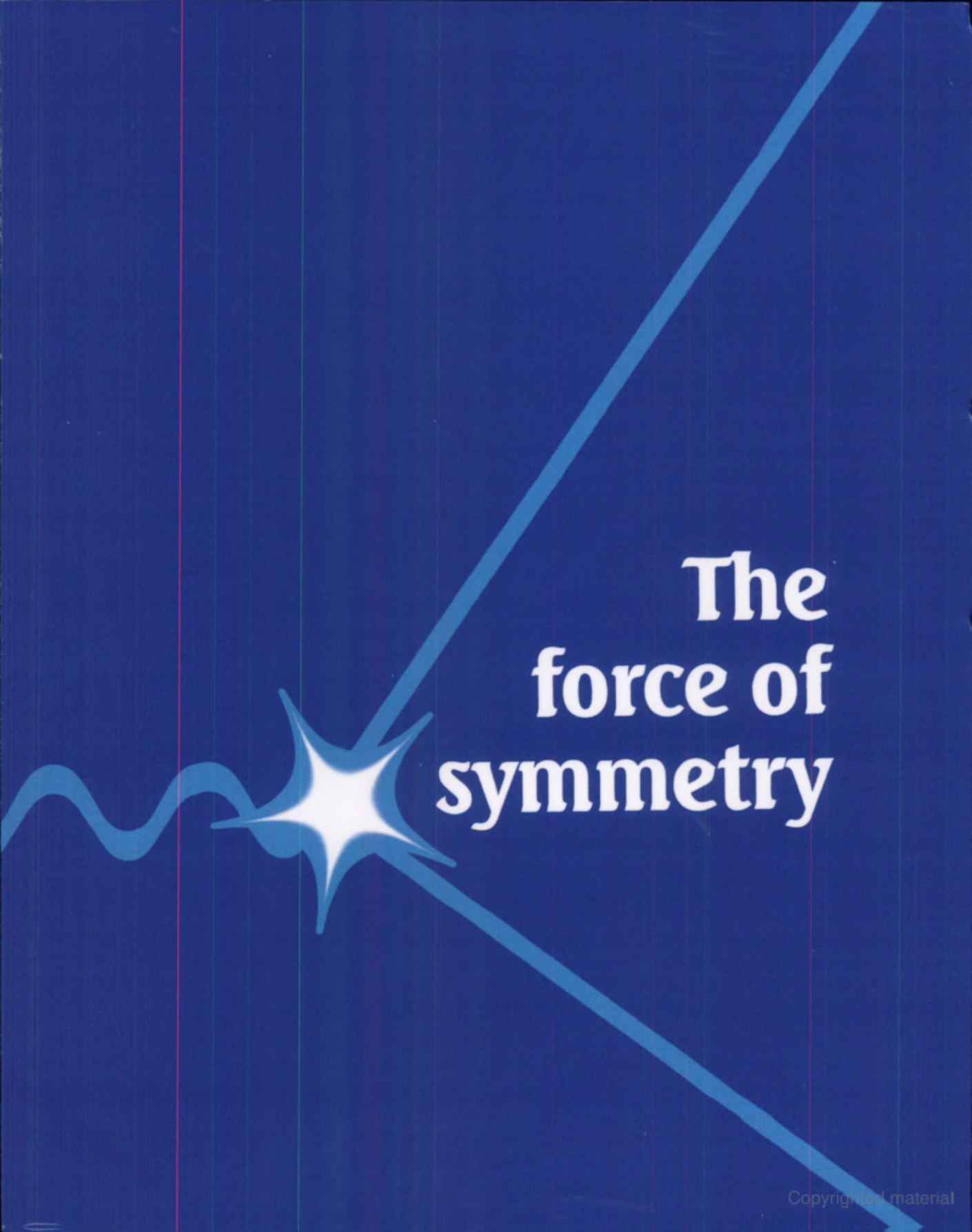

THE FORCE OF SYMMETRY

VINCENT ICKE
Leiden University

'But if anybody says he can think about quantum problems without getting giddy, that only shows that he has not understood the first thing about them.'

Niels Bohr

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge CB2 1RP, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, United Kingdom
40 West 20th Street, New York, NY 10011-4211, USA
10 Stamford Road, Oakleigh, Melbourne 3166, Australia

First published 1995

Reprinted with corrections 1997

Typeset in Monotype Times [TAG]

A catalogue record of this book is available from the British Library

Library of Congress Cataloguing in Publication data

Icke, Vincent.
The force of symmetry / Vincent Icke.
p. cm.
ISBN 0-521-40495-9. – ISBN 0-521-45591-X (pbk.)
1. Symmetry (Physics). I. Title.
QC174.17.S9I25 1994
539–dc20 94-26237 CIP

ISBN 0 521 45591 X paperback

Transferred to digital printing 1999

TAG

1

A matter of force

1.1 The law of inertia

The way the world works is mostly the way things move: *Where is what when*? is just about the most basic question one can ask about the Universe. Everyday experience gives us a rough-and-ready answer: the motion of matter is governed by forces. A puck may lie still on the ice until it is struck with a stick, after which it glides straight along until it hits something else. Without being struck, bumped, caught, or otherwise interfered with, it will follow its own path.

This description is horribly vague. On the ice, the puck moves with very nearly constant speed in a straight line. But the same object, struck in the same way, moves very differently on the pavement: almost as soon as the blow that sets it in motion is over, the puck lies still again. At the very least, then, it is unclear what an object's true path is: the smooth gliding along the ice, or the state of rest on the pavement, or what?

We cannot specify what we mean by 'force' until we have specified what ideal state of motion that force is supposed to perturb. Some four centuries ago, it was generally assumed that motion with constant speed along a circle is the ideal motion that can maintain itself indefinitely without external influence. This idea (although we now know it to be wrong) is not in itself absurd: we cannot deduce from first principles whether or not uniform circular motion is ideal in the above sense. Indeed, if we were to build our own universe, maybe we could arrange it that way.

But we *observe* that in the actual Universe a circular motion cannot maintain itself indefinitely, as is evident in the operation of a slingshot (Fig. 1.1). In another universe, constructed along the lines of the classical philosophers, maybe David's stone would have continued to buzz around in a circle, rather than fly off on a tangent to strike Goliath down.

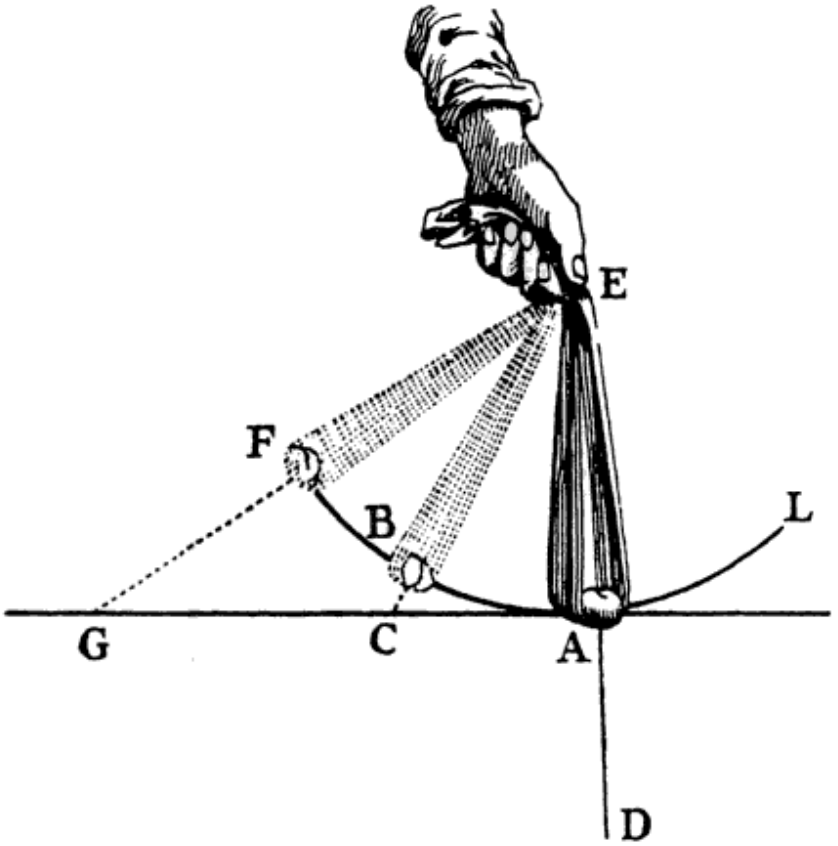

Fig. 1.1 Engraving illustrating the action of a slingshot, showing the law of inertia at work: the stone, when released from the constraining force of the sling, moves away in a straight line. (From Descartes's *Principes de la Philosophie*, part two, Art.39.)

A major advance towards today's theories of Nature was made by Descartes, who formulated what we now know as the Law of Inertia: *a piece of matter moves in a straight line with constant speed, unless a force acts on it*. This introduces force as something that causes the state of motion of an object to become different from the ideal constant-velocity motion. The pivotal question is then: what is that something?

Descartes himself thought that a force comes about through the immediate physical contact between objects. To anyone who limits the observation of Nature's workings to an occasional billiards game, this idea is self-evident. Descartes knew that the motions of the planets must be governed by some sort of force, even though they do not seem to be in contact with anything (pinned to crystalline spheres, or whatever). Planetary orbits are *curved*, so the law of inertia implies that there is a force which acts on them constantly.

