfect bodies as the Sun, the Moon and planets, and the crystal spheres to which they were supposedly attached was brought into the polyhedral scheme – this substance appearing, to the Greeks, to satisfy very different laws from the ones found on Earth, having a seemingly eternal motion, rather than the universal tendency of familiar substances to slow down and stop. Perhaps there is some lesson here about the way in which even modern sophisticated theories, having been initially presented in a supposedly definitive form, may then become significantly altered and their original doctrines stretched to lengths not previously conceived, in the face of new theoretical or observational evidence. As I understand it, the view that the ancient Greeks held was that somehow the laws governing the motions of stars, planets, Moon, and Sun were indeed quite different from the laws that governed things on Earth. It took Galileo, with his understanding of the relativity of motion, and then Newton, with his theory of universal gravitation – strongly influenced also by Kepler's understanding of planetary orbiting – to appreciate that the same laws actually apply to celestial bodies as to those on Earth.

When I first became acquainted with these ancient Greek ideas, they struck me as sheer romantic fantasy, with no mathematical (let alone physical) rationale. But more recently I learned that there is a bit more of a theory underlying these ideas than I had initially imagined. Some of these polyhedral shapes can be cut up into pieces, and then suitably recombined to make others (as, for example, two cubes can be cut to make two tetrahedra and an octahedron). This might be related to physical behaviour and used as a geometrical model for the basis of transitions which can occur between these different elements. At least here was a bold and imaginative guess as to the nature of material substance, which was not really an unreasonable suggestion at a time when so little had been established as to the actual nature and behaviour of physical material. Here was an early attempt to find a basis for real materials in terms of an elegant mathematical structure – very much in the spirit of what theoretical physicists are still striving to do today – in which theoretical consequences of the model could be tested against actual physical behaviour. Aesthetic criteria were clearly also at work here, and the ideas certainly seem to have appealed to Plato. But, needless to say, the details of the ideas cannot have stood the tests of time too well – otherwise we should surely not have abandoned such a mathematically attractive proposal!

Let us consider a few other things. The Ptolemaic model of planetary motion – in which the Earth was taken to be fixed, located at the centre of the cosmos

– was extremely successful, and remained unchallenged for many centuries. The motions of Sun, Moon, and planets were to be understood in terms of *epicycles*, according to which planetary movements could be explained in terms of the superposition of uniform circular motions upon one another. Though the scheme had to be rather complicated, in order to get good agreement with observation, it was not altogether lacking in mathematical elegance, and it was able to provide a reasonably predictive theory of the future motions of planets. It should be mentioned that epicycles do have a genuine rationale, when one considers external motions relative to a stationary Earth. The motions that we actually directly see from the Earth's perspective involve the composition of the Earth's rotation (so there is a perceived circular motion of the heavens about the Earth's polar axis), which must be composed with the general apparent motions of the Sun, Moon, and planets that are roughly constrained to the ecliptic plane, which appears to us as a fairly closely circular motion about a *different* axis. For good geometrical reasons, we already perceive something of the general nature of epicycles – circular motions upon other circular motions – so it was not so unreasonable to suppose that this idea might be extended more generally in the more detailed motions of the planets.

Moreover, epicycles themselves provide some interesting geometry, and Ptolemy was himself a fine geometer. In his astronomical work, he employed an elegant and powerful geometrical theorem that he possibly discovered, as it now bears his name. (This theorem asserts that the condition for four points in a plane A, B, C, D to lie on a circle – taken in that cyclic order – is that the distances between them satisfy $AB \cdot CD + BC \cdot DA = AC \cdot BD$.) This was the accepted theory of planetary motion for around fourteen centuries, until it was superseded, and eventually completely overturned, through the wonderful work of Copernicus, Galileo, Kepler, and Newton, and it is now regarded as thoroughly incorrect! It must certainly be described as a fashionable theory, however, and it was an extraordinarily successful one, for around fourteen centuries (from the mid second to the mid sixteenth), fairly closely accounting for all observations of planetary motion (when appropriate improvements were introduced from time to time), up until the more precise measurements of Tycho Brahe towards the end of the sixteenth century.

Another famous theory that we do not now believe, though it was very fashionable for over a century between 1667 (when it was put forward by Joshua Becher) and 1778 (when effectively disproved by Antoine Lavoisier), was the phlogiston theory of combustion. According to this theory, any inflammable substance contained an element called *phlogiston*, and the process of burning involved that substance giving up its phlogiston into the atmosphere. The phlogiston theory accounted for most of the facts about burning that were known at the time, such as the fact that when burning took place in a reasonably small sealed container, it would tend to stop before all of the inflammable material was used up, this being explained by the air in the container becoming saturated with phlogiston and unable to absorb any more. Ironically, it was Lavoisier who was responsible for another fashionable but false theory, namely that heat is a material substance, which he referred to as *caloric*. That theory was disproved in 1798 by Count Rumford (Sir Benjamin Thompson).

