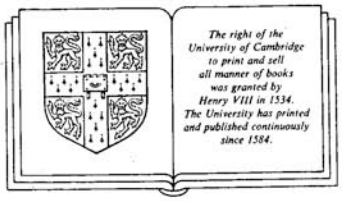
The New Physics

Edited by
Paul Davies
Professor of Theoretical Physics
University of Newcastle upon Tyne



CAMBRIDGE UNIVERSITY PRESS
Cambridge
New York *New Rochelle* *Melbourne* *Sydney*
1989

16 Effective quantum field theories

H. M. Georgi

In Chapter 15 I discussed the idea of Grand Unified Theories (GUTs) of particle interactions. I still find that subject very interesting, despite the disappointing failure of experiments to find proton decay. But I have mixed emotions. I feel about the present state of GUTs as I imagine that Richard Nixon's parents might have felt had they been around during the final days of the Nixon administration. I am very proud that the grand unification idea has become so important. After all, at first it was something of an ugly duckling, roundly ridiculed by everyone. But proud as I am, I cannot help being very disturbed by the things which GUTs are doing now.

GUTs were motivated by the physics of $SU(2) \times U(1)$ and colour SU(3) and the desire to predict the value of the weak mixing angle (a parameter in the $SU(2) \times U(1)$ model that measures the relative strength of the SU(2) and U(1) interactions) and to explain the quantisation of electrical charge. They were certainly not an attempt to emulate Einstein and produce an elegant geometrical unification of all interactions including gravity, despite the parallels which have been drawn in the semipopular press. Einstein's attempts at unification were rearguard actions which ignored the real physics of quantum mechanical interactions between particles in the name of philosophical and mathematical elegance. Unfortunately, it seems to me that many of my colleagues are repeating Einstein's mistake. It is primarily for this reason that I want to address the larger picture.

The language of relativistic quantum mechanics is called 'quantum field theory'. In a quantum field theory, a field (like the magnetic field surrounding a magnet) is assigned quantum mechanical properties and is seen to be associated with a type of particle (the magnetic field, for example, is associated with the photon, the particle of light). In the last fifteen years, our understanding of quantum field theory has changed considerably. I believe that we have arrived at a mature and satisfying view of the subject. It is this modern view of quantum field theory that I will discuss here. The view of many field theorists today is that the most appropriate description of particle interactions in the language of quantum field theory depends on the energy at which the interactions are studied.

Table 16.1. *Logarithmic energies and related distances*

Energy ($M_p \times c^2$)	Distance (cm)	Associated physics
		Quantum
10^{19}	10^{-33}	Gravity?
10^{14}	10^{-28}	GUTs?
10^{2}	10^{-16}	W and Z particles
		Quarks
10^{-1}	10^{-13}	Nuclei
10^{-6}	10^{-8}	Atoms
		Molecules
10^{-11}	10^{-3}	Amoeba
10^{-16}	10^{2}	People

The description is in terms of an 'effective field theory' that contains explicit reference only to those particles that are actually important at the energy being studied. The effective theory is the most useful means of extracting the physics in a limited range of energies, but it changes as the energy changes, to reflect changes in the relative importance of different particles and forces. I should say at the outset that my own role in the development of this view has been minor compared to that of giants like Ken Wilson and Steve Weinberg. But the invention of GUTs was important because it forced us to think seriously about particle physics at extremely short distances. But I am getting ahead of myself. I am going to begin by making some rather obvious statements about the nature of physics.

16.1 Dimensional analysis

Dimensional analysis is one of the oldest and most important physical ideas. The key principle in dimensional analysis is that physics should not depend on the units in which physical parameters are measured. Because of this principle, the dimensions of a physical quantity can tell you a lot about the physics itself. To some degree, this principle is incorporated into our common sense, that uneven distillation of instinct, experience and learning that we carry with us in our study of the limited physical universe of our everyday lives. When we hear that something is measured in square centimetres, we immediately think of an area. Kilometres per hour identifies a velocity. Calories measure energy (often in a tempting form that produces guilt when we eat it). We know from experience that the actual value of each of these quantities depends on comparison with a set of units that has no fundamental significance except perhaps historically. We know that we can convert from one set of units to another, as long as the quantities measured are the same. But while we can compare centimetres and inches, we know that it doesn't make sense to compare centimetres with square inches. Inches and centimetres represent the same dimensional property in different units. Centimetres and square inches are dimensionally different so they always measure physically different things.

Some other units are less familiar. Momentum, which has units of mass times distance over time, grammes times centimetres per second for example, measures the tendency of a moving object to keep going in a straight line. Momentum times distance (for example grammes times square centimetres per second) is the unit of something called angular momentum, the unit of spin. It is the physics of angular momentum that an ice skater uses to perform a rapid revolution. For a fixed angular momentum (and it is fixed unless the skater is given a twist by some outside force such as the push or pull of the skates on the ice) the rate of the skater's turning is inversely proportional to the area over which his mass is spread. When he lifts his arms over his head, this area decreases and his rate of revolution increases, even though his angular momentum stays about the same (see figure 16.1).

We all understand in our bones that dimensional quantities, those with units, have values that depend on the system of units in which they are measured. That is very different from a pure number like 1 or π. This is common sense. But dimensional analysis is much more general than common sense. Common sense applies only in the bounded domain of human experience. Classical physics, the physics of Newton, is built on common sense and quantifies and extends it in a precise mathematical language to explain in great detail such different phenomena as the fall of an apple and the motion of a planet. It works so impressively well in the domain of our everyday experience and at larger, astronomical distances, that in the nineteenth century, many physicists were confident that they knew most of what was worth knowing about the way the world works.