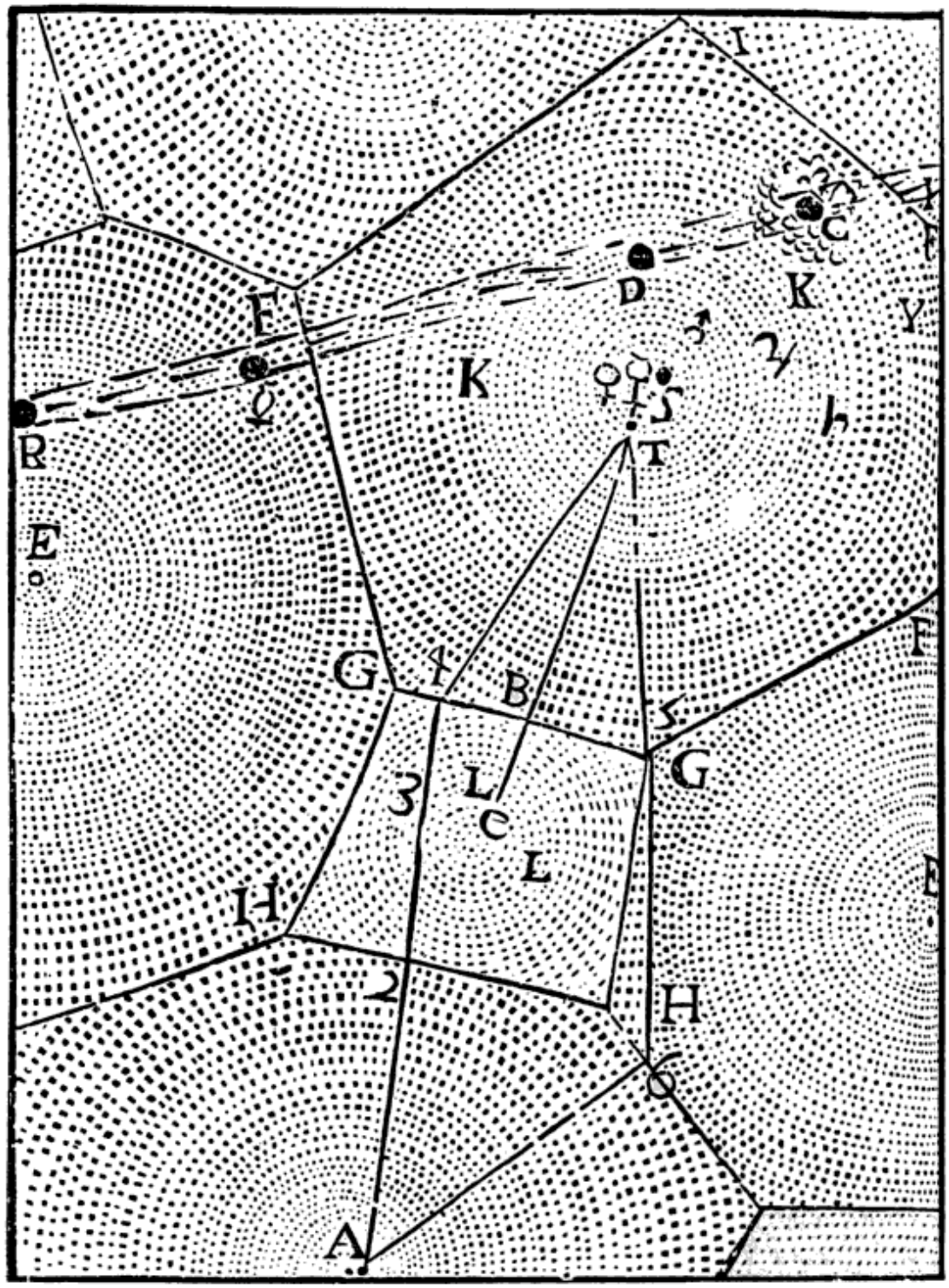

Fig. 1.2 Vortices in the 'subtle matter' or aether which, according to Descartes, was responsible for transmitting the forces that keep the universe together. (From *Le Monde, ou Traité de la Lumière.*)

Accordingly, Descartes assumed that all of space was filled, to the complete exclusion of emptiness, with a novel kind of substance, a subtle matter that transmitted the forces in the Solar System through direct contact. This matter was supposed carry the celestial bodies around in its swirling eddies (Fig. 1.2). In the second part of his *Principes de la Philosophie*, Article 16, Descartes wrote:

> ... from the sole fact that an object is extended in length, height and depth, we have reason to conclude that it is a substance, [and so] we must conclude the same about space which is supposedly empty: namely that, because it possesses [spatial] extent, it also has substance.

In other words: space has physical attributes, namely its three dimensions, so it must be regarded as real stuff. This powerful notion lay hidden for three hundred years, until it was rediscovered independently by Einstein.

One could make some objections of principle, for example that the constituents of Descartes's subtle matter must not have any internal structure, and hence must be infinitely small, wherefore – by Descartes's own definition – they have no spatial extent and consequently do not exist. Indeed, this problem had been spotted in antiquity, and gave rise to the hypothesis that the world is made of atoms: small particles that have no inner structure and therefore cannot be further divided, but that do have a finite extent. Of course, the atoms we now know do not fit that description at all, even though they have the same name. Later I will return in detail to the consideration of smallness.

Unfortunately for its inventor, the Cartesian hypothesis about the direct-contact origin of forces did not lead to calculable results. Descartes proposed that the planets are kept in their orbits by swirling, vortical motions in the subtle matter between them; but this assumption did not give a quantitative prescription for the behaviour of the force. Thus, nothing could be calculated; for example, Kepler's laws (which were known to describe planetary motion very well) could not be explained in terms of the motion of the subtle matter.

It was Newton who first realized that, for the description of planetary motions, it is not necessary to know what a force 'is', as long as one can give a precise description of what it does, i.e. formulate how it depends algebraically on physical quantities. In fact, it had already been pointed out by Hooke that Kepler's Third Law implies that the force between the Sun and a planet acts along the line connecting them and decreases in inverse proportion to the square of the distance between them. Newton extended this with the prescription that the force be proportional to the product of

the masses of the attracting bodies. This was an important advance, because it established a symmetry between the objects involved. It isn't as if one object, for example the Sun, is the boss that does all the attracting; in the Newtonian description of a force, both objects attract *each other*, so that we can truly speak of an interaction. Moreover, in keeping with Descartes's hint that there is only one force in Nature, Newton presumed that the dominant force in the Solar System (to be known as gravitation) is universal and acts between all objects.

The success of this approach is well known, and it has been praised beyond measure. And yet the Newtonian idea of force had some uncomfortable features. It was conceived to be an instantaneous interaction: a mutual working, at exactly the same moment, between two spatially separated objects. The instantaneous nature of the action was not, at first, recognized as a problem; but the objections to the 'action at a distance', across supposedly empty space, were loud from the beginning. Still, it worked. Planetary motions could be calculated; Kepler's laws were explained in terms of the force of gravity. The Cartesian hypothesis of direct contact was completely eclipsed by the Newtonian action at a distance.

It was clear even in Newton's days that there must be more forces in Nature than gravity alone. Whereas Descartes's forces could appear in many guises – attractive or repulsive – depending on the detailed workings of the subtle matter, the gravitational force is always attractive, and hence cannot make stable objects: all things always fall down, so to speak. Thus, all many-particle systems in our Universe, when acted upon by gravity alone, must inevitably collapse, even though this might require a very long time. The apparent solidity and stability of matter is proof of the existence of other forces, so there was scope for an extension of the Newtonian system by finding those forces. Some were found, such as the magnetic and electrostatic forces; the experiments of Cavendish and Coulomb even showed that the electrostatic force can be described by exactly the same mathematical form as the force of gravity. But nobody questioned the underlying concept of instantaneous action at a distance any more.

1.2 The speed of light

A dramatic and fundamental step forward was made by Maxwell. This advance was wholly within the Newtonian world view but, interestingly, was also one of the first nails in the coffin of that view. Maxwell showed that the forces of electricity and magnetism are not two totally different ones, but instead are two aspects of one force (albeit a more complex one), the

asking: 'If c is really invariant, then how does one light ray see another one move?' Well, if there is no conceivable superseagull that can fly in such a way that a light wave appears to be standing still around it, the answer must be: 'Any light ray sees any other one move with the speed c.'

If a light ray, travelling at speed c, encounters another one head-on, also travelling at c, then the above can be written symbolically (not algebraically!) as "$c+c=c$." Similarly, if two light rays that travel in the same direction see each other move with the speed c, we must, in some sense, likewise require that $c-c=c$. If it is really true that c is invariant, then we cannot escape the conclusion that $c+c=c-c=c$! *Most* remarkable: one would have expected $2c$, or 0, or something in between, depending on circumstances. But the Michelson–Morley experiment has shown that the speed c does *not* depend on circumstances.

Einstein, who first asked the question about the relative motion of light rays, had the courage and the insight not to reject $c+c=c$ out of hand as absurd; all it means is that the invariance of the speed of light compels us to accept that speed is a more complicated beast than we suspected. However, the implications are staggering: we can, in fact, add speeds in such a seemingly contradictory way, but only at the expense of a drastic revision of the classical concepts of space and time.

Because it is required that $c+c=c-c=c$, speed cannot be a simple algebraic number for which the normal rules of addition and subtraction hold, as in the case of money or apples in the market. We must reconsider what exactly is meant by the addition of speeds. Note that this is no cause for dismay; even in everyday experience, we deal with quantities for which two plus two does not always equal four. Your position on Earth is a case in point: if you walk two kilometres, then another two, and then two more, you are not necessarily six kilometres from your point of departure. In fact, if you have walked along an equilateral triangle, you are back where you started. If distances added like money, you could never mail a letter or walk the dog: you would never get back home.

1.3 Relativity and fields

Einstein showed how we ought to define addition in such a way that the addition of any two velocities leaves the speed of light invariant. In order to be able to do this, he had to abandon the idea that time and space can be measured independently: if speeds add in a curious way, then this is due to the underlying behaviour of the ingredients of speed, namely space and time (remember, speed = distance/time!) If space and time cannot

be measured independently of one another, then there is no such thing as universal time.