In each of these two main examples, the success of the theory may be understood by its close relation to the more satisfactory scheme which superseded it. In the case of Ptolemaic dynamics, we can transform to the more satisfactory heliocentric picture of Copernicus by a simple geometric transformation. This involves referring motions to the Sun as centre, rather than the Earth. At first, when everything was described in terms of epicycles, this made little difference – except that the heliocentric picture looked much more systematic, with the more rapid planetary motions being those for planets that were closer to the Sun [Gingerich 2004; Sobel 2011] – and there was a basic equivalence between the two schemes at this stage. But, when Kepler found his three laws of *elliptic* planetary motion, the situation changed completely, since a geocentric description of this kind of motion made no good geometrical sense. Kepler's laws provided the key that opened the way to the extraordinarily precise and broad-ranging Newtonian picture of *universal gravity*. Nevertheless, we might not today regard the geocentric perspective as quite so outrageous as would have been the case in the nineteenth century, in the light of the *general covariance principle* of Einstein's general relativity (see §§1.7, A.5, and 2.13), which allows us to adopt massively inconvenient coordinate descriptions (like a geocentric one in which the Earth's coordinates do not change with time) as nevertheless legitimate. Likewise, the phlogiston theory could be made to correspond closely to the modern perspective on combustion in which the burning of some material is normally taken to involve the taking up of oxygen from the atmosphere, where phlogiston would simply be regarded as "negative oxygen". This provides us with a fairly consistent translation between the phlogiston picture and the now conventional one. But when detailed mass measurements by Lavoisier demonstrated that phlogiston would have to have negative mass, the picture began to lose support. Nevertheless, "negative oxygen" is not such an absurd concept from the perspective of modern particle physics, where every type particle in nature (including a composite one) is supposed to have an anti-particle – an "anti-oxygen atom" is therefore completely in accord with modern theory. It would not, however, have a negative mass!

Sometimes theories that have been out of fashion for some while can come back into consideration in view of later developments. A case in point is an idea that Lord Kelvin (William Thompson) put forward in about 1867, in which atoms (the elementary particles of his day) were to be regarded as being composed of tiny knot-like structures. This idea attracted some considerable attention at the time, and the mathematician J. G. Tait began a systematic study of knots on the basis of this. But the theory did not lead to any clear-cut correspondence with the actual physical behaviour of atoms, so it became largely forgotten. However, more recently, ideas of this general kind have begun to find favour again, partly in view of their connection with string-theoretic notions. The mathematical theory of knots has also encountered a revival, since around 1984, starting with the work of Vaughan Jones, whose seminal ideas had their roots in theoretical considerations within quantum field theory [Jones 1985; Skyrme 1961]. The methods of string theory were subsequently employed by Edward Witten [1989] to obtain a kind of quantum field theory (called a *topological quantum field theory*) which, in a certain sense, encompasses these new developments in the mathematical theory of knots.

As a revival of a far more ancient idea for the nature of the large-scale universe, I might mention – though not entirely seriously – a curious coincidence that occurred at about the time that I was presenting my Princeton lecture on which this particular chapter is based (on 17 October 2003). In that talk, I referred to the ancient Greek idea that the aether was to be associated with the regular dodecahedron. Unbeknown to me, at that time, there were newspaper reports of a proposal, due to Luminet et al. [2003], that the 3-dimensional spatial geometry of the cosmos might actually have a somewhat complicated topology, arising from the identification (with a twist) of opposite faces of a (solid) *regular dodecahedron*. Thus, in a sense, the Platonic idea of a dodecahedral cosmos was also being revived in modern times!

The ambitious idea of a theory of everything, intended to encompass all physical processes, including a description of all the particles of nature and their physical interactions, has been commonly mooted in recent years, especially in connection with string theory. The idea would be to have a complete theory of physical behaviour, based on some notion of primitive particles and/or fields, acting according to some forces or other dynamical principles precisely governing the motions of all constituent elements. This may also be regarded as a revival of an old idea, as we shall see in a moment.

Just as Einstein was producing the final form of his general theory of relativity, towards the end of 1915, the mathematician David Hilbert put forward his own method of deriving the field equations of Einstein's theory,[1] using what is known as a *variational principle*. (This very general type of procedure makes use of the Euler–Lagrange equations, obtained from a Lagrangian, this being a powerful notion, referred to by name in §1.1; see, for example, Penrose [2004, chapter 20]; henceforth this book will be referred to as "TRtR".) Einstein, in his own more direct approach, formulated his equations explicitly in a form which showed how the gravitational field (as described in terms of space-time curvature) would behave, as influenced by its "source", namely the total mass/energy densities of all the particles, or matter fields, etc., collected together in the form of the energy tensor **T** (referred to in §1.1).

Einstein gave no specific prescription for the detailed equations governing how these matter fields were to behave, these being supposed to be taken from some other theory specific to the particular matter fields under consideration. In particular, one such matter field would be the electromagnetic field, whose description would be given according to the wonderful equations of the great Scottish mathematical physicist James Clerk Maxwell in 1864, which fully unified electric and magnetic fields, thereby explaining the nature of light and much of the nature of the forces governing the internal constitution of ordinary materials. This was to be considered matter in this context, and to play its appropriate part in **T**. In addition, other types of field, and all sorts of other kinds of particles could also be involved, being governed by whatever equations as might turn out to be appropriate, would also count as matter and contribute to **T**. The details of this were not important to Einstein's theory, and were left unspecified.