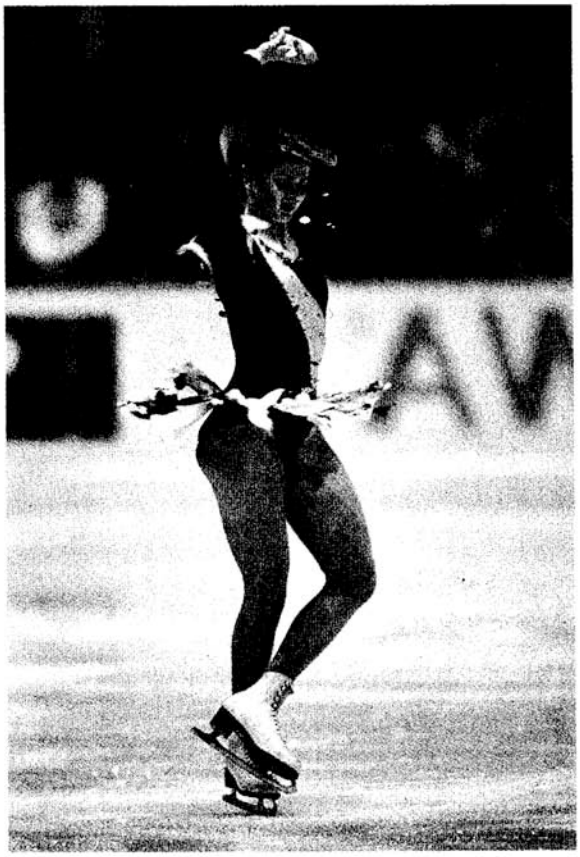


Figure 16.1. Conservation of angular momentum is well illustrated by the spinning ice skater. By the drawing in of the arms, so concentrating the skater's mass closer to the axis of rotation, the rate of rotation increases.

16.2 Biology is not a branch of physics

In fact, however, classical and nonrelativistic physics are only approximate theories which work well for velocities much smaller than the speed of light, c, which is roughly equal to $3 \times 10^{10}\,\mathrm{cm\,s^{-1}}$ and angular momentum much larger than Planck's constant, $\hbar$, which is about $10^{-27}\,\mathrm{cm^2\,s^{-1}}$. c and $\hbar$ are fundamental constants which mark the boundaries between different appropriate descriptions of the world. The word 'appropriate' is crucial here. It is easy to say that classical and nonrelativistic physics have simply been replaced by quantum mechanics and relativistic physics, the theories that we know

are needed to understand the world at small angular momenta and large velocities. But it is not true that they have simply been replaced. It is not true in a sense which is similar to the sense in which the statement 'chemistry and biology are branches of physics' is not true. It *is* true that in chemistry and biology one does not encounter any new physical principles. But the systems on which the old principles act differ in such a drastic and qualitative way in the different fields that it is simply not *useful* to regard one as a branch of another. Indeed the systems are so different that 'principles' of new kinds must be developed, and it is the principles which are inherently chemical or biological which are important. In the same way, to study phenomena at velocities much less than c and angular momentum much greater than $\hbar$, it is simply not useful to regard them as special cases of phenomena for arbitrary velocity and angular momentum. In fact, we usually put the logic the other way around. The correspondence principle is the statement that our quantum mechanical description of the world must reduce to the simpler classical description in the appropriate domain. We don't need relativity and quantum mechanics for small velocities and large angular momenta. It's just as well, too, because if we had had to discover the laws of relativistic quantum mechanics from the beginning, we probably would never have gone anywhere.

Particle physicists, like me, tend to forget all this, because we are interested almost exclusively in velocities nearly equal to the speed of light, c, and angular momenta not much larger than $\hbar$. We don't need the fundamental constants c and $\hbar$ as boundaries because we are always in the same domain, so we just set them equal to one and measure all dimensional quantities in units of mass, or whatever. For example, when we set c equal to one, a second can be either a time of one second or the distance that light travels in one second. When we set $\hbar$ equal to one, one centimetre can be either a distance of one centimetre or the inverse of the momentum required to produce an angular momentum of $\hbar$ at a distance of one centimetre from the axis. Thus we can measure energy and momentum in units of mass, time and distance in units of inverse mass, force in units of mass squared, etc. This habit is so ingrained that we tend to use these units interchangeably. For example, I will often convert the mass of a particle, M, into a distance, $1/M$, called the Compton wavelength of the particle, the length at which we see the particle's quantum mechanical properties. For example, the Compton wavelength of an electron is about 4×10^{-11} cm (40 millionths of a millionth of a centimetre). The proton, which is almost 2000 times heavier, has a Compton wavelength about 2000 times *smaller*, or 2×10^{-14} cm. This is a trivial exercise in dimensional analysis, but it illustrates a general feature of the quantum mechanical world. The heavier a particle is, or the higher its energy, the smaller is the distance at which its quantum effects appear.

Having said these obvious (?) things, I will now proceed with a brief review of the history of quantum field theory.

16.3 Local quantum field theory

Field theory developed in the late 1920s and early 1930s to describe the interactions of electrons and photons. It was the natural synthesis of quantum mechanics and relativistic wave equations like Maxwell's equations (the equations that describe the properties of classical electric and magnetic fields) and the Dirac equation (an attempt to describe the properties of the field associated with the electron). Today we would say that this particular synthesis was more than just natural, it was inevitable. Local quantum field theory is the only way to combine a quantum mechanical theory of particles with special relativity consistent with causality. *Causality* is the general principle that causes should always happen before their effects. The word 'local' here is the crucial one. A local quantum field theory is one in which the interactions which cause scattering or creation or annihilation of particles take place at single space-time points. Locality is important because action at a distance causes trouble with causality in relativistic theories.

Obviously, the assumption of locality is an act of incredible hubris. After all, a 'space-time point' is not a physical thing. It is a mathematical abstraction – infinitely small. To really know how particles interact at a single point you have to understand how the world works down to arbitrarily small distances. That is ridiculous! Only a particle theorist would have the infernal gall to even propose such a thing!

Nevertheless, early quantum field theory yielded many important results, although, from our modern vantage point, the logic often seems confused. The Dirac equation was the first relativistic treatment of electron spin. Goudsmit and Uhlenbeck discovered in the 1930s that the electron behaves like a spinning top. Its angular momentum is exactly half of Planck's constant, $\hbar/2$. A spinning electrically charged particle should act as a magnet. The strength of the magnetism of such a particle is measured by its g factor, which is the ratio of the actual strength of the particle's internal magnet to that of a point particle with the same charge and mass moving around a fixed axis to give the same angular momentum. Dirac's theory not only incorporated the spin of the electron in a way which was consistent with Einstein's relativity principle, but it also 'predicted' that the g factor of the electron is 2, close to its experimental value. In fact, however, this was really only an aesthetic argument. The arguments which led to the Dirac equation also allow an additional term in the equation called a Pauli term with an arbitrary coefficient, which gives an arbitrary g factor. It is really only the 'simplest' Dirac equation which gives $g=2$.