Subsequently, Einstein found a number of other results that are, to our intuition conditioned by always moving much slower than light about as bizarre as $c + c = c$. One of these results is that c is an absolute *maximum* speed: nothing can travel faster than light. This conclusion can be glimpsed from the above. If the addition of speeds is defined such that $c+c = c$, then we must also have $c+c+c = c$, and so forth: we can never exceed the speed of light. Notice, by the way, that I have hereby shown that c must be an upper limit because of the observed fact that the speed of light is always the same; I do *not* say that c is the maximum speed because I have tried hard to exceed it and have failed! The c limit is an inescapable consequence of the experimental fact of c invariance, and so we can prove that any attempt to exceed the speed of light must fail. Because c is finite and maximal, and because it is the same no matter what you do, it serves as a universal standard of speed. There is an *absolute* meaning to the expressions fast and slow: motion with a speed that is much smaller than the speed of light is slow, any other motion is fast.

Another result is that mass and energy are essentially the same. This, too, can be appreciated on the basis of the above. I will show this by means of a `space-time diagram`, which is a graphical summary of 'where is what when' in a particular case (Fig. 1.4).

Suppose that we have a particle that has no internal structure, and that is acted upon by one force only. We know that the more energy is transferred by the force to the particle, the faster it travels. But because c is the maximum speed, I can transfer all the energy I want, I will never exceed the speed of light. Then where did all that energy go? Recall from everyday mechanics that the transfer of a given amount of energy gives a large boost to the speed of a low-mass object, and a small speed increment to one with a high mass. Let us look at the behaviour of a football and a railway engine (Fig. 1.5) when each is given a standard kick, for example one delivered by the goalkeeper of the Dutch national team.

The ball is propelled to a high speed, which in a space-time diagram means that it lies on a line that is inclined strongly with respect to the time axis. But the same kick, delivered to a railway engine, has almost no effect; even if all eleven team members were to hurl themselves at the machine, its speed would barely differ from zero, and its space-time path would be practically parallel to the time axis. We summarize these experimental facts by saying that a football has a small mass and a railway engine has a large mass.

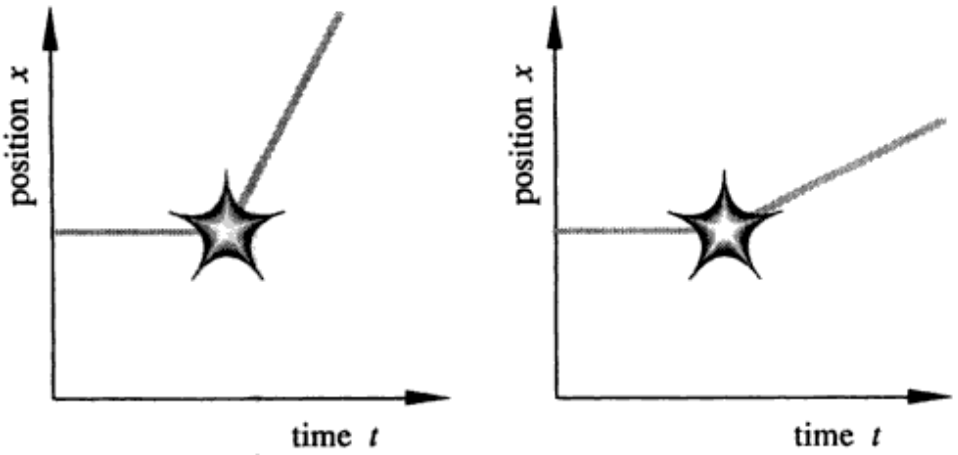

Fig. 1.4 Sequence of events of two different objects during a kick, as shown by their tracks in space-time. Such space-time diagrams occur frequently in this book, but the axes will usually be omitted. The convention I will use is that time runs to the right and spatial distance increases upwards on the page. This is different from what you usually find in the physics literature: there, time runs upwards on the page. That convention is typographically inept and I will almost never use it in this book.

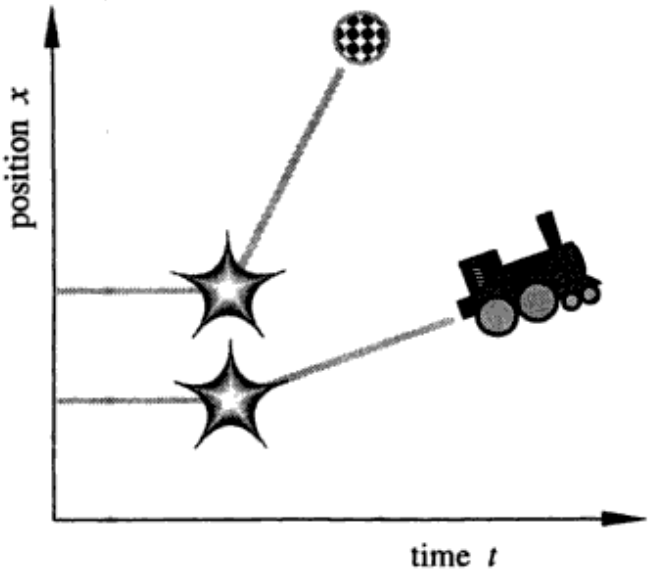

Fig. 1.5 Space-time diagrams of two different objects during a kick. The top track shows a football being kicked; its low mass shows up in a large change in velocity (i.e. a large acceleration, producing a sharp kink in the space-time track). The other track is not as sharply deflected, showing what happens when the same kick is delivered to a railway engine.

Now if c is to be a strict upper limit, the space-time paths of objects cannot be arbitrarily steep. If the football has a small speed, the kick delivered by our goalie makes a noticeable bend in its space-time path. But if the ball moves close to c, the same kick must make a much smaller bend: because c is a strict upper limit, the closer we get to the speed of light, the smaller the increase of the velocity becomes for a given addition of energy. Therefore, the more energy a football has, the more it resembles a railway engine: its energy acts as if it were mass! Because the c limit forces us to accept that energy is essentially the same as mass, we must also accept that even the mass m of an object at rest is equivalent to a certain quantity of energy E. A precise calculation of this equivalence leads to the famous $E = mc^2$.

These consequences of the experimental fact that the speed of light is the same for all observers are described precisely in the theory of `relativity`. In our discussion of forces, the facts of relativity have a very profound influence, mostly because of the findings that: first, the speed of light cannot be exceeded; second, mass is equivalent to energy; third, because a universal time does not exist, the order in time of events can be different for different observers; fourth, the speed of light can be kept constant only if we treat space and time on an equal footing. These points will be discussed in more detail later.

The facts and conclusions of relativity slash all support from under the Newtonian concept of instantaneous action at a distance. Because the speed of light is a maximum, there is no such thing as an instantaneous connection between spatially separated points. Accordingly, if there is to be any influence across a spatial distance, we must accept that that influence is underway for a while. Relativity prevents a force from acting instantaneously, so that there must be something that transmits it, some sort of messenger substance that carries the information about the action of the force from one point to another. This something is called a `field`, and because of the c limit we are compelled to accept the field as a physically real object, not merely as an aid to calculation. It is beginning to look after all as if we need some sort of direct contact to transmit a force (at this point, Descartes smiles).

Fields can have different forms, from very simple to very complex. We may imagine a field as follows: at every instant in time, each point in space is provided with a little label, on which we can read the strength of the force. When the field is simple, the label contains just one number ('scalar' field; Fig. 1.6). In a more complex case, the tag contains more numbers.

On this two-dimensional sheet of paper, I can represent a scalar field by a greyscale: the darker the picture, the stronger the field, and a single number (the percentage of paper covered by ink) describes the field at each

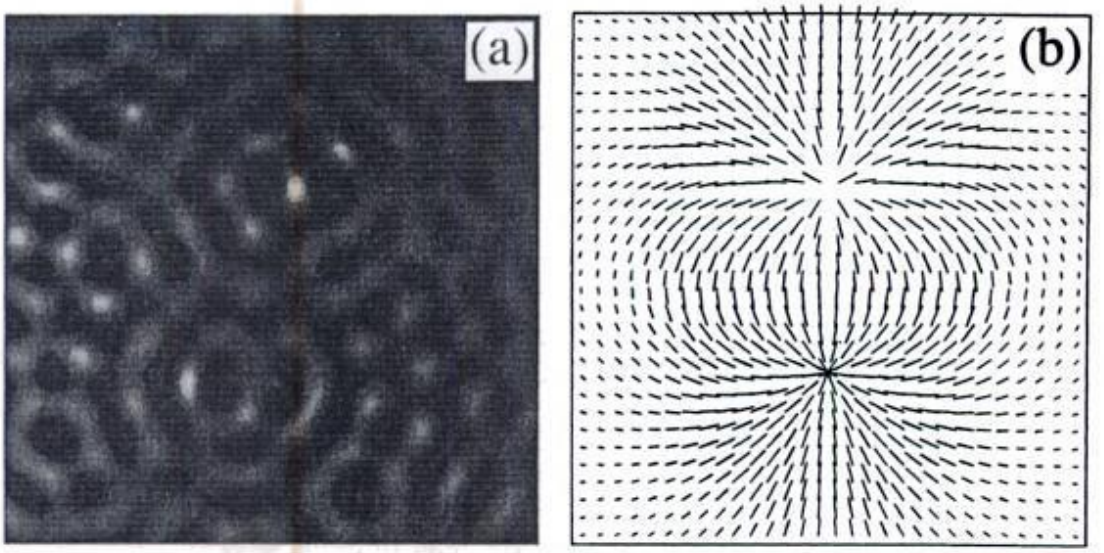

Fig. 1.6 (a) A scalar field. The field amplitude is indicated by a single quantity which, in this diagram, is represented by the intensity of the grey shading, as on a black-and-white photograph: the blacker the print, the stronger the field. (b) A vector field. Here, the field amplitude has two components: a strength and a two-dimensional direction. The strength is indicated by the length of the line segments, the direction by their orientation.

point. A somewhat more complicated field requires more numbers (called components) to describe it at each point. In two dimensions, I could represent a two-component field by a colour scale, where each component is shown as the percentage of paper covered by a different ink (say, blue for one component and yellow for the other – this would allow us to speak truly of 'green fields'). In practice, it is much clearer to represent the field at a given point by means of an arrow, which is why this type is called a *vector field*. The base of the arrow is the point where the field is measured; the directional angle is the first field component, and the arrow's length represents the second component. Thus, we obtain a field of arrows.