On the other hand, in his own proposal Hilbert was attempting to be more all-embracing. For he put forward what we might now refer to as a *theory of everything*. The gravitational field was to be described in just the same way as in Einstein's proposal, but rather than leaving the source term **T** unspecified, as Einstein had done, Hilbert proposed that this source term should be that of a very specific theory that was fashionable at the time, known as *Mie's theory* [Mie 1908, 1912a,b, 1913]. This involved a nonlinear modification of Maxwell's electromagnetic theory, and it had been proposed by Gustav Mie as a scheme intended to incorporate *all* aspects of matter. Accordingly, Hilbert's all-embracing proposal was supposed to be a complete theory of matter (including electromagnetism) as well as gravity. The strong and weak forces of particle physics were not understood at the time, but Hilbert's proposal could indeed have been viewed as what

[1] On the controversial issue of who was first, see Corry et al.'s [1997] commentary.

we now frequently refer to as a theory of everything. Yet, I think it likely that not a great many physicists today will have even heard of the once fashionable Mie's theory, let alone the fact that it was explicitly part of Hilbert's theory-of-everything version of general relativity. That theory plays no part in the modern understanding of matter. Perhaps there is a lesson of caution here for theoreticians of today, intent on proposing their own theories of everything.

## 1.3. PARTICLE-PHYSICS BACKGROUND TO STRING THEORY

One such theoretical proposal is string theory, and many theoretical physicists today do indeed still regard this proposal as providing a definite route to such a theory of everything. String theory originated with some ideas which, when I first heard about them in around 1970 (from Leonard Susskind), I found to be strikingly attractive and of a distinctive compelling nature. But before describing these ideas, I should put them in the appropriate context. We should try to understand why replacing the notion of a point particle by a little loop or curve in space, as was indeed the original idea of string theory, should have any promise as the basis for a physical picture of reality.

In fact there were more reasons than one for the attraction of this idea. Ironically, one of the most specific reasons – having to do with the observational physics of the interactions between hadrons – seems to have become completely left behind by the more modern developments in string theory, and I'm not sure that it has any status in the subject at all now, beyond a historical one. But I ought to discuss it, nevertheless (as I shall, more particularly, in §1.6), as well as some of the other elements of the background of fundamental particle physics which motivated the underlying principles of string theory.

First let me say what a hadron is. We recall that an ordinary atom consists of a positively charged nucleus, and negatively charged electrons orbiting around it. The nucleus is composed of protons and neutrons – collectively called *nucleons* (N) – where each proton has a positive electric charge of one unit (the unit of charge being chosen so that the electron's charge is the negative of one unit) and where each neutron has zero electric charge. The attractive electric force between positive and negative charges is what holds the negatively charged electrons in their orbits around the positively charged nucleus. But if electric forces were the only ones of relevance, then the nucleus itself (apart from that of hydrogen, which has just a single proton) would explode into various constituents, because the protons, all having charges of the same sign, would repel one another. Accordingly, there must be another, stronger, force which holds the nucleus together, and this is what is called the *strong* (nuclear) force. There is, in addition, something called the *weak* (nuclear) force, which has particular relevance in relation to nuclear decay, but this is not the major component in the forces between nucleons. I shall be saying something about the weak force later.

Not all particles are directly affected by the strong force – for example, electrons are not – but those which are so affected are the comparatively massive particles called *hadrons* (from the Greek *hadros*, meaning *bulky*). Accordingly, protons and neutrons are examples of hadrons, but there are many other kinds of hadron now known to exist. Among these others are the cousins of protons and neutrons called *baryons* (from *barys*, meaning *heavy*), which in addition to neutrons and protons themselves include the lambda ($\Lambda$), sigma ($\Sigma$), xi ($\Xi$), delta ($\Delta$), and omega ($\Omega$), most of which come in different versions with different values for the electric charge, and also in a sequence of excited (more rapidly spinning) versions. All these other particles are more massive than the proton and neutron. The reason that we do not find these more exotic particles as parts of ordinary atoms is that they are highly unstable and rapidly decay, ultimately into protons or neutrons, giving up their excessive mass in the form of energy (in accordance with Einstein's famous $E = mc^2$). The proton, in turn, has the mass of about 1836 electrons, and the neutron of around 1839 electrons. Intermediate between the baryons and electrons is another class of hadrons, called *mesons*, the most familiar ones being the pion ($\mu$) and the kayon (K). Each of these comes in a charged version ($\mu^+$ and $\mu^-$, each with a mass of about 273 electrons; $K^+$ and $\bar{K}^-$, each with a mass of about 966 electrons) and an uncharged version ($\mu^0$ has a mass of about 264 electrons; $K^0$ and $\bar{K}^0$, each with a mass of about 974 electrons). The practice here is to use a bar over the particle symbol to denote the anti-particle; we note, however, that the anti-pions are again pions, whereas an anti-kayon differs from a kayon. Again, these particles have many cousins and excited (more highly spinning) versions.

You can begin to see that all this is very complicated – a far cry from the heady days of the early twentieth century when the proton, neutron, and electron (and one or two massless ones such as the photon, the particle of light) had seemed to represent, more or less, the sum total of it all. As the years rolled by, things got more and more complicated, until eventually a unified picture of it all – called the *standard model of particle physics* – took shape [Zee 2010; Thomson 2013] between about 1970 and 1973. According to this scheme, all hadrons are composed of quarks and/or the anti-particles of quarks, known as *anti-quarks*. Each baryon is now taken to be composed of three quarks, and each (ordinary) meson, of a quark and an anti-quark. The quarks come in six different flavours, referred to (rather oddly and unimaginatively) as *up*, *down*, *charm*, *strange*, *top*, and *bottom*, and they have the respective electric charges $\frac{2}{3}, -\frac{1}{3}, \frac{2}{3}, -\frac{1}{3}, \frac{2}{3}, -\frac{1}{3}$. The fractional charge values seem, at first, to be distinctly odd, but for the observed free particles (such as baryons and mesons), the total electric charge always has a value that is an integer.