The 'prediction' of the positron was confused in a different way. The Dirac equation has negative energy solutions. Dirac realised brilliantly that, if nearly all the negative energy states were filled, the few 'holes' would behave like particles with positive charge. But at first he interpreted these as protons! The work of Weyl, Oppenheimer and others convinced Dirac that

this interpretation was untenable and that his equations predicted the existence of a genuinely new particle with the same mass as the electron but with positive charge. Then the positron was discovered, and eventually it was recognised by all that it was Dirac's hole particle. Today, of course, we believe that all particles have antiparticles with the same mass and opposite charges (particles with no charge, like the photon, can be their own antiparticles).

Perhaps the most impressive successes of early quantum field theory were the calculations of scattering probabilities for a variety of processes. These were possible because the theory contained a small dimensionless parameter, the 'coupling constant', $e^2/\hbar c = \alpha \simeq 1/137$ (see Chapter 15). Scattering probabilities could be calculated unambiguously to first order in α. This means that the contribution to the probability that is proportional to the small parameter α is calculated while the contributions proportional to α^2 and higher powers of α are ignored. The processes which were studied included (where e^- is an electron and e^+ a positron):

electron–photon scattering	Klein and Nishina, 1929
e^+e^- annihilation into photons	Dirac, 1930
electron–electron scattering	Moller, 1932
$e^- \rightarrow e^- +$ photon in nuclear field	Bethe and Heitler, 1934
photon $\rightarrow e^+e^-$ in nuclear field	
$e^+e^- \rightarrow e^+e^-$	Bhabha, 1936

They were in reasonable agreement with experiment.

16.4 The tragic flaw

But not all was well. For one thing, in the 1930s and 40s, it became clear that a theory of electrons and photons could not be the whole story of particle interactions because there were other particles and interactions: the proton, the neutron, the pion, the muon, the neutrino . . ..

But what was worse was that the formalism of quantum field theory itself seemed to have a tragic flaw. When nontrivial calculations were attempted to higher order in α, the results were infinite! This meant that the theory defined in the most naïve way simply did not make sense. These infinities arose precisely because of the local nature of the interaction. They were, in fact, the punishment imposed on particle theorists for the hubris of locality. And they worried people a lot. The possibility that it might be possible to absorb the infinities by 'renormalisation' of the physical parameters was discussed, but not completely understood.

16.5 Quantum electrodynamics

So matters stood until the theoretical physics community reassembled after World War II. Stimulated by exciting experiments (such as the measurement by Willis Lamb of the small energy shift between two states of the hydrogen atom that were predicted to have the same energy to first order in α), theorists used renormalisation to do finite calculations of quantum corrections to the first order results (such as the g factor of the electron and the Lamb shift). In renormalisation, the theory is defined by a limiting procedure. First the physics at distances shorter than some cut-off length is modified so that the calculations make sense. This involves giving up one or more of the cherished principles which led to local interactions in the first place. Then the parameters in this modified theory are expressed in terms of physical, measurable quantities (masses, scattering probabilities, etc.). Finally, the renormalised theory is defined by taking the cut-off length to zero. Presumably, this restores all the nice properties that a local quantum field theory should have. With the new theoretical tool of Feynman diagrams, it was possible to show that renormalisation was sufficient to absorb all the infinities in a quantum field theory into renormalised physical parameters. However, unless the theory is carefully constructed, the number of parameters required is infinite! The special theories in which only a finite number of physical parameters are required to unambiguously define the physics were called 'renormalisable'. Fortunately, the simplest theory of electrons and photons had the special property of renormalisability. This was exactly the form of the theory in which the electron's g factor came out right. *The local quantum field theory that made sense was also the one that accurately described the world. Quantum electrodynamics had come of age.*

Note that the logic here is a bit peculiar. The infinities in local quantum field theory, which were regarded as a disaster when they were first uncovered, became an asset with the development of renormalisation and quantum electrodynamics. Particle physicists now had another principles, renormalisability – another constraint to impose on their theories. An extra constraint is always very useful because it decreases the number of theories that you have to think about, which, in turn, decreases the amount of work that you have to do. But many physicists were uneasy about it, because this particular asset still had its roots in an apparent disaster. Here is a similar situation. Suppose I need a used car and I go to the only used car dealer in town and find that every car, when I start the engine, makes a horrible scraping noise and, after a minute or two, starts smoking and smelling awful and stops running. But the dealer tells me that there is one car which starts the same way, but if I gun the engine and pound on the dashboard the scraping sound goes away and the car runs beautifully. Well, I try it out and it works! Terrific! I don't even have a decision to make. So I buy the car and it's a great car, just what I need. But

somehow I can never stop wondering about what causes the horrible scraping noise.

Nevertheless, the success of renormalisable quantum electrodynamics in the following decade was spectacular. The electron magnetic moment (g factor) and other quantities were calculated to incredible accuracy and agreed well with increasingly precise experiments. Quantum electrodynamics became the paradigm of a successful physical theory. It began to look as if hubris were justified.

16.6 New physics

The quantitative success of renormalisable quantum electrodynamics sustained the prestige of quantum field theory for nearly ten years. But in the late 1950s and the 1960s, renormalisable quantum field theory began to run out of steam. It was not that there was anything wrong with quantum electrodynamics. There were just a lot of other things going on. Experimental physics marched forward. Meson and baryon resonances were discovered and studied. By definition, resonances are incredibly short-lived structures, so that it barely makes sense to call them particles at all. They decay as fast as they are produced through so-called strong interactions into more stable things such as pions and protons. Strange particles were also discovered and studied. These are comparatively long-lived particles, produced in particle–antiparticle pairs by the strong interactions but decaying singly much more slowly by so-called weak interactions.

At first, neither the strong nor the weak interaction could be usefully described using renormalisable quantum field theory. The problem with the strong interactions was that they were strong. There was no small dimensionless parameter, like $\alpha = 1/137$ of quantum electrodynamics. Reliable dynamical calculations were impossible. Progress *was* made by using symmetry arguments to extract information which was independent of the dynamical details, but that was all.

The weak interactions as seen in the decays of various particles, on the other hand, could apparently be described very well by a quantum field theory, to lowest order in a parameter, the Fermi constant, G_F, which is about equal to 10^{-5} in units of inverse square of the proton mass. The trouble was that the constant had dimensions of inverse mass squared. As I will show you in a minute, this means that the theory is *not* renormalisable. That in turn means that you must specify an infinite number of physical input parameters to determine the physics unambiguously. Most physicists felt, not unreasonably, that this was an unsatisfactory situation.