Complicated fields are more difficult to represent graphically, so that we usually prefer to work with an array of numbers that indicate the values of the various field components. Notice that we have a certain amount of leeway as to how we choose the components; for example, instead of taking the direction and the length of an arrow as the field components, we could have taken the lengths of the projections of the arrow in two different directions.

Fields are like weather maps. The temperature at ground level is a scalar field: single numbers all over the map suffice to specify it. The wind velocities

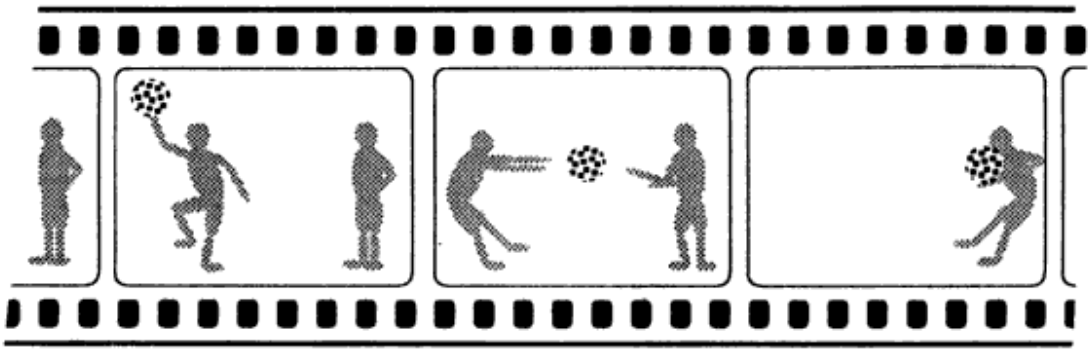

Fig. 1.8 Sequence of snapshots indicating the exchange of a heavy beanbag between two skaters standing on a frictionless surface.

A force appears upon the exchange of a field quantum as follows (Fig. 1.8). Let two skaters glide on a frictionless ice surface; one skater throws a beanbag at the other. This changes the velocity of the thrower; when the other skater catches the bag, the velocity of the recipient also changes.

An observer, far above the ice surface, does not see the beanbag, but does notice that the skaters change their velocities; hence the observer concludes that there is a force between them. In this analogy, we only see a repulsive force. Later we will note that the same exchange mechanism can produce attractive forces as well; in fact, repulsion is the exception and attraction is the rule.

Pictorially, the exchange event is described by a `Feynman diagram`. In such a diagram (Fig. 1.9) the three spatial dimensions have been collapsed into one (the vertical direction on the page). The dimension of time is represented by the horizontal direction (from left to right on the page). This layout is exactly the same as the one I used in the space-time diagrams discussing footballs and railway engines. The space-time tracks of the skaters are indicated by continuous lines, the track of the beanbag by a wavy line.

The point of the throw or the catch, where three lines come together, is called a `vertex`. A quantum is not a 'minimal' parcel of energy or matter or whatever. There's nothing quantized about a quantum: you can make it as big (e.g. high-energy) or as small (low-energy) as you like, and continuously to boot. It is the *interaction* which is quantized, in the sense that it is all-or-nothing. This is symbolized by the vertex in the Feynman diagrams.

You will get a clearer picture of what happens during the exchange of a field quantum if you transform the Feynman diagrams in this book into motion pictures. That can be done by making a Feynman diagram scanner

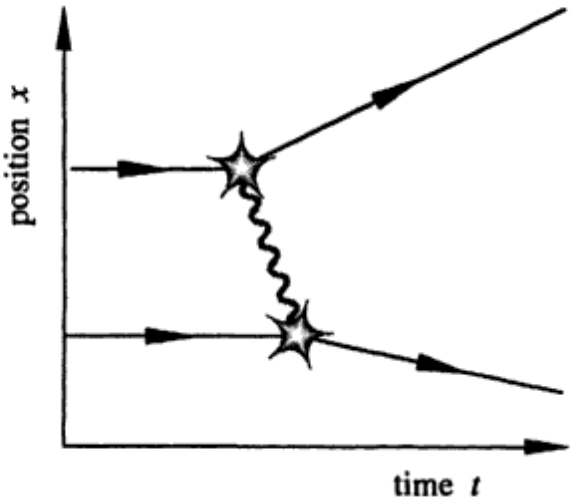

Fig. 1.9 First-order Feynman diagram. The snapshots in the preceding diagram can be derived from this space-time sequence by looking at what happens at various times. A snapshot corresponds to the situation on a vertical line in this diagram.

(Fig. 1.10). Take a piece of thin white cardboard, about 10 by 20 centimetres in size, and in the middle cut a 5 cm slot across it, with a width of a millimetre or so. Place the scanner with the slot vertically on the leftmost side of a Feynman diagram, and then slide it to the right. In the slot will appear the positions of the various quanta as time goes on. This gives a dynamic picture of what is happening. Especially with the more complex diagrams we will encounter later (involving antiparticles and all), this trick is very helpful; you are urged to scan each Feynman diagram you encounter with the FD scanner.

At this point, the analogies used for the description of a force have perhaps been stretched to the extent that you start to make objections, probably along the lines of those listed below (each of these will be discussed in detail later, but it is proper to at least mention them here).

First, why do we never see the beanbag? The reason is precisely the quantization I have invoked to describe the force. *Either you've got a quantum, or you don't; a fraction of a quantum is never observed.* If you want to see the quantum that the particles exchange, you must absorb it in its entirety; and if you do, the field quantum will not arrive at the second vertex. You may arrange to pass the beanbag along after inspection, but that would also spoil the connection between the vertices: the process you have created by intervening with your observation is not the same as the process that gave rise to the force.

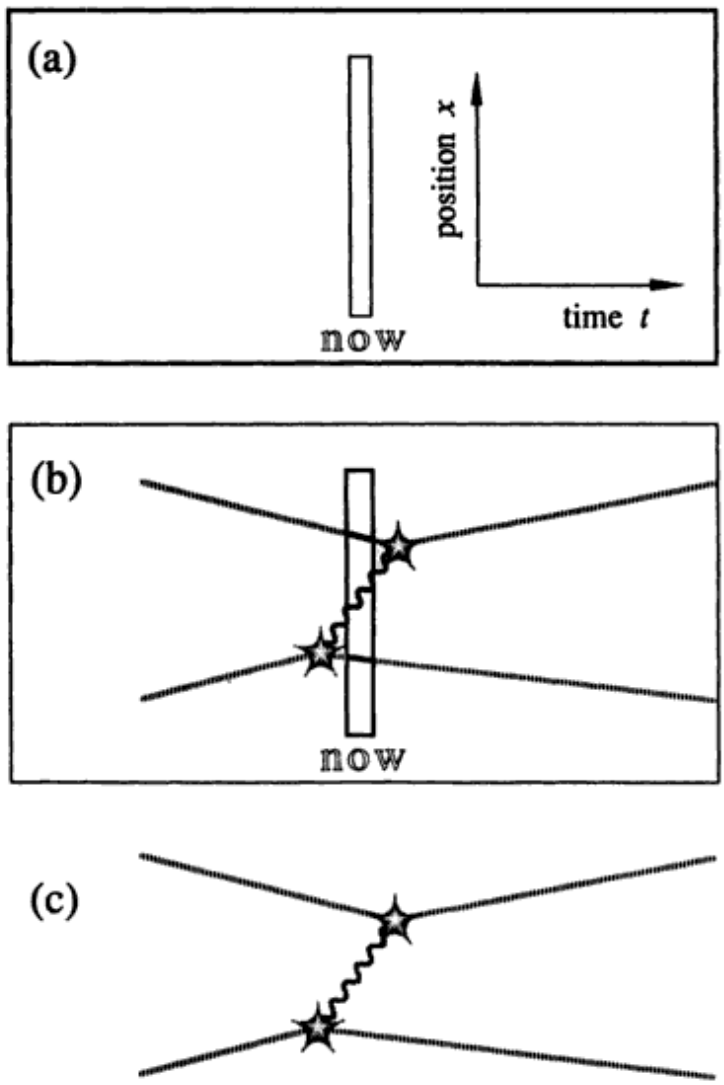

Fig. 1.10 (a) A Feynman diagram scanner, made by cutting a narrow slot in a piece of cardboard. The slot indicates the situation at the present time. By looking at the Feynman diagrams in this book through this scanner, you may produce snapshots as in the previous figures, or motion pictures by sliding the scanner in the direction of increasing time (from left to right on the page). Diagram (b) shows the scanner in action. The Feynman diagram in (c) can be used for practice.

It isn't that the intermediate quantum is unobservable; it can be observed just fine, but your observation does not give you a fair view of the quantum exchange. With quanta, it is impossible to take a careful look: either you look or you don't. Here the beanbag analogy goes grossly wrong because Nature behaves in an essentially different way on an atomic small scale than on a human large scale (we will presently see that small is in fact defined as that scale of size below which quantum behaviour becomes overt). In

summary: we see the consequences of the exchange of the field quantum, but – unless we want to mess up the force – we cannot see the quantum itself. Accordingly, the latter is called a `virtual quantum`.

Second, you have probably noticed that the beanbag-throwing mechanism works for skaters, but cannot be expected to work exactly as stated for subatomic particles. An electron, say, doesn't have an internal store of energy that can provide the work done in the throw. Thus, energy cannot be conserved when a virtual quantum is emitted. That appears to be a shocking statement, and yet it does not matter: as we have seen, the force comes about through *exchange* of a virtual quantum, so we are only interested in the overall conservation of energy after the quantum has been caught by the target particle. We couldn't possibly observe whether or not energy is conserved at a single vertex, because the virtual quantum must be absorbed entirely in any process designed to measure its energy; and this, as before, would devastate the effect we were trying to observe.