The standard model not only systematizes the seeming menagerie of the basic particles of nature, it also provides a good description of the main forces that influence them. Both the strong and the weak force are described in terms of an elegant mathematical procedure – referred to as *gauge theory* – that makes crucial use of the notion of a *bundle*, for which a brief description is given in §A.7, and to which I shall return, particularly in §1.8. The *base space* $\mathcal{M}$ of the bundle (for which notion, see §A.7) is space-time and, in the case of the strong force (which is the more mathematically transparent case), the fibre $\mathcal{F}$ is described in terms of a notion referred to as *colour*, which is assigned to the individual quarks (there being three basic alternative colours available to each quark). The theory of strong-interaction physics is referred to, accordingly, as *quantum chromodynamics* (QCD). I do not want to go into a proper discussion of QCD here because it is difficult to describe properly without using more mathematics than I can provide here [see Tsou and Chan 1993; Zee 2003]. Moreover, it is not "fashionable" in the sense that I mean to use the term here, because the ideas, while sounding exotic and strange, actually work extraordinarily well, not only forming a consistent and tightly knitted mathematical formalism, but finding excellent confirmation in experimental results. The QCD scheme would be studied in any physics research department concerned in a serious way with the theory of strong interactions, but it is not simply fashionable, in the sense intended here, because it is widely studied for very good scientific reasons!

For all its virtues, however, there are also powerful scientific reasons for striving to go beyond the standard model. One of these is that there are some thirty or so numbers in the standard model, for which its theory provides no explanation whatever. These include things like quark and lepton masses, quantities referred to as *fermion mixing parameters* (such as the Cabibbo angle), the Weinberg angle, the theta angle, gauge couplings, and parameters connected with the Higgs mechanism. Related to this issue is another serious drawback, which had already been very much present in other schemes that were around before the emergence of the standard model, and which is only partly resolved by it. This is the disturbing issue of the *infinities* (which are nonsensical answers arising from divergent expressions, like those exhibited in §A.10) that arise in quantum field theory (QFT) – QFT being the form of quantum mechanics that is central not just to QCD and other aspects of the standard model, but to all modern approaches to particle physics, and also to many other aspects of basic physics.

I shall have to say a good deal more about quantum mechanics, generally, in chapter 2. For the moment, let us restrict attention to one very specific but fundamental feature of quantum mechanics, which may be regarded as a root of the problem of the infinities in QFT, and we shall also see how the conventional method of dealing with these infinities precludes any complete answer to the issue of deriving the thirty or so unexplained numbers in the standard model. String theory is largely driven by an ingenious proposal to circumvent the infinities of QFT, as we shall be seeing in §1.6. It therefore appears to offer some hope of providing a route to resolving the mystery of the unexplained numbers.

## 1.4. THE SUPERPOSITION PRINCIPLE IN QFT

A foundation stone of quantum mechanics is the superposition principle, which is a feature common to *all* of quantum theory, not just QFT. In particular, it will be central to the critical discussions of chapter 2. In the present chapter, in order to provide some insight into the source of the problem of the infinities of QFT, I shall need to introduce this principle briefly here, though my main discussion of quantum mechanics will take place in chapter 2 (see, in particular, §§2.5 and 2.7).

To bring out the role of the superposition principle in QFT, let us consider situations of the following kind. Suppose that we have some physical process leading to a particular observed outcome. We shall suppose that this outcome could have arisen via some intermediate action $\mathbf{\Psi}$, but there is also another possible intermediate action $\mathbf{\Phi}$ which could also have resulted in essentially the same observed outcome. Then, according to the superposition principle, we must consider that, in an appropriate sense, *both* $\mathbf{\Psi}$ and $\mathbf{\Phi}$ could well have taken place *concurrently* as the intermediate action! This, of course, is very non-intuitive, since at an ordinary macroscopic scale we do not find distinct alternative possibilities taking place at once. Yet, for submicroscopic events, where we do not have the possibility of directly observing whether one intermediate activity has occurred as opposed to another, then we must allow that both could have occurred together, in what is referred to as a *quantum superposition*.

The archetypal example of this sort of thing occurs with the famous two-slit experiment, often used in introductions to quantum mechanics. Here we consider a situation where a beam of quantum particles (say, electrons or photons) is

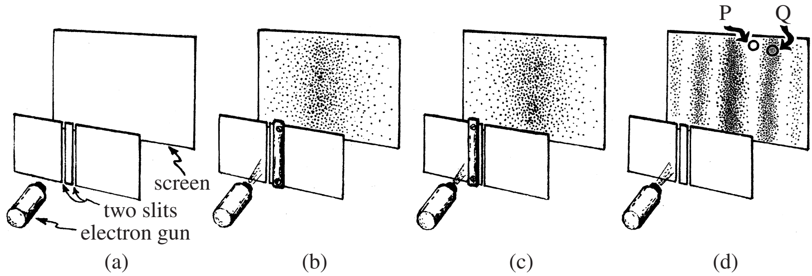


Figure 1-2: The two-slit experiment. Electrons are aimed at a screen through a narrowly separated pair of slits (a). If just one slit is open (b), (c), then a random-looking pattern of impacts is registered at the screen, scattered about the direct route through the slit. However, if both slits are open (d), the pattern acquires a banded appearance, where some places (e.g. P) cannot now be reached by the particle, although they could have been if just one slit were open; moreover, at other places (e.g. Q) there is *four* times the intensity of reception than for a single slit.