16.7 Dimensional parameters and renormalisability

That dimensionless parameters are OK while parameters with dimensions of $1/m^n$ lead to trouble can be seen, more or less, by dimensional analysis. The physical idea which underlines renormalisation is as follows. In the limiting procedure, in which the cut-off length is taken to zero, the physics itself must not depend on the cut-off, but our description of the physics may. In fact, the parameters of the modified, cut-off theory *must* be chosen to depend on the cut-off in order to keep physics at distances much larger than the cut-off length fixed when the cut-off changes. It follows that any dependence on the cut-off that is important for very small cut-off lengths must be associated with some parameter in the theory. To understand the cut-off dependence, we must understand these parameters. The parameters in a quantum field theory are the masses of the particles it describes and the relativistically invariant amplitudes for the various elementary scattering processes which describe the interactions. Because the theory is quantum mechanical, we expect these parameters to determine the quantum mechanical amplitudes whose squares are related to the probabilities for particle scattering.

The essential fact is that the local nature of the interactions requires the mass dimension of the amplitudes that describe the scattering processes to become more negative as the number of particles involved in the scatterings increases. The reason is just Heisenberg's uncertainty relation. In a quantum mechanical theory, there is a limit to the accuracy with which one can determine simultaneously the position of a particle and its momentum. The product of the uncertainties in the position and the momentum of any particle must always be greater than Planck's constant $\hbar$. But if several particles interact at the same point in space, we know their relative positions exactly. At the instant of the interaction, they are all sitting at the same point. Therefore, according to Heisenberg, we can't know anything about their relative momenta. Apart from the difference between the total momentum before and after the scattering (which is fixed by energy and momentum conservation), all the momenta can be varied continuously over the whole kinematically allowed region for the process. That, in turn, implies that the probability that the scattering will take place for any definite value of the momenta is infinitesimally small.

Here is a similar situation. Suppose that I put a penny somewhere at random inside a cubic box of side one metre which is full of sand. Now I can offer to pay you a billion dollars if you guess exactly the position of the centre of the penny in the box. I am completely safe because the probability of the penny being exactly in any given spot is zero. To give you a chance (which I would never do because I can't pay off the bet!) I would have to let you specify a volume in which you guess that the centre of the penny lies. The probability that your guess will be

correct is then the ratio of the volume you specify to the total volume of the box ($1\ m^3$). This means that the quantity that describes the likelihood that the centre of the penny will be at a given point is dimensional. It has units of inverse cubic metres. I know that because I must multiply it by a volume with units of cubic metres to get a dimensionless probability.

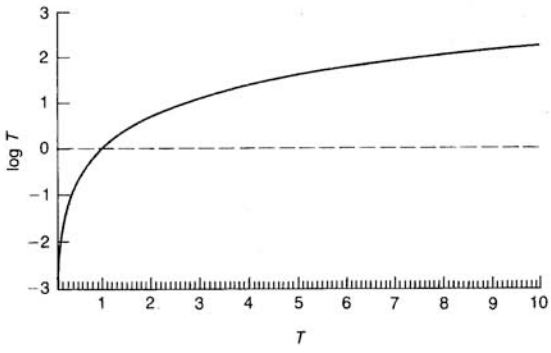
In the same way, to get a finite value for the probability that a local interaction will cause particle scattering, we must multiply the square of the amplitude by a 'phase space' factor that gives the 'volume' or range of momentum, with units of mass cubed. We must do this for each particle momentum. The result is a probability which is dimensionless. That means that the amplitudes, which contain the parameters that describe the scatterings, must have mass dimension which decreases (in this case, becomes more negative) as the number of particles involved in the scattering increases. It turns out that only the simplest scattering processes involving two, three or four particles at a time can be associated with parameters with dimensions of mass to the zero or a positive power. So you see that the parameters are asymmetrical. There are only a finite number of possibilities for parameters with dimensions of mass to the zero or a positive power. But the number of possible parameters with dimensions of mass to a negative power is infinite. The more particles involved, the more negative the mass dimension.

When we calculate quantum corrections in perturbation theory in these parameters, the results can depend on the cut-off length which we will eventually take to zero or alternatively on a cut-off mass which we will eventually take to infinity. We worry about contributions that go to infinity as M goes to infinity. Any dangerous term will involve a product of two or more of the parameters of the quantum field theory times some increasing function of the cut-off mass, M. The increasing function may be either a positive power of M or a logarithm of M. But the effects of the cut-off at very short distances is just to redefine or 'renormalise' the parameters which describe the physics. Thus each correction term proportional to M^k or $\log M$ is associated with some physical parameter with the same mass dimension as the correction. In particular, only those combinations with the dimensions of one of the possible parameters actually appear. We have to worry about only a finite number of M dependent corrections with positive mass dimension, because there are only a finite number of parameters with positive mass dimension for such corrections to renormalise. But there could be an infinite number of dangerous terms with negative mass dimension.

Notice that if all the parameters have positive or zero mass dimension, so will all of the dangerous quantum correction terms. This is precisely the situation in which there are only a finite number of dangerous terms, all of which can be absorbed into a finite number of renormalised physical parameters. Thus, theories in which all the physical parameters have zero or positive mass dimension are renormalisable. But if any of the

The logarithmic function, $\log M$, is something that increases with M more slowly than any power (even more slowly than a fractional power like the square root, $M^{\frac{1}{2}}$). Physicists and mathematicians like to use what is called a natural logarithm that is defined by $e^{\log M} = M$, where e is an irrational number about equal to 2.71828. A graph of this function is shown in figure 16.2. Here is a physical picture of the log function. Suppose that you are trying to stop your car but your brakes are wearing out so that they never bring you to a complete stop. Instead, the speed of the car at time T (measured in hours) is 1 km divided by T (this gives the speed in kilometres per hour). When T is 1 h, your speed is $1\ km\ h^{-1}$. When T is 2 h, your speed is $0.5\ km\ h^{-1}$. When T is 10 h, your speed is $0.1\ km\ h^{-1}$. And so on. How far does your car travel between time T equals 1 h and a later time T? The answer (in kilometres) is $\log T$. From the graph, you can read that by $T = 2$, one hour after $T = 1$, the car has gone about 0.7 km. After two hours at $T = 3$, it has gone about 1.1 km. The car never stops, so the distance it travels is increasing, no matter how large T is. But the rate of increase keeps going down because the car keeps slowing down. For T less than one hour, the distance travelled is negative (as shown in the graph) because the car has not yet got to where it will be at $T = 1$.