Third, one might ask: what's in the bag? As it happens, this is one of the main themes in this book. The beans in the bag represent information of some sort that is exchanged between the vertices, and by putting lots of databeans in the bag we may construct forces with very complex and subtle behaviour. We can expect that the laws of relativity and quantization place restrictions on the data which the bag can carry. Moreover, we may hope to discover some general principle, over and above these laws, that prescribes what quantum beanbags may or may not contain. We will find that such a principle exists: `symmetry`. Current speculations in theoretical physics suggest that the contents of the beanbags are in fact stored in higher dimensions outside the common four of space-time (Chapter 14).

1.5 Matter and force

In the above, we saw that the laws of relativity and quantization lead us to consider exchanged field quanta as the carriers of a force. In the analogy given, the beanbag represents the force and the skaters represent the matter on which the force acts. With the powerful bias that gross everyday physics produces in our minds, it seems natural to think that matter and force are totally different things. And yet, having carefully considered how relativity and quantization led to the concept of the exchange of quanta, you may wonder why we shouldn't occasionally expect to see two beanbags throwing a skater at each other!

What is it, then, that produces the radically different behaviour of matter and force in our large-scale world? This is the thrust of the next seven

chapters, but I think that it is important to discuss the distinction between matter and force briefly here. As we will see, the laws of relativity and quantization, together with certain properties of space-time, imply that all quanta can be divided in two classes: the Fermi–Dirac particles or `fermions`, and the Bose–Einstein particles or `bosons`. To which class a particle belongs depends on the amount of rotation it carries. This amount is indicated by a quantity called spin angular momentum, or `spin` for short. Spin is quantized. In suitably chosen units, the only values that the spin s of a quantum can assume can be written as $s = n/2$, where n is a whole number: $0, 1, 2, 3, \cdots$. If n is an *odd* number, the particle is a *fermion*; if n is *zero or even*, it is a *boson*. Thus, we have s(fermion)$= \frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \cdots$, and s(boson)$= 0, 1, 2, 3, \cdots$. A list of the more common bosons and fermions is given in the table below.

name	*symbol*	*spin*	*el.charge*
photon	γ	1	0
weak photon	W^+, Z, W^-	1	1, 0, -1
gluon	g	1	0
electron	e	$\frac{1}{2}$	-1
neutrino	ν	$\frac{1}{2}$	0
proton	p	$\frac{1}{2}$	1
neutron	n	$\frac{1}{2}$	0

In most Feynman diagrams (Fig. 1.11) we see fermions exchange bosons, such as in the scattering of one electron off another. However, it is perfectly possible to have bosons exchange fermions, for example in photon-photon scattering.

The existence of spin is intimately associated with the fact that, in three-dimensional space, objects can be rotated about an axis that lies inside that space. For example, the rotational axis of Earth points to the star Polaris, and not towards some point outside our Universe. If Earth were a flat circle, it could still rotate, but the rotation axis would *not* lie in the two-dimensional universe of such a flat Earth: it would be perpendicular to it, and be 'out of this world'.

Because spin is the amount of rotation that a quantum carries, it is plausible to expect that the way in which a particle behaves when it is

different, but are the names that we have given (guided by mere large-scale behaviour) to collections of fermions and bosons.

In summary, the two classes of particles show the following behaviour: fermions act like infants. You always have to keep them out of each other's range. The one in the sandbox won't tolerate number two; you must put it in a stroller. That won't hold more, so the third infant must ride a bicycle. Number four takes a motorbike, number five a car, and so forth, all the way up to the speed of light. Thus fermions, like infants, take up an amount of space that is gigantic compared with their size. Bosons are quite the opposite. They behave rather like rugby players. As soon as one hits the ground, both fifteens pile themselves right on top, forming a wriggling heap. These are chummy particles which squeeze themselves in the smallest possible volume, never taking up much space but capable of acting as a coherent team.

Fermions collect in definite lumps, but cannot produce a coherent field; a batch of bosons collapses without much resistance, but can act coherently when exchanged. Therefore, fermions appear to us as matter, whereas bosons provide what we call force. Thus, I was *very* amused to see the clash of light sabres in the *Star Wars* film trilogy: since a light sabre is presumably made of particles of light (which are bosons) it will hardly stand up to impact! The law of spin and statistics gives an excellent reason to make swords out of fermions (such as atoms of iron and carbon). Our understanding of Nature involves two closely related quests: the search for the free fermions that occur (often loosely called fundamental particles), and the search for the bosons that they can exchange (occasionally called field particles).

2

Stalking the wild rainbow

2.1 Colours and spectra

There is a striking similarity between the struggle of today's physicists with particles and the pursuit by their predecessors of a hot topic: spectroscopy. A lot of the jargon of the present relativistic quantum field theories comes directly from spectroscopic descriptions. Our search for order in the bewildering array of particle masses is like the efforts of scientists who, towards the end of the preceding century and in the first decades of the the twentieth, tried to find some order in the arrays of light waves that can be emitted by atoms.

It has been known since time immemorial that there are colours, but it wasn't until the seventeenth century that it became clear that all colours are different manifestations of the same phenomenon: light. The behaviour of light began to yield to quantitative descriptions through the brilliant work of Snell and Huygens. The colours and behaviour of the rainbow, first correctly explained by Descartes, were then no longer a religious mystery. Newton worked systematically on the splitting of sunlight by glass prisms (Fig. 2.1) into its coloured components, and he showed that the colours can be recombined to yield white light.

Huygens's wave theory of light reached its apotheosis in the work of Fresnel, who proved that the colour of a light wave is determined by but one quantity, the wavelength. Fresnel's work also paved the way for a new gadget with which white light can be split into its constituent colours: the `diffraction grating`. Because few things are so convincing as experiments done by oneself, we will take some time out to construct a device that splits light into colours: a `spectroscope`. Seriously: with some cardboard and a small diffraction grating (Fig. 2.2), you will be doing

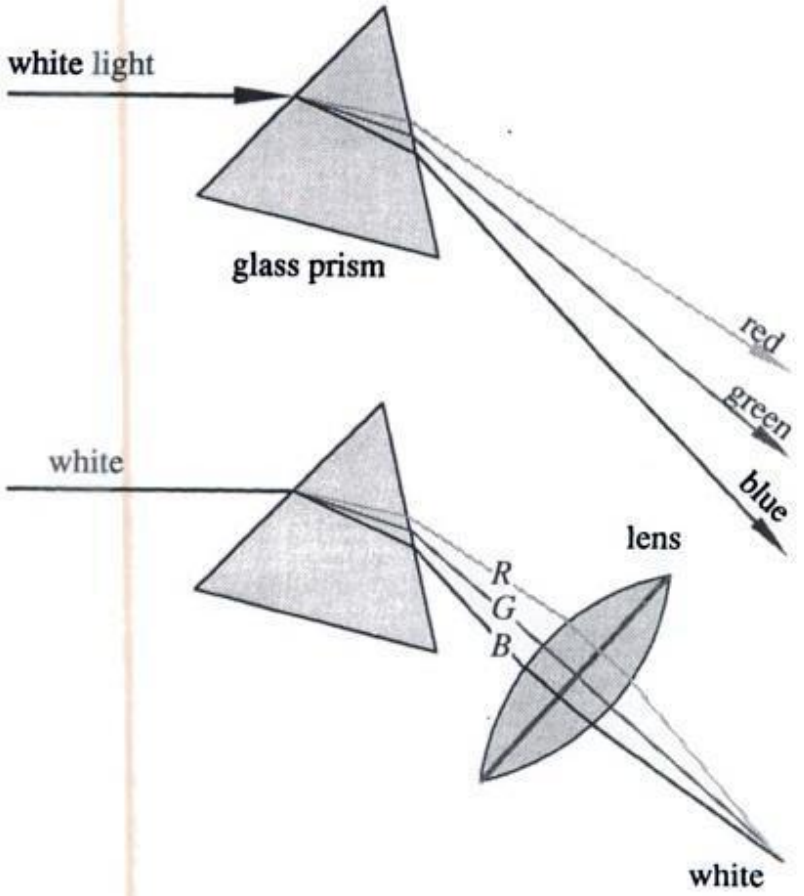

Fig. 2.1 Two basic experiments with a prism. When white light passes through a glass prism, the light is split into a band of colours. Each colour corresponds to a certain wavelength. When the colours are combined by means of a lens, the eye sees the result as white light again.

with light waves the exact same thing that giant mass analysers do in the world's biggest particle accelerators.

A diffraction grating is a transparent or reflecting piece of material, on the surface of which are cut a large number of thin, parallel grooves. Such gratings can be bought at many science museums and suppliers of scientific equipment. A grating is a multimillion-aperture version of the double-aperture diffraction experiment which we will discuss in detail in Chapter 5. Light passing through the grating is deflected by a process called diffraction, and the amount of the deflection is proportional to the wavelength of the radiation. Thus, a diffraction grating sorts light waves by wavelength.