directed at a screen, where the beam must pass through a pair of close parallel slits on its way from source to screen (figure 1-2(a)). In the situation that is being considered, upon reaching the screen each particle makes a distinctive dark mark at an individual location at the screen, indicative of the particle's actual *particulate* nature. But after many such particles have passed through, an interference pattern of light and dark bands is built up, the dark bands occurring where many particles reach the screen and the light bands, where relatively few reach it (figure 1-2(d)). A standard careful analysis[2] of the situation leads one to conclude that each individual quantum particle must, in some sense, pass through *both* slits at once, in the manner of a strange kind of superposition of the two alternative possible routes that it might take.

The reasoning behind such an odd conclusion comes from the fact that if either one of the slits is covered up, while the other remains open (figure 1-2(b),(c)), we get no bands, but just a fairly uniform illumination which is darkest at the centre. When both slits are open, however, there are lighter regions at the screen

[2] This is what may be regarded as the conventional analysis of the situation. As might be expected, for such a strange-seeming conclusion, there are various other ways of interpreting what happens in this intermediate stage of the particle's existence. The most noteworthy alternative perspective is that of the de Broglie–Bohm theory according to which the particle itself always goes either through one slit or through the other, but there is also an accompanying "carrier wave" which guides the particle and which must *itself* "feel out" the two alternatives which the particle might have adopted [see Bohm and Hiley 1993]. I briefly discuss this viewpoint in §2.12.

situated between darker bands, these lighter regions occurring at places which are perfectly dark when just one of the slits is open. Somehow, when both routes are available to the particle, those lighter places become *inhibited*, whereas the dark places are *enhanced*. If each particle simply either did what it could do when only one of the slits is open or what it could do when only the other is open, then the effects of the routes would just add together, and we wouldn't get these strange interference stripes. This happens only because both of the possible routes are available to the particle, both these alternatives being felt out by the particle to give the ultimate effect. In some sense these routings *coexist* for the particle when it is between source and screen.

This, of course, is very much at odds with our experience of the behaviour of macroscopic bodies. For example, if two rooms are connected to each other by two different doors, and if a cat is observed to have started in one room and is later observed to be in the other room, then we would normally infer that it had passed through one door or through the other door, not that it could, in some strange way, have passed through both doors at the same time. But with an object of the size of a cat, it would be possible, without disturbing its actions significantly, to make continual measurements of its location and thereby ascertain which of the two doors it actually passed through. If we were successful in doing this for a single quantum particle in the two-slit experiment described above, we would have to disturb its behaviour to a degree that would result in the interference pattern at the screen being destroyed. The wave-like behaviour of an individual quantum particle that gives rise to the interference bands of light and dark at the screen depends upon our *not* being able to ascertain which of the two slits it actually went through, thereby allowing for the possibility of this puzzling intermediate superposed state of the particle.

In this two-slit experiment, we can see the extreme strangeness of the behaviour of single quantum particles most particularly by concentrating our attention at a point P of the screen at the middle of a gap between the dark bands, where we find that the particle is simply unable to reach P when both slits are open to it, whereas if only one of the slits is open, the particle could quite readily reach P via the open slit. When both slits are open, the two possibilities that are available to the particle in order to reach P have somehow cancelled each other; yet, at another place on the screen, say Q, where the interference pattern is at its darkest, we find that instead of cancelling, the two possible routes seem to reinforce each other so that when both slits are open, the likelihood of the particle reaching Q is four times as great as it would have been if just one of the two slits were open, not just twice, as would have been the case with an ordinary classical object, rather than a quantum particle. See figure 1-2(d). These strange features are a consequence of what is known as the *Born rule*, which relates the intensities in the superpositions to actual probabilities of occurrence, as we shall come to shortly.

The word *classical*, incidentally, when used in the context of physical theories, models, or situations, simply means *non-quantum*. In particular, Einstein's general theory of relativity is a classical theory, in spite of its having been introduced after many of the seminal ideas of quantum theory (such as the Bohr atom) had come about. Most particularly, classical systems are *not* subject to the curious superpositions of alternative possibilities that we have just encountered above, and which indeed characterizes quantum behaviour, as I come to briefly next.

I shall delay my full discussion of the basis of our present understanding of quantum physics until chapter 2 (see, in particular, §2.3 onwards). For the moment, I recommend that we simply accept the strange mathematical rule whereby modern quantum mechanics describes such intermediate states. The rule turns out to be extraordinarily accurate. But what is this strange rule? The quantum formalism asserts that such a superposed intermediate state, when there are just two alternative intermediate possibilities $\boldsymbol{\Psi}$ and $\boldsymbol{\Phi}$, is to be expressed mathematically as some kind of a sum $\boldsymbol{\Psi} + \boldsymbol{\Phi}$ of the two possibilities or, more generally, as a *linear combination* (see §§A.4 and A.5),