Figure 16.2. A graph of the natural logarithmic function.



parameters have negative mass dimension, the corrections can have *arbitrarily negative mass dimension* because the parameters with negative mass dimension can appear many times in the product. Thus you can never stop with a finite set of such parameters. The cut-off dependent quantum corrections will always have pieces with more negative mass dimension than any parameter in your set. Only with an infinite number of physical parameters can you absorb all the cut-off dependence. This is why theories with parameters with negative mass dimension are not renormalisable. Thus, amplitudes whose coefficients have dimensions of inverse mass to a power are called 'nonrenormalisable interactions' because they destroy renormalisability.

It is possible to define what you mean by a theory that is not renormalisable by establishing a set of arbitrary rules for dealing with the cut-off dependence. But that doesn't help. If you construct, in this way, a theory with a finite set of parameters, some of which have negative mass dimension, you are just fooling yourself. The theory really depends on an infinite number of physical parameters but you have fixed all but a finite number of them according to an arbitrary prescription. There is no reason to expect such a construction to have anything to do with the world.

16.8 The sociodynamics of particle theory

So far, I have been discussing history which I read about in books, or learned from talking to my older colleagues. In 1947, while quantum electrodynamics was being created, I was busy being born in the post-war baby boom. But, by the mid 1960s, I was an undergraduate at Harvard, getting interested in particle physics myself. By this time, things had deteriorated to the point that, at Harvard, no course in quantum field theory was taught! Julian Schwinger, one of the heros of quantum electrodynamics and soon to be a Nobel prize winner, had given up on quantum field theory in favour of what he felt was a more phenomenological formalism which he taught in a course called 'relativistic quantum mechanics'. He convinced me at the time because he was a masterful lecturer and I was an impressionable undergraduate. But I now believe that he was pulling the wool over my eyes. Relativistic quantum mechanics *is* quantum field theory, properly defined. Schwinger gave up too early.

At any rate, though I didn't realise it at the time and didn't fully appreciate it until 1971, the seeds of the explosive 1970s were being sown all around me during the comparatively boring 1960s.

This may be a good time to tell you my theory about how theoretical particle physics works as a sociological and historical phenomenon. The progress of the field is determined, in the long run, by the progress of experimental particle physics. Theorists are, after all, parasites. Without our experimental friends to do the real work, we might as well be mathematicians or philosophers. When the science is healthy, theoretical and experimental particle physics track along together, each reinforcing the other. These are the exciting times. But there are often short periods during which one or the other aspect of the field gets way ahead. Then theorists tend to lose contact with reality. This can happen either because there are no really surprising and convincing experimental results being produced (in which case I would say that theory is ahead – this was the situation in the late 1970s and early 1980s, before the discovery of the W and Z) or because the experimental results, while convincing, are completely mysterious (in which case I would say that experiment is ahead – this was the situation during much of the 1960s). During such periods, without experiment to excite them, theorists tend to relax back into their ground states, each doing whatever comes most naturally. As a result, since different theorists have different skills, the field tends to fragment into little subfields. Finally, when the crucial ideas or the crucial experiments come along and the field regains its vitality, most theorists find that they have been doing irrelevant things. But the wonderful thing about physics is that good theorists don't keep doing irrelevant things after experiment has spoken. The useless subfields are pruned away and everyone does more or less the same thing for a while, until the next boring period.

This theory explains, I hope, how I can say that the 1960s were boring despite the fact that many of the pieces of the puzzle of the $SU(3) \times SU(2) \times U(1)$ theory of strong and electroweak interactions were discovered in the 1960s (and even the 1950s). There were people, like Feynman, Gell-Mann, Glashow, Weinberg, Ken Wilson and others, who had many of the right ideas all along, but they were isolated islands in a sea of confusion, unable to convince everyone else that what they were doing was right, and frequently unable even to convince themselves. After all, the original papers of Glashow and Weinberg on the $SU(2) \times U(1)$ theory were pretty much ignored even by the authors themselves until 1971. It wasn't obvious that the theory was renormalisable because it still apparently contained interactions proportional to a dimensional parameter, the inverse of the W mass.

While the decade of the 1960s was a difficult time to live through, it is fascinating to look back on. Other articles in this collection will go into this history in more detail. What I find particularly amusing about it is the peculiar interplay between the attempts to understand the weak interactions and the attempts to understand the strong interactions. These had to be closely connected because many of the particles which decay by weak interactions are produced by, or otherwise participate in, the strong interactions. But, in fact, the interplay was much more subtle. Many of the ideas that were developed in an attempt to understand the strong interactions eventually found their most important application to the weak interactions instead and vice versa. I regard this curious historical fact as significant and mysterious. The crucial concepts that were developed during this period include the renormalisation

group, gauge theories, flavour SU(3) symmetry, quarks, charm, colour, spontaneous symmetry breaking, scaling, the parton model, and some others that I will discuss later. All of these are essential components of our present understanding. It is not necessary that you know in more detail what all of these words mean (you can learn more about them in some of the other articles in this collection). But what I do want to convey is the sense that during the 1960s, while we had almost all the pieces of the puzzle of the strong and weak interactions, they were scattered among a myriad of other ideas which would soon be forgotten. Missing were the crucial theoretical insights and experimental results needed to prune away the dead wood.

16.9 Spontaneously broken gauge theories

The missing theoretical idea, provided by 'tHooft in 1971, was the demonstration that spontaneously broken gauge theories can be renormalisable. In particular, he showed the renormalisability of the $SU(2) \times U(1)$ gauge theory written down by Glashow and Weinberg and Salam in which the weak interactions are due to the exchange of W and Z particles. The W and Z are gauge bosons like the photon, but heavy because of spontaneous symmetry breaking. The main reason that this was nontrivial is that, in gauge theories, the symmetry and the dynamics are intertwined in such a way that you need the symmetry to perform the renormalisations. But the physical states in the theory do not respect the symmetry because the vacuum breaks it spontaneously. Thus the absorption of the infinites into the physical parameters does not work in any simple way. But what 'tHooft, Ben Lee and a few other people realised was that the important thing in the renormalisation programme was to get rid of the dependence of physical results on the physics at very short distances. This should have nothing to do with the spontaneous symmetry breaking which gives mass to the W and Z because the spontaneous symmetry breaking affects the structure of the physics only at distances longer than the scale at which the symmetry is broken, in this case the Compton wavelength of the W and Z. With the help of Veltman, 'tHooft was able to disentangle the symmetry from the dynamics. They also introduced a new renormalisation scheme in which the short distance effects are simply thrown away by a completely automatic procedure that makes no explicit reference to the physical parameters. With these, and a few other, bits of cleverness he was able to show that spontaneously broken gauge theories are renormalisable.