Now take a piece of grating, hold it close to your eye, and look through it at a small incandescent light bulb in the distance. You will see the bulb, with

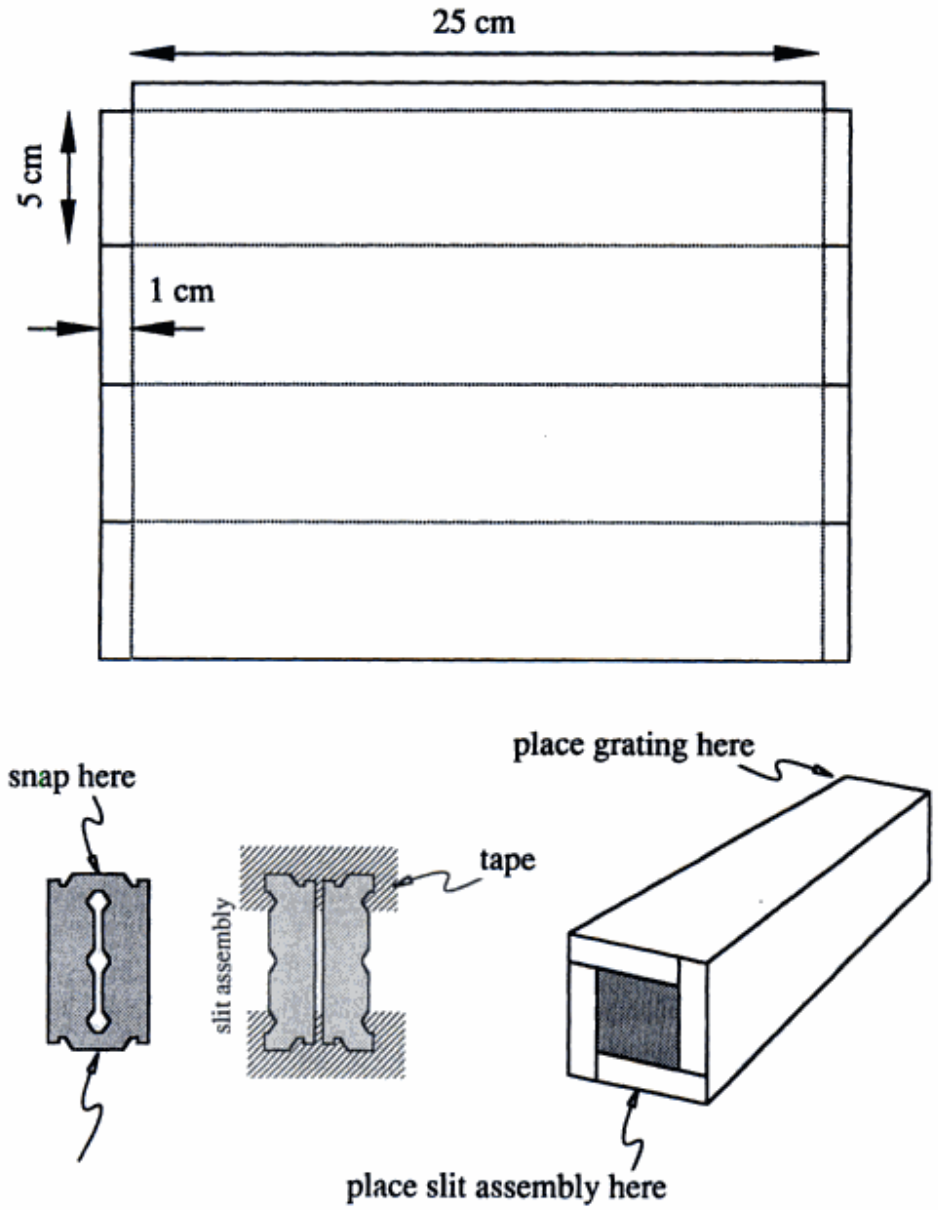

Fig. 2.2 Plan of a spectroscope box. Cut the pattern shown here out of cardboard (preferably black). Assemble it to make a long box. Make a narrow slit from the two halves of a razor blade (*be careful!*) or from two straight strips of folded aluminium foil. Place the slit over one end. On the other end, place a plastic diffraction grating, as sold by various science supply shops, or in some science museums. Make sure that the grating is aligned properly with the slit.

on each side an image of it, smeared out into the colours of the rainbow. This band is called the `spectrum` of the light. We can improve on this experiment by constructing a box for the grating which blocks stray light

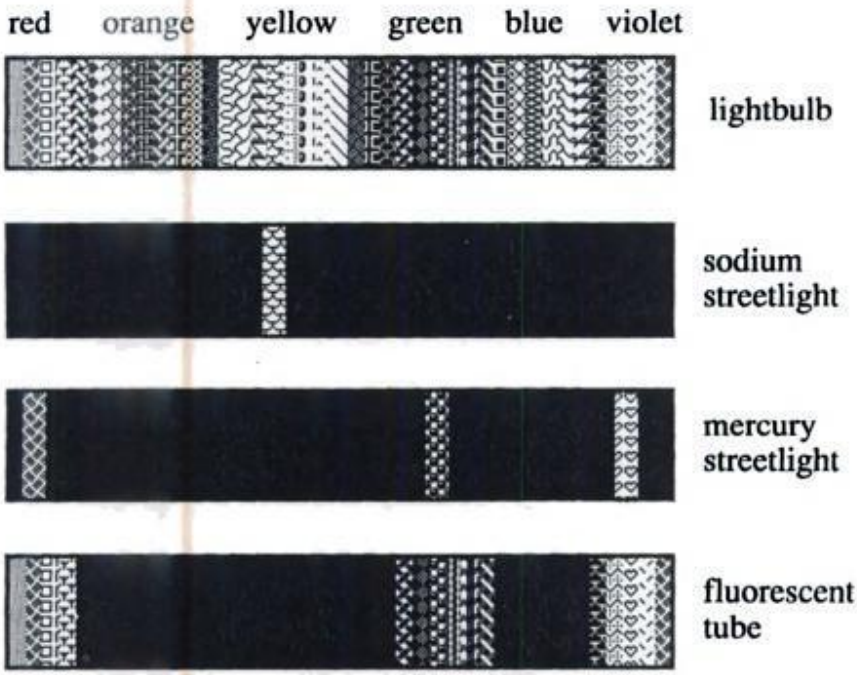

Fig. 2.3 Sketches of continuous and line spectra. Red is on the left, blue and violet on the right. The funny textures represent the various colours, which are irreproducible in a black and white book. All the more reason to do the experiment!

from the surroundings. Moreover, the box can be fitted with an entrance slit to sharpen the image of the spectrum. The slit and the box effectively act like a pinhole camera, but here the pinhole is small only in one direction and elongated in the other. Fix the slit on to the box, and rotate the grating until the spectrum is broadest. The grooves of the grating are then parallel to the slit. Keep the grating close to your eye, and look at a light source through the slit at the other end of the box.

2.2 Spectral lines

Now look at as many different light sources as you can find. In particular, look at the neon signs in shop windows and at sodium or mercury city lights (see Fig. 2.3 for a rough sketch of what to expect). You will notice that the light from many of these sources is not a continuous band but contains a number of individual streaks of pure colour. These are called `spectral lines`. Investigate the lines from different sources. Are there regularities, similarities, differences?

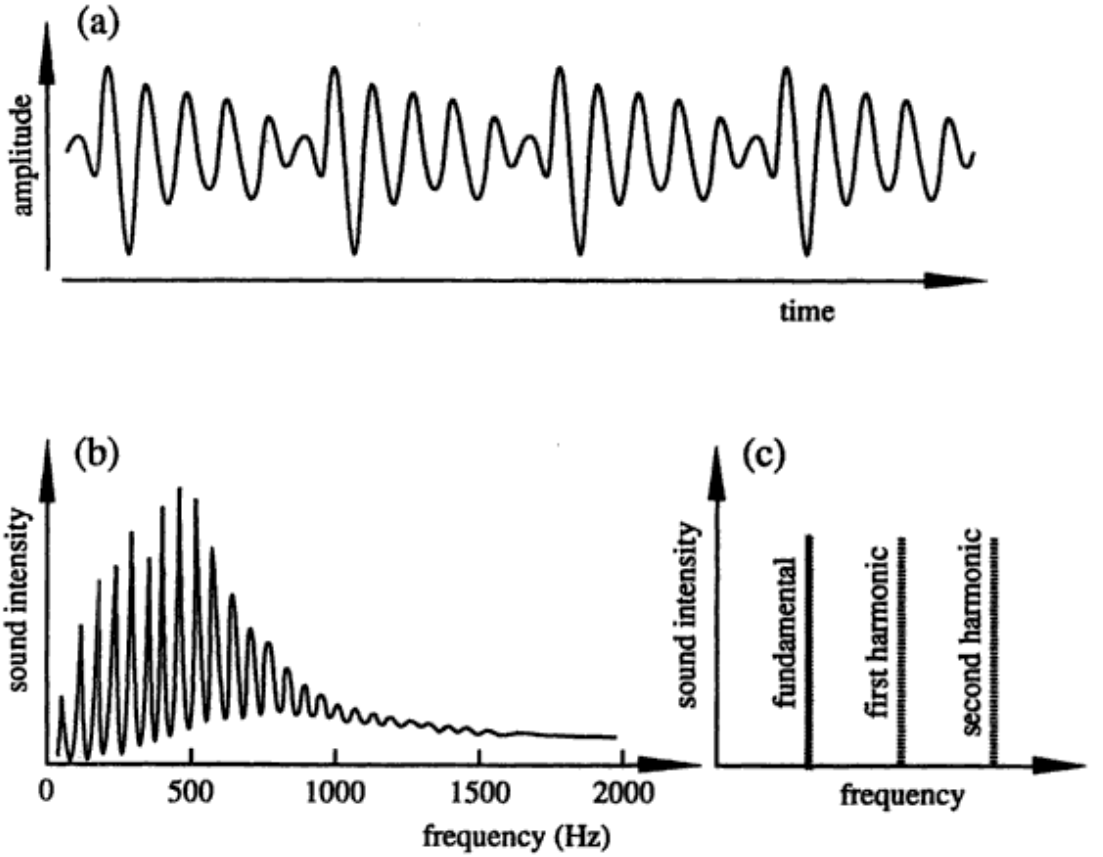

Fig. 2.6 (a) The wave form of a French horn is not a pure oscillation but contains many higher harmonics, also called overtones. These are shown in the spectrum of the sound (b); the curve gives the strength of the sound at any particular frequency. The unit of frequency is the hertz (Hz), the number of oscillations per second. (c) An indication of the place of a series of overtones in a schematic sound spectrum.

It was discovered that the differences in frequency correspond to differences in energy of the atom. This energy (Fig. 2.7) is proportional to the frequency of the emitted light. Thus, the various spectral lines could be attributed to transitions inside atoms, due to some (then mysterious) internal degrees of freedom of the atom which somehow fail to obey the rules of classical oscillators such as guitar strings.