$$w\boldsymbol{\Psi} + z\boldsymbol{\Phi},$$

where $w$ and $z$ are *complex numbers* (the numbers involving $\mathrm{i} = \sqrt{-1}$, as described in §A.9), not both being zero! Moreover, we shall be forced to consider that such complex superpositions of states have to be allowed to *persist* in a quantum system, right up until the time that the system is actually observed, at which point the superposition of alternatives must be replaced by a probability mixture of the alternatives. This is indeed strange, but in §§2.5–2.7 and 2.9 we shall be seeing how to use these complex numbers – sometimes referred to as *amplitudes* – and how they tie in, in remarkable ways, with probabilities, and also with the time evolution of physical systems at the quantum level (Schrödinger's equation); they also relate, fundamentally, to the subtle behaviour of the spin of a quantum particle, and even to the 3-dimensionality of ordinary physical space! Although the precise connections between these amplitudes and probabilities (the *Born rule*) will not be addressed fully in this chapter (since for this we need the notions of *orthogonality* and *normalization* for the $\boldsymbol{\Psi}$ and $\boldsymbol{\Phi}$, which are best left until §2.8), the gist of the Born rule is as follows.

A measurement, geared to ascertaining whether a system is in state $\boldsymbol{\Psi}$ or in state $\boldsymbol{\Phi}$, when presented with the superposed state $w\boldsymbol{\Psi} + z\boldsymbol{\Phi}$, finds:

$$\text{ratio of probability of } \boldsymbol{\Psi} \text{ to probability of } \boldsymbol{\Phi} = \text{ratio of } |w|^2 \text{ to } |z|^2.$$

We note (see §§A.9 and A.10) that the squared modulus $|z|^2$ of a complex number $z$ is the sum of the squares of its real and imaginary parts, this being the squared distance of $z$ from the origin in the Wessel plane (figure A-42 in §A.10). It may also be remarked that the fact that probabilities arise from *squares* of the moduli of these amplitudes accounts for the *fourfold* increase in intensity, as noted earlier, where contributions reinforce each other in the two-slit experiment (see also the end of §2.6).

We must be careful to appreciate that this notion of *plus*, in these superpositions, is quite different from the ordinary notion of *and* (despite a common modern use of *plus* in ordinary conversation simply to mean *and*), or even with *or*. What is meant here is that, in some sense, the two possibilities are actually to be thought of as being *added* together in some abstract mathematical way. Thus, in the case of the two-slit experiment, where $\boldsymbol{\Psi}$ and $\boldsymbol{\Phi}$ represent two distinct transient locations of a single particle, then $\boldsymbol{\Psi} + \boldsymbol{\Phi}$ does *not* represent two particles, one in each location (which would be "one particle in the $\boldsymbol{\Psi}$ position *and* one particle in the $\boldsymbol{\Phi}$ position" – implying *two* particles in total), nor must we think of the two as just being ordinary alternatives, one *or* the other of which actually happened, but where we don't know which. We must indeed think of just a single particle somehow occupying both locations at once, *superposed* according to this strange quantum-mechanical "plus" operation. Of course this looks extremely odd, and the physicists of the early twentieth century would not have been driven to consider such a thing without having some very good reasons to do so. We shall be exploring some of these reasons in chapter 2, but for now I am just asking that the reader simply accept that this formalism indeed works.

It is important to appreciate that, according to standard quantum mechanics, this superposition procedure is taken to be *universal* and, accordingly, applies also if there are more than just two alternatives for the intermediate state. For example, if there are three alternative possibilities, $\boldsymbol{\Psi}$, $\boldsymbol{\Phi}$, and $\boldsymbol{\Gamma}$, then we have to consider triple superpositions like $w\boldsymbol{\Psi} + z\boldsymbol{\Phi} + u\boldsymbol{\Gamma}$ (where $w$, $z$, and $u$ are complex numbers, not all of which are zero). Correspondingly, if there were four alternative intermediate states, we would need to consider quadruple superpositions and so on. Quantum mechanics demands this, and there is excellent experimental support for such behaviour at the submicroscopic level of quantum activity. Strange it is, indeed, but it makes good consistent mathematics. This is, so far, just the mathematics of a *vector space*, with complex-number scalars, as considered in §§A.3, A.4, A.9, and A.10, and we shall be seeing more of the ubiquitous role of quantum superpositions in §2.3 onwards. However, matters are considerably worse in QFT, because we frequently have to consider situations in which there are infinitely many intermediate possibilities. Accordingly, we are led to having to consider infinite sums of alternatives, and then the issue looms large of the possibility that such an infinite sum might provide us with series whose sum may actually *diverge* to infinity (in the sort of way exhibited in §§A.10 and A.11).

## 1.5. THE POWER OF FEYNMAN DIAGRAMS

Let us try to understand in a bit more detail how such divergences actually come about. In particle physics, what we have to consider are situations in which several particles come together to make other particles, where some of them may split apart to make still others, and where pairs of these might join together again, etc., etc., so that they may well be involved in very complicated processes of this kind. The types of situation that particle physicists are frequently concerned with involve some given collection of particles coming together – often at relative speeds close to that of light – and this combination of collisions and separations results in some other collection of particles emerging from it all. The total process would involve a vast quantum superposition of all the possible different kinds of intermediate processes which might take part and are consistent with the given input and output. An example of such a complicated process is illustrated in the *Feynman diagram* of figure 1-3.

We do not go far wrong if we think of a Feynman diagram as a space-time diagram of the particular collection of particle processes involved. I like to represent time as proceeding upwards along the page, being someone who works in relativity theory as opposed to being a professional particle physicist or QFT expert; the professionals usually have time progressing from left to right. Feynman diagrams (or Feynman graphs) are named after the outstanding American physicist Richard Phillips Feynman. Some very basic diagrams of this kind are shown in figure 1-4. Here, figure 1-4(a) shows the splitting of a particle into two and figure 1-4(b) shows the combining of two to make a third.