There are many nice things about 'tHooft's scheme. One thing I liked very much (when I finally understood it) was the fact that in 'tHooft's procedure the M^n divergences do not even have to be thrown out. They never show up at all. In fact, I think that this is good because these divergences have no physics in them. They can be renormalised away completely, so it is nice that they never appear at all. On the other hand, the logarithmic divergences are interesting because they must be associated with a log of some other dimensional parameter, a momentum or mass. This is dimensional analysis again. The log function is dimensionless. You can find the log of a dimensionless number unambiguously, but the log of a mass depends on the units in which the mass is measured. Thus the logarithmic divergences must have the form:

$$\log (M/m),$$

where m is some physical parameter with the dimensions of mass so that M/m is dimensionless and independent of the choice of units. m may be the mass or energy or momentum of one of the particles involved in the process you are calculating. When the logarithmic divergence is absorbed, the rest of the logarithm remains. Thus renormalisation requires a dimensional parameter to set the scale of the logarithm. This is called the renormalisation scale μ. It is arbitrary in principle, but in fact, for any given calculation, some renormalisation scales are more convenient than others because of the presence of logs of m/μ. If all the masses and momenta in a process are of the same order of magnitude, it pays to choose μ in the same range to minimise the effects of the logs and make the perturbation theory better behaved. It is this logarithmic dependence on the renormalisation scale which is responsible for the renormalisation group dependence of parameters on the distance or momentum scale, first discussed by Gell-Mann and Low. In the renormalisation group, you use the physics at one scale μ to figure out what the physics will look like at a very nearby scale. But by putting together many of these small steps, you can understand how the physics changes under large changes of the scale.

We now know that all this solved the problem of the weak interactions. Fortunately, it wasn't obvious in 1971 because it wasn't clear that the Glashow–Weinberg–Salam $SU(2) \times U(1)$ theory was the right one. So theorists had a good excuse to explore the vast new class of renormalisable theories which 'tHooft had opened up for us. Five years later, the experimental evidence had settled down to the point where we could be confident that $SU(2) \times U(1)$ was right, but, in the mean time, a great deal was learned about the properties of the new kind of theory.

16.10 Scale dependence

One of the most important properties was discovered by Sidney Coleman and Eric Weinberg. They thought about quantum field theories which naïvely have no dimensional parameters and only a single dimensionless parameter, such as quantum electrodynamics with a zero mass electron. They realised that, because the renormalisation scale μ has to be introduced to define the quantum theory, the physics is actually determined by a dimensional parameter instead of a dimensionless one. The point is that the renormalised dimensionless coupling is a

function of the renormalisation point, but by dimensional analysis it must therefore be a function of μ/Λ, where Λ is some fixed dimensional parameter. Thus

$$\alpha(\mu)=f(\mu/\Lambda).$$

Furthermore, the μ dependence of $\alpha(\mu)$ is determined by the theory. Therefore f is some fixed, computable function. The Λ is the only thing that can actually be varied. They called this process, in which a dimensionless parameter α is traded for a dimensional parameter Λ, 'dimensional transvestism'. This was too much for the editors of the *Physical Review*, who consider it part of their job to keep the language of particle physics as boring as possible, so the effect is now called 'dimensional transmutation'.

In quantum electrodynamics, the coupling constant is an increasing function of μ, which looks approximately like

$$\alpha(\mu)=b/\log(\Lambda/\mu)$$

for some constant b. Since $\alpha=e^2/\hbar c$ must be greater than zero, this only makes sense for Λ greater than μ where the log is positive. Indeed, the quantum electrodynamics theory probably only makes sense, in principle, for μ less than Λ. This used to worry some people. It no longer bothers us in quantum electrodynamics. Because α is quite small at ordinary scales, the log must be big. That means that Λ is a truly enormous mass. Since we do not believe that quantum electrodynamics is a complete theory of the world, we don't worry too much that it doesn't make sense at energies much larger than anything that we care about. However, it was thought, at one time, that all interesting quantum field theories behave as quantum electrodynamics does. That is they are sick at very short distances. But David Politzer and others showed that this is not true. Gauge theories based on groups such as SU(2) and SU(3) (which are called non-Abelian because their group multiplication laws are not commutative, see my chapter on GUTs) have exactly the opposite property. Asymptotic freedom! In these theories, the coupling constant is again dimensionless, and again dimensional transmutation occurs so that the actual parameter which determines the physics is a mass Λ. But here the coupling decreases with scale:

$$\alpha(\mu)=b/\log(\mu/\Lambda).$$

This makes sense only for μ greater than Λ. For scales much larger than Λ the theory can be simply described in terms of the gauge couplings. But, for scales of order Λ and smaller, the description of the theory in terms of perturbation theory in the coupling α no longer makes sense. In this region, the character of the theory must change in some way.

In the colour SU(3) theory of the strong interactions, quantum chromodynamics, we believe that the change in the character of the theory at large distances is associated with colour and quark confinement. The Λ parameter in quantum chromodynamics is the length scale at which confinement becomes important. At shorter distances, the theory can be described accurately in terms of the interactions between quarks and gluons (the gauge particles of quantum chromodynamics, like the photon in quantum electrodynamics). But at longer distances this picture breaks down. The colour force between quarks, which behaves much like electromagnetism at distances shorter than $1/\Lambda$, does not drop off further for distances longer than $1/\Lambda$. This was not at all obvious before 1973, because most of our understanding of the strong interactions was based on experiments at long distances, where the nature of the physics is dominated by confinement and the underlying dynamics is obscured.