One could try to resort to extremes in a desperate effort to rescue classical mechanics, as follows. Suppose that the atomic laws of motion are such that only exactly pure oscillations occur, without any harmonics whatever. In this view, each line corresponds to a 'fundamental' or 'elementary' light wave, unrelated to the others. This is bizarre, but not strictly forbidden, and even though it does not explain Ritz's law we may hope to get away with postulating one perfect oscillator for each spectral line. But there is a problem with this. If we shake an oscillator around, it begins to vibrate;

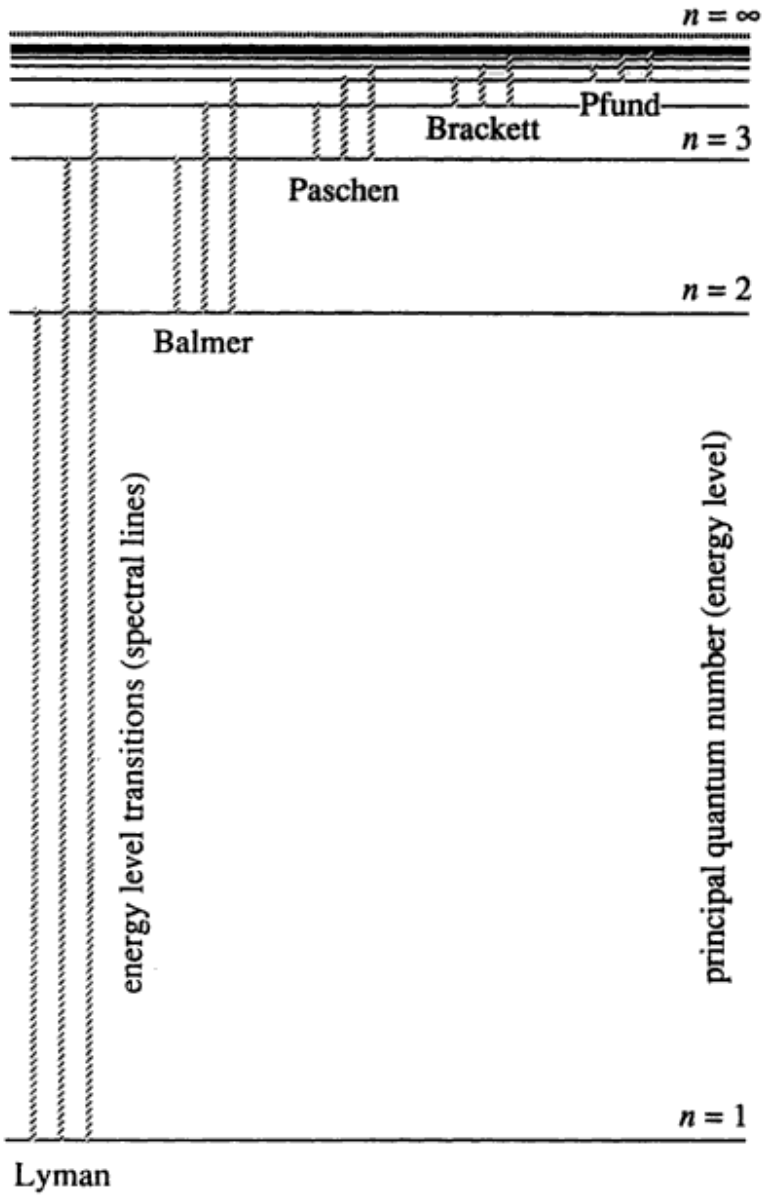

Fig. 2.7 The energy levels of the hydrogen atom. The higher the principal quantum number n, the higher the energy. Some transitions between energy levels are indicated. These are responsible for the spectral series shown in Fig. 2.5.

some of the energy used in the shaking is absorbed by the vibrational motion. Thus, if each spectral line is due to an independent oscillator, each one soaks up a little energy. But the number of lines observed in a given species of atom is enormous (in fact, it turns out to be infinite). If each line oscillator is an energy sink, we would expect to be able to add very

large amounts of energy to matter without it getting appreciably hotter. This is not observed. Somehow, the oscillators corresponding to the spectral lines do not soak up energy. This is in blatant contradiction with classical mechanics.

3

Light

3.1 Waves of light

Light behaves like a wave. In the laboratory, we can measure the vibrations that are set up in matter when a light wave comes by. We can also observe the peculiar light-and-dark patterns that occur when two light rays are made to act simultaneously, an unmistakable sign of wave behaviour.

Let us make a small excursion into wave motion. Throughout this exposition, you are encouraged to experiment as much as possible with waves (preferably real ones, or at least those in the diagrams), to become familiar with their fascinating behaviour. Water waves in the bathtub, or those seen on open water from a high tower or an airplane, are especially instructive.

Waves are periodic; if you pick a point in space (e.g. the surface of a pond at the point where a reed pokes through) and you watch the motion of the wave carrier (the water that bobs up and down), then you will see that the state of motion of the carrier repeats itself at regular intervals of time. This interval is the `period` of the wave. A free wave has a `velocity`, i.e. a direction and a speed. At a given point, a wave alternates periodically between a certain maximum and a minimum. Half the height between wave crest and wave trough is called the `amplitude` of a wave (Fig. 3.1). The distance from crest to crest is the `wavelength`, and the number of wave crests going by a fixed point in one second is the `frequency` of the wave.

The extent to which a wave has completed any oscillation cycle (Fig. 3.2) is called the `phase`. For example, we may (arbitrarily) start counting at a wave crest, and call that phase zero; when the phase is half a cycle, there will be a wave trough. The phase is often expressed in degrees, so that 360° corresponds to one full oscillation cycle. The reason is that uniform circular motion can be used to generate a wave (Fig. 3.3). Imagine a series of circular wheels, each with a dot painted on its circumference. If these

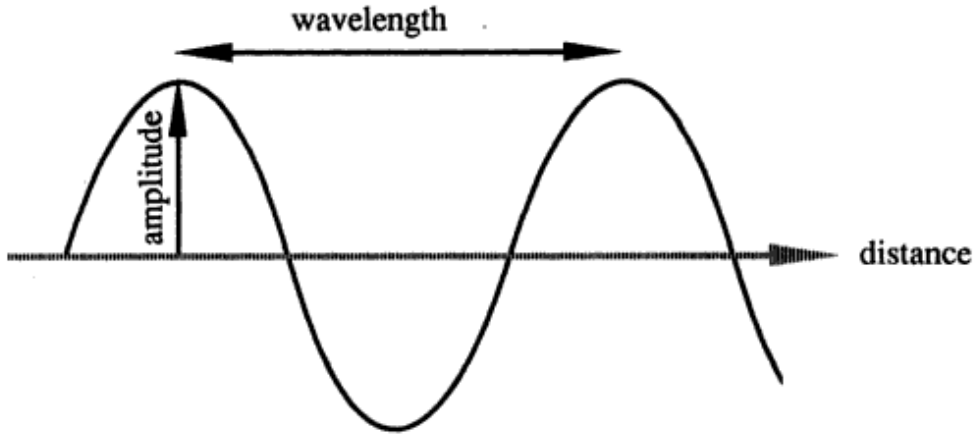

Fig. 3.1 Two of the main properties of a wave: the wavelength and the amplitude.

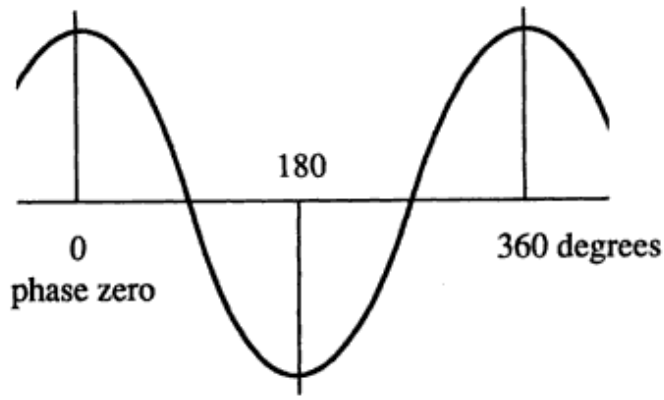

Fig. 3.2 The phase of a wave. Starting at a fixed but otherwise arbitrary point (here the point of maximum height), the phase measures the extent to which a wave has completed an oscillation cycle.

rotate in unison, we perceive the motion of the dots as a progressing wave, provided that the dot on each wheel has a constant offset (called `phase shift`) with respect to those on adjacent wheels. The undulating curves generated by dots on revolving circles are called *sine waves*. A *cosine wave* is the same as a sine wave, but with a phase shift of $-90°$. In a later chapter, we will consider the possible paths that particles can take, and we will see how the wave properties produced by a rolling wheel allow us to describe the world on an atomic scale.

A wave is not a material thing. We clearly see ripples propagate along the water surface, but closer inspection shows that a cork is not moving

We will assume that the speed of the wave is the same everywhere; the consequences of differences in speed can also be studied by means of Huygens's principle,† and the effects of refraction, for example – as expressed in Snell's Law – can thus be calculated; but we don't need any of that here. Suppose that the wave has reached a certain point in space. In the next wave period, each point on the wave crest emits a ripple with a radius equal to the wavelength. All those ripples together generate the next leading crest, and in this way the wave front advances. You can easily see from this construction that the direction of the velocity of a wave is always perpendicular to the wave front, so that a straight wave front remains straight. This type is called a `plane wave`. Also, a spherical wave remains spherical (Fig. 3.6). An irregular wave smooths itself out.