In figure 1-4(c), we see the exchange of a particle (say a photon, the quantum of electromagnetic field or light, indicated by the wiggly line) between two particles. The use of the term *exchange* for this process, though common among particle

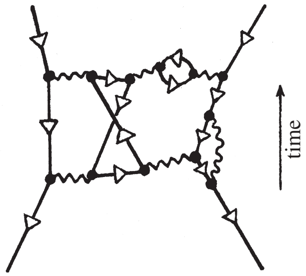


Figure 1-3: A Feynman diagram (drawn with upward time direction) is a schematic space-time picture (with a clear-cut mathematical interpretation) of a particle process often involving creation, annihilation, and exchanges of intermediate particles. Wavy lines indicate photons. Triangular arrows here denote electric charge (positive if the arrow points upwards and negative if downwards).

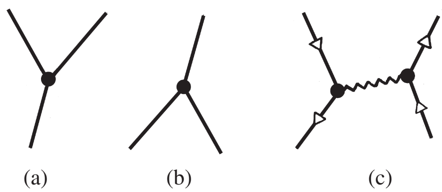


Figure 1-4: Elementary Feynman diagrams: (a) a particle splits into two; (b) two particles combine to make another particle; (c) two oppositely charged particles (e.g. an electron and a positron) “exchange” a photon.

physicists, is perhaps a bit odd here, since a *single* photon simply passes from one external particle to the other – albeit in a way that (deliberately) does not make clear which particle is the emitter and which the receiver. The photon involved in such an exchange is usually what is called *virtual* and its speed is not constrained to be consistent with the requirements of relativity. The usual colloquial use of the term *exchange* might apply more appropriately to the situations depicted in figure 1-5(b), though such processes as shown in figure 1-5 tend to be referred to as the exchange of two photons.

We may think of the general Feynman diagram to be composed of many basic ingredients of this general kind, pieced together in all sorts of combinations. However, the superposition principle tells us not to think of what actually happens in some such particle collision process as being represented by just *one* such Feynman diagram, because there are many alternatives, and the actual physical process is represented as some complicated linear superposition of many different such

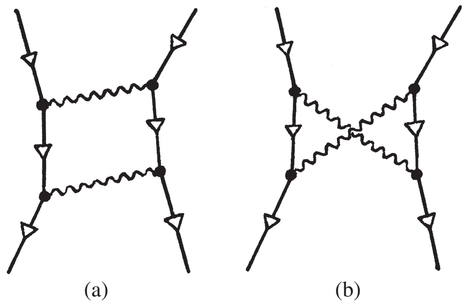


Figure 1-5: Two-photon exchanges.

Feynman diagrams. The magnitude of the contribution to the total superposition from such a diagram – essentially a complex number such as the $w$ or $z$ that we encountered in § 1.4 – is what we need to calculate from any particular Feynman diagram, these numbers being called complex *amplitudes* (see §§ 1.4 and 2.5).

We must bear in mind, however, that the mere arrangement of the connections in the diagram does not tell us the whole story. We also need to know the values of the energies and momenta of all the particles involved. For all the external particles (both incoming and outgoing), we may take these values to be already assigned, but the energies and momenta of the intermediate – or internal – particles could generally take many different values, consistent with a constraint that energy and momentum have to add up appropriately at each vertex, where the momentum of an ordinary particle is its velocity multiplied by its mass; see §§A.4 and A.6. (Momentum has the important property that it is *conserved*, so that in any collision process encountered by particles, the total of the momenta going in – added together in the sense of vector addition – must be equal to the total momentum coming out.) Thus, complicated as our superpositions may appear to be, merely from the elaboration of the succession of increasingly complicated diagrams appearing in the superposition, things are really much *more* complicated than this, because of the generally *infinite* numbers of different possible values that the energies and momenta might take for the internal particles in each diagram (consistent with the given external values).

Thus, even with a single Feynman diagram, with given input and output, we may expect to have to add together an infinite number of such processes. (Technically, this adding together would take the form of a continuous integral, rather than a discrete sum (see §§A.7, A.11, and figure A-44), but the distinctions are not important for us here.) This kind of thing happens with Feynman diagrams containing a *closed loop*, such as occurs in the two examples of figure 1-5. With a

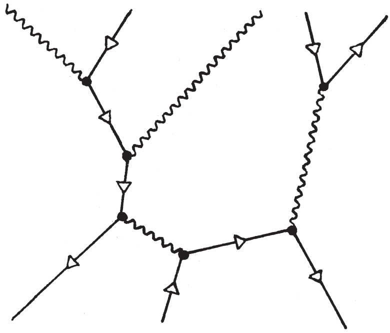

Figure 1-6: A tree diagram, i.e. containing no closed loops.

*tree diagram*, such as those of figure 1-4 and figure 1-6, where there are no closed loops, the values of the internal energies and momenta turn out to be simply fixed by the external values. But these tree diagrams do not probe the genuinely quantum nature of particle processes; for this we actually need to bring in the closed loops. And the trouble with the closed loops is that there is no limit to the energy-momentum that can, in effect, circulate around the loop, and adding all these up provides us with a *divergence*.