16.11 Grand Unified Theories

Once we understood $SU(2)\times U(1)$ and quantum chromodynamics, Grand Unified Theories were a simple step. The motivation for the simplest GUT, SU(5), was not any mystical desire to follow in Einstein's footsteps and unify everything. Shelley Glashow and I were just trying to understand $SU(2)\times U(1)$ better. For several years, we had realised that if we could incorporate the $SU(2)\times U(1)$ gauge symmetry into a single simple group it would give us some extra information. It would fix the value of the weak mixing angle, a free parameter in the ordinary $SU(2)\times U(1)$ theory and it would explain why all the electric charges we see in the world are multiples of the charge of the electron. But we were having great difficulty doing it. The quarks never seemed to fit in properly. When he heard about quantum chromodynamics, Shelley suggested that we might have to incorporate colour to get it to work. When I pointed out to him that the strong interactions are strong, he replied that we only know that they are strong at long distances where confinement is important. At sufficiently short distances, they could be as weak as $SU(2)\times U(1)$. With that hint, it was easy for me to find the SU(5) theory, into which SU(3) colour and electroweak $SU(2)\times U(1)$ fit very neatly, basically because $2+3=5$. Of course, this theory also predicted proton decay, but that just meant the scale at which the SU(5) symmetry was broken had to be extremely large.

A few months later, Helen Quinn, Weinberg and I figured out how to actually calculate the scale of SU(5) breaking in the simplest SU(5) model and we discovered that the scale really is very large, about 10^{14-15} GeV.

16.12 Effective field theories

What I want to emphasise about all this is the following. In our understanding of the weak interactions and strong interactions based on $SU(2)\times U(1)$ and SU(3) and in the attractive speculation of GUTs, based on SU(5), there is a crucial role played by dimensional parameters, the confinement scale of quantum chromodynamics and the breaking scales of $SU(2)\times U(1)$ and SU(5). This was a dramatic change from quantum elec-

trodynamics in which the physics seemed (at least to the naïve observer) to be mostly in the dimensionless coupling constant α. Furthermore, we were getting used to incorporating physics at short distances (such as GUTs) without disturbing our understanding of physics at longer distances. Many physicists began to verbalise and answer a question which had been nagging at them for a long time. If there is all this wonderful stuff going on at short distances, how come quantum electrodynamics worked so well? Of course, the reason is obvious and had been known, in some sense, for a long time. Quantum electrodynamics works extremely well for the electron because the distances at which other stuff is happening are very small compared to the electron's Compton wavelength. It was easy to see this explicitly in theories such as $SU(2) \times U(1)$ in which quantum electrodynamics was embedded in a more complicated but still renormalisable theory at a smaller distance scale. But physicists were slow in appreciating the full power of the idea, which is unleashed only when quantum electrodynamics is thought of as an 'effective field theory', approximately valid at long distances.

The point is this. At distances of the order of the electron Compton wavelength, the only particles we really have to think about are the electron and the photon. All other charged particles are heavier, and, at such large distances, there is not enough energy to produce them, so we do not have to include them in our theory. There are light neutral particles, neutrinos, but at these distances they are so weakly interacting that they don't matter much, so we can ignore them as well. Thus we can describe the electron–photon interaction at these large distances by an effective field theory involving only the electron and the photon. This has to work. With a completely general quantum field theory, we can describe the most general possible interactions consistent with relativity, quantum mechanics and causality. We do not give up any *descriptive* power by throwing out the heavier particles and going to an effective theory.

It might seem, though, that we have given up *predictive* power. After all, an arbitrary effective theory has an infinite number of nonrenormalisable interactions and thus an infinite number of parameters. But this is not quite right for two reasons, one quantitative and one qualitative. Quantitatively, if we know the underlying theory at shorter distances, we can calculate all the nonrenormalisable interactions. Indeed, there is a straightforward and useful technology for performing these calculations. Thus quantitative calculations can be done in the effective theory language.

The qualitative message is even more interesting. All of the nonrenormalisable interactions in the effective theory are due to the heavy particles which we have ignored. Therefore, the dimensional parameters that appear in the nonrenormalisable interactions in the effective theory are of the order of the heavy particle masses. If these masses are all very large compared to the electron mass and the photon and electron energies, the effects of the nonrenormalisable interactions will be small. They will be suppressed by powers of the small mass or momenta over the large masses.

Thus, not only do we not lose any quantitative information by ignoring the heavy particles and going to an effective field theory language, but we gain an important qualitative insight. When the heavy particle masses are large, the effective theory is approximately renormalisable. It is this feature that explains the success of renormalisable quantum electrodynamics.

To extract the maximum amount of information from the effective theory with the minimum effort, we should renormalise the theory to minimise the logarithms that appear in perturbation theory. We can do this by using 'tHooft's scheme and choosing the renormalisation scale, μ, appropriately. If all the momenta in a process of interest are of order μ, there will be no large logarithms. The standard techniques of the renormalisation group can be used to change from one μ to another as required.

In the extreme version of the effective field theory language, we can associate each elementary particle Compton wavelength with a boundary between two effective theories. For distances larger than its Compton wavelength, the particle is omitted from the theory. For shorter distances, it is included. The connection between the parameters in the two effective theories on either side of the boundary is simple. They must be related so that the description of the physics just below the boundary (where no heavy particles can be produced) is the same in the two effective theories. These relations are called 'matching conditions' for obvious reasons. They are calculated with μ equal to the mass of the boundary particle to eliminate large logs.

If we had a complete renormalisable theory at infinitely short distances, we could work our way up to the effective theory at any larger distance in a totally systematic way. Starting with the mass M of the heaviest particles in the theory, we could set $\mu = M$ and do the matching to find the parameters of the effective theory with the heaviest particles omitted. Then we could use the renormalisation group to scale μ down to the next heaviest mass and repeat the matching calculations to produce the next effective theory. *And so on!* In this way we get a tower of effective theories, each with fewer particles and more small nonrenormalisable interactions than the last. We simply have to continue this procedure until we get to the large distances in which we are interested.

There is another way of looking at it, however, which corresponds more closely to what we actually do in studying physics. We can start at long distances and try to build up each member of the tower of effective theories stretching down to arbitrarily short distances only as it becomes relevant to our understanding of physics. In this view, we do not know what the renormalisable theory at short distances is, or even that it exists at all. In fact, we can dispose of the requirement of renormalisability altogether and replace it with a condition on the nonrenormalisable interactions in the effective theories. The condition is this:

> In the effective theory which describes physics at a scale μ, *all the nonrenormalisable interactions must have dimensional couplings* less than $1/\mu$ to the appropriate power. If there are nonrenormalisable interactions with couplings $1/M$ to a power, for some M greater than μ, there must exist heavy particles with a mass m less than or about equal to M that produce them. In the effective theory including the heavy particles, the nonrenormalisable interactions must disappear.