Let us look at this a little more closely. One of the simplest situations where a closed loop occurs is that shown in figure 1-5(a) in which two particles are exchanged. The trouble arises because although at each vertex in the diagram the values of the energy and of the three components of momentum must add up properly (i.e. the sum in equals the sum out), this does not result in enough equations to fix the internal values of these quantities. (For each of the four components of the energy-momentum, separately, there are three independent equations since each of the four vertices provides a conservation equation, but one is redundant, merely re-expressing the overall conservation for the entire process – yet there are four independent unknowns per component, one from each internal line, so there are not enough equations to fix the unknowns, and the redundancy must be summed over.) There is always the freedom to add (or subtract) the same energy-momentum quantity all the way around the loop in the middle. We need to add all these infinitely many possibilities together, involving potentially higher and higher values for the energy-momentum, and this is what leads to the potential divergence.

Thus, we see that the direct application of the quantum rules is indeed likely to give us a divergence. Yet, this does not necessarily mean that the "correct" answer to that quantum field theoretic calculation is actually $\infty$. It would be useful to keep in mind the divergent series shown in §A.10, where a finite answer can sometimes be assigned the series despite the fact that simply adding the terms up leads to the answer "$\infty$". Although the situation with QFT is not exactly like this, there are some distinct similarities. There are many calculational devices that QFT experts have developed over the years in order to circumvent these infinite answers. Just as with the examples of §A.10, if we are clever about it, we may be able to unearth a "true" finite answer that we do not get simply by "adding up the terms". Accordingly, QFT experts are frequently able to squeeze finite answers out of the wildly divergent expressions that they are presented with, although many of the procedures that are adopted are far less straightforward than simply the method of analytic continuation, referred to in §A.10. (See also §3.8 for some of the curious pitfalls that even the "straightforward" procedures can lead into.)

One key point about the root cause of many of these divergences – those referred to as *ultraviolet divergences* – should be made note of here. The trouble basically arises because, with a closed loop, there is no limit to the scale of energy and momentum that can circulate around it, and the divergence arises from contributions of higher and higher energy (and momentum) having to be added up. Now, according to quantum mechanics, very large values of energy are associated with very tiny times. Basically this comes from Max Planck's famous formula $E = h\nu$, where $E$ is the energy, $\nu$ is the frequency, and $h$ is Planck's constant, so high values of energy correspond to large frequencies and therefore to tiny time intervals between one beat and the next. In the same way, very large values of the momentum correspond to very tiny distances. If we imagine that something strange happens to space-time at very tiny times and distances (as, indeed, most physicists would be inclined to agree would be an implication of quantum-gravity considerations), there might be some kind of effective "cut-off", at the high end of the scale, to the allowed energy-momentum values. Accordingly, some future theory of space-time structure, in which drastic alterations occur at very tiny times or distances, might actually render finite the currently divergent QFT calculations that arise from closed loops in Feynman diagrams. These times and distances would have to be far tinier than those which are relevant to ordinary particle-physics processes, and are frequently taken to be something of the order of those quantities of relevance to quantum-gravity theory, namely the *Planck time* of some $10^{-43}$ s or the *Planck length* of around $10^{-35}$ m (referred to in §1.1), these values being something like $10^{-20}$ of the usual small quantities of direct relevance to particle processes.

It should be mentioned here that there are also divergences in QFT referred to as *infrared divergences*. These occur at the other end of the scale, where energies

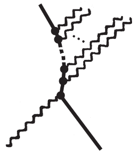

Figure 1-7: Infrared divergences occur when there are indefinitely large numbers of "soft" photons emitted.

and momenta are extremely tiny, so that we are concerned with extraordinarily large times and distances. The problems here are not to do with closed loops, but with Feynman diagrams like those of figure 1-7 in which an unlimited number of *soft photons* (i.e. photons of very tiny energy) might be emitted in a process, and adding all these together again produces a divergence. Infrared divergences tend to be regarded by QFT experts as less serious than the ultraviolet divergences, and there are various ways of sweeping them under the carpet (at least temporarily). However, in recent years their importance is perhaps beginning to be faced up to more seriously. For my own discussions here, I shall not pay too much attention to the infrared problems and concentrate instead on how the problem of the ultraviolet divergences – resulting from the closed loops in the Feynman diagrams – is tackled in standard QFT, and how the ideas of string theory appear to offer hope of providing a resolution of this conundrum.

Of particular note, in this connection, is the standard QFT procedure of renormalization. Let us try to glimpse how this operates. According to various direct QFT calculations we get an infinite scale factor between what would be called the *bare charge* of a particle (such as an electron) and the *dressed charge*, the latter being what would be actually measured in experiments. This comes about because of contributions due to processes like that shown in the Feynman diagram of figure 1-8, which serve to damp down the measured value of the charge. The trouble is that the contribution from figure 1-8 (and many another like it) is "infinity". (It has closed loops.) Accordingly, we find that the bare charge would have to have been infinite in order that we can find a finite value for the observed (dressed) charge. The underlying philosophy of the renormalization procedure is to accept that QFT might not be completely right at very tiny distances, which is where the divergences appear, and some unknown modification of the theory might supply the necessary cut-off that leads to finite answers. The procedure thus involves our giving up on attempting to calculate nature's actual answer for these scale factors (for charge and for other things like mass, etc.), where we instead collect together all such infinite scale factors that QFT burdens us with, and we

(continued...)

# Index