Note that an effective field theory, like any nonrenormalisable theory, depends on an infinite number of parameters (which are related at shorter distances). But the above condition insures that only a finite number of them are actually important in any physical situation because all the nonrenormalisable interactions are suppressed by powers of μ/M where μ is less than M. Thus, as we go down in distance through the tower of effective field theories, the effects of nonrenormalisable interactions grow and become interesting on the boundaries between theories, at which point they are replaced by renormalisable (or, at least, less nonrenormalisable) interactions involving heavy particles.

This condition on the effective theories is, I believe, a weaker condition than renormalisability. One can imagine, I suppose, that the tower of effective theories goes down to arbitrarily short distances in a kind of infinite regression. This is a peculiar scenario in which there is really no complete theory of physics – just a series of layers without end. More likely, I think, the series does terminate, either because we eventually come to the final renormalisable theory of the world, or (most plausible of all) because, at some very short distance, the laws of relativistic quantum mechanics break down and an effective quantum field theory is no longer adequate to describe the physics.

Renormalisability is still very important. When the gap between two neighbouring mass scales is large, the effective field theory near the lower scale is approximately renormalisable, because the nonrenormalisable interactions have a very small effect. This is the situation in quantum electrodynamics near the electron mass scale. But we no longer have to assume that renormalisability is a fundamental property.

In this picture, the presence of infinities in quantum field theory is neither a disaster, nor an asset. It is simply a reminder of a practical limitation – we do not know what happens at distances much smaller than those we can look at directly.

Whatever happens at short distances, it doesn't affect what we actually *do* to study the theory at the distances we can probe. We have purged ourselves of the hubris of assuming that we understand infinitely short distances. This is the great beauty of the effective field theory language.

16.13 Dollars and direction

Some of you are probably saying to yourselves, by this time, that the whole idea of effective field theories is rather simple and obvious, so why have I subjected you to an article on the subject? One reason is that it makes a difference in dollars. Since our understanding of physics is organised by distance scale (we understand physics at distances greater than 10^{-16} cm, but not at smaller distances), we must push for experimental information at short distances. But short distances mean large energies which mean larger and more expensive accelerators. More specifically, our general understanding of the connection between small nonrenormalisable effects and the heavy particles which produce them is important in the planning of future accelerators.

But I have another reason for talking about effective field theories. As I suggested at the beginning of this chapter, I am somewhat concerned about the present state of particle theory. The problem is, as I mentioned before, that we are in a period during which experiment is not pushing us in any particular direction. As such times, particle physicists must be especially careful.

We now understand the strong, weak and electromagnetic interactions pretty well. Of course, that doesn't mean that there isn't anything left to do in these fields any more than the fact that we understand quantum electrodynamics means that there is nothing left to do in atomic physics. The strong interactions, quantum chromodynamics, in particular will rightly continue to absorb the energies of lots of theorists for many decades to come. But it is no longer frontier particle physics in the sense that it was fifteen years ago.

What then is there to do? If we adopt the effective field theory point of view, we must try to work our way down to short distances from what we know at longer distances, working whenever possible in the effective theory which is *appropriate* to the scale we are studying. We should not try to guess the ultimate theory at infinitely small distances. Even if we could do it, it would probably be about as useful as explaining biology in the language of particle physics. This seems to me to be an extremely important bit of common sense, a useful antidote to the Einstein complex (that is a desire to work on difficult and irrelevant theoretical questions just because Einstein did it) to which most theoretical particle physicists are very susceptible.

Thus, for example, one subject that certainly deserves the attention of all theorists is the question of what causes the spontaneous breaking of the $SU(2) \times U(1)$ symmetry of the electroweak interactions. This is the physics of the *next* effective theory and it *will* be explored by experiment in the near future, if we have the strength and the will to build the superconducting supercollider and push the experimental frontier to the next

scale at energies many tens of thousands of times higher than the proton mass!

It is not so obvious that GUTs are interesting things to study. Some years ago, in a panel discussion, Feynman presciently asked me what I would think about SU(5) if proton decay was not observed at the predicted level. In my youthful enthusiasm, I replied that I would believe that it is right anyway. It is too pretty to be wrong. I think that I still believe that. But what I didn't see at the time was that SU(5) or closely related GUTs could be right but not very interesting. If proton decay is actually observed, they become extremely interesting. But until then, apart from a few numbers which express the relations between parameters in our low energy world which follow from unification, their only connection with reality is through cosmology. Cosmology is fun, but it seems unlikely to me that we will know enough about it to extract much quantitative information about physics at very short distances, at least not anytime soon.

I am particularly suspicious of attempts to guess the structure of physics below the Planck length (the length at which quantum gravitational effects are expected to become important, about 10^{-33} cm). If there is any scale at which we might expect quantum field theory to break down, *this is it*, because there is no satisfactory quantum theory of gravity based on conventional relativistic quantum mechanics in ordinary space-time. Indeed, most of the popular theories (such as Kaluza–Klein theories or string theories) assume that physics changes rather dramatically here and that space-time actually has more than four dimensions.

Apparently, the mathematics of these ideas is so appealing that no one is immune. Steve Weinberg, one of the heros of the effective field theory idea, has become so wrapped up in it that he came to Harvard recently to give us a series of talks on differential geometry. I was so moved that I composed the following poem for the occasion:

> Steve Weinberg, returning from Texas
> brings dimensions galore to perplex us.
> But the extra ones all
> are rolled up in a ball
> so tiny it never affects us.

One problem with all this, of course, is that 'it never affects us'. These theories probably have no experimental consequences at all in the practical sense, because we will never probe small enough distances to see their effects. But there is another subtler objection to this kind of speculation. Once you start relaxing the assumptions of relativistic quantum mechanics, where do you stop? In practice, theorists have considered only theories which they happened to know something about for purely accidental historical reasons. That does not seem to me to be a good enough reason to look at them. Theoretical physics must be more than an historical accident.

My personal suspicion is that Nature is much more imaginative than we are. If we theorists approach her study with the proper respect, if we recognise that we *are* parasites who must live on the hard work of our experimental friends, then our field will remain healthy and prosper. But if we allow ourselves to be beguiled by the siren call of the 'ultimate' unification at distances so small that our experimental friends cannot help us, then we are in trouble, because we will lose that crucial process of pruning of irrelevant ideas which distinguishes physics from so many other less interesting human activities.