If a wave – say it is a plane wave – encounters an obstacle, Huygens's principle shows that the wave front curls around it. This is called `diffraction`; it explains why we can hear sounds from the other side of a building, for the diffraction allows the waves to sneak around a corner, even in the absence of reflection. You are urged to try out Huygens's principle in a variety of situations; suggestions are given in Fig. 3.7 and Fig. 3.8. An especially important case is a wave impinging on a wall with a small hole in it (Fig. 3.5). Please convince yourself by applying the Huygens construction that the wave beyond the wall is spherical if the hole is small compared with the wavelength, and that it becomes less and less spherical if the hole is made bigger. This is *very* important, as will be seen later.

3.3 Interference

In what follows, we will mostly be interested in the ways in which waves can be added together, because the resulting effects are the hallmark of wave motion. Moreover, we will see that the way in which Nature adds quantum processes corresponds exactly with the addition of waves. Consider a long train of wave crests and troughs, going from left to right. Now take a wave that is precisely the same, except that it travels at a slight upward slant. What do these waves look like if we combine them? That depends on what is meant by 'combine'.

† Huygens's principle is not a fundamental 'law of nature', or anything of that sort. Rather, it is a geometrically and intuitively easy rule (which can be rigorously derived from the equation of wave propagation) that summarizes the most important aspects of the motion of waves. Use of this principle allows us to describe many of the concepts of wave mechanics without having to go through heavy mathematics. It is a remarkable reflection on the genius of Huygens that he found this powerful principle long before the wave equation (a partial differential equation that strictly governs wave motion) was discovered!

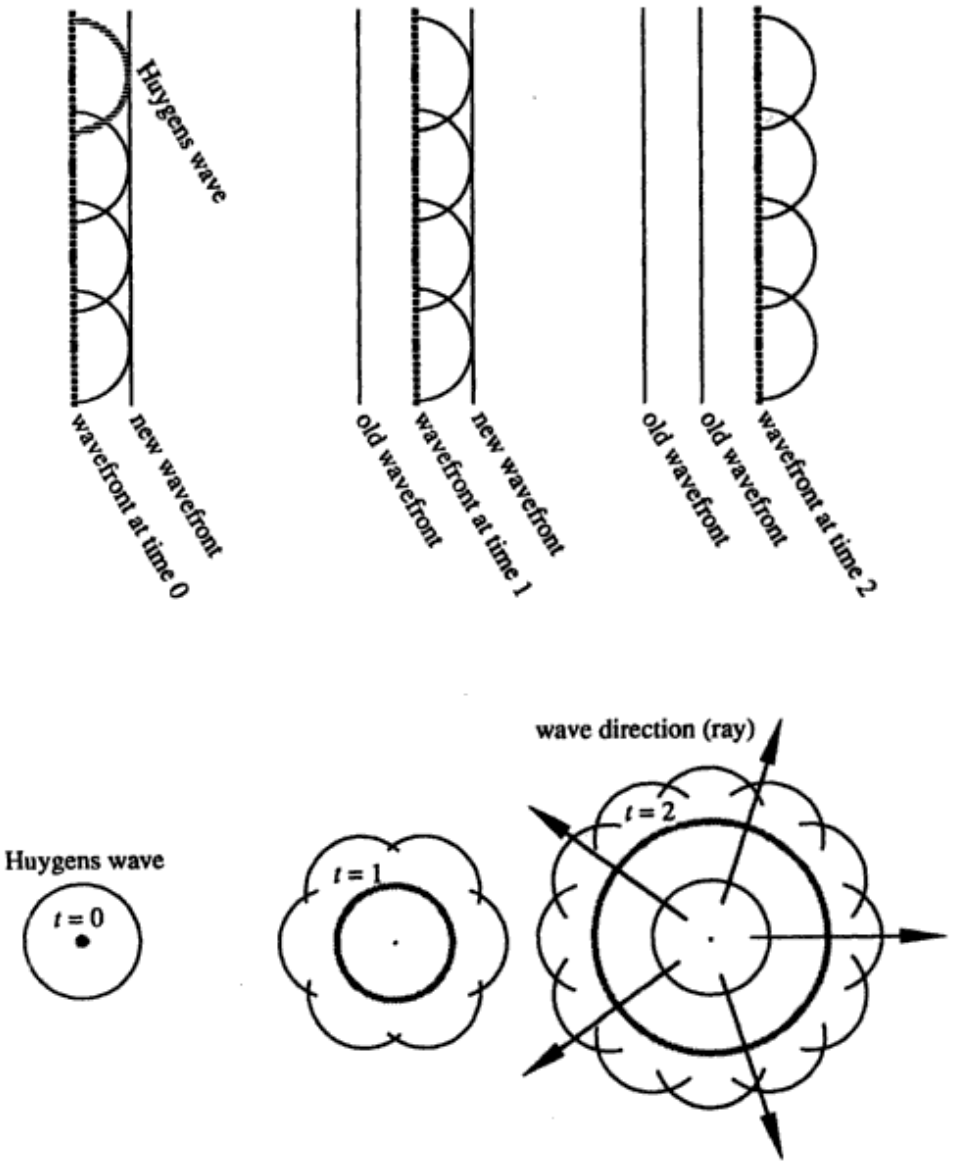

Fig. 3.6 Propagation of a plane and a spherical wave according to the Huygens construction. Each new wave front is found from the previous one by letting small wavelets propagate from each point on the old front. The direction of the wave is everywhere perpendicular to the wave front. Lines which trace these directions are called rays.

As it happens, classical waves of which the amplitude is very small compared with the wavelength (so-called `linear waves`) can be combined by simply adding the wave heights together (Fig. 3.9). This is called linear superposition, or `superposition` for short. Immediately we see that the superposition of waves generates peculiar patterns, because in some places the wave crests coincide with other crests (leading to a doubling

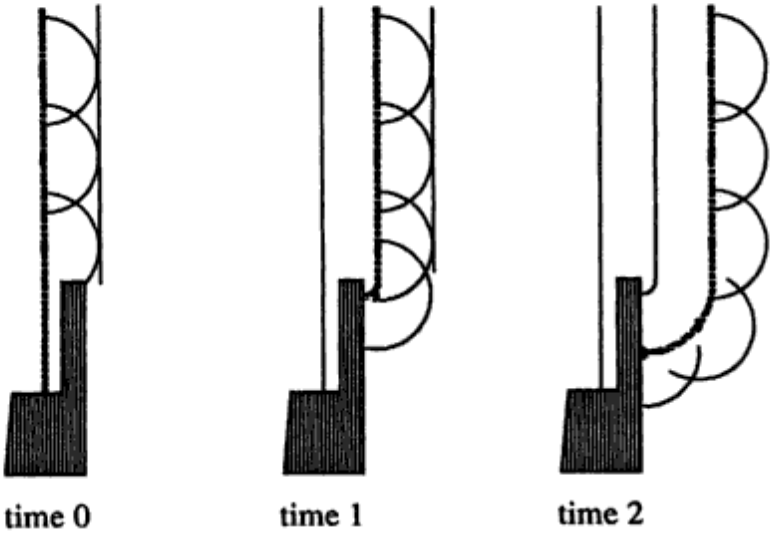

Fig. 3.7 Huygens's construction for the propagation of a wave around a flat obstacle. Notice how the construction predicts the bending of the wave around the corner; this is called diffraction. Curiously, Huygens never appears to have used this remarkable insight explicitly.

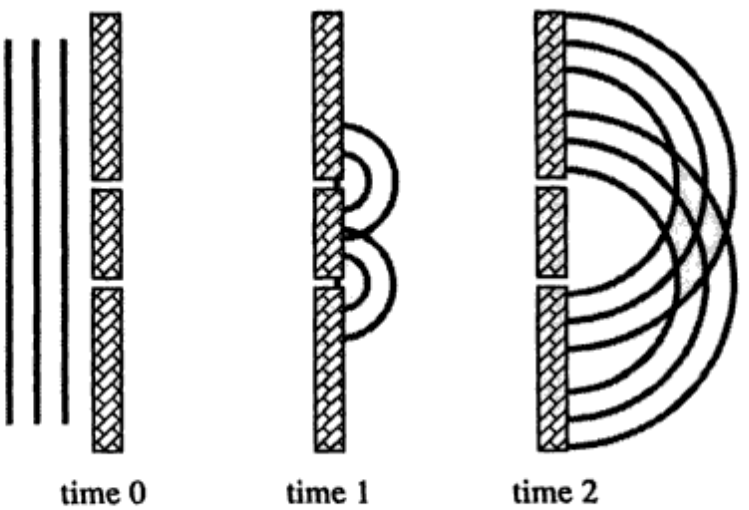

Fig. 3.8 Huygens's construction for the propagation of a wave through a screen with two holes. Notice the appearance of a zone where the two waves on the far side of the screen intersect; this is called interference.

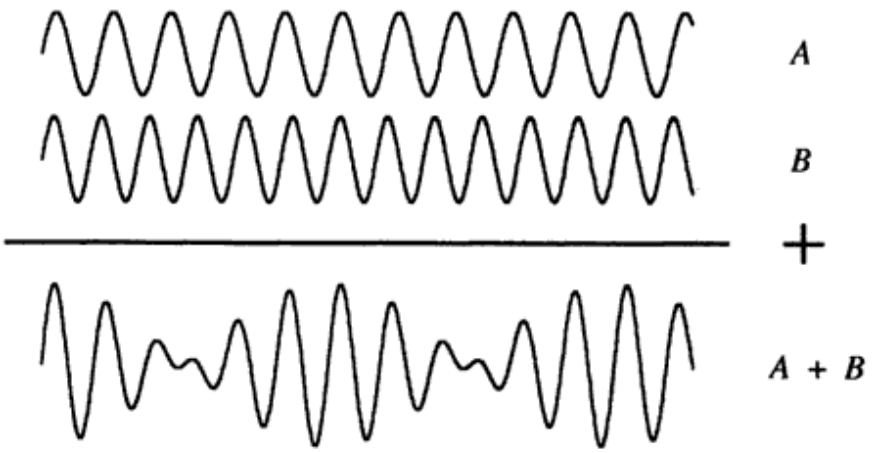

Fig. 3.9 The result of adding two one-dimensional waves. In places where two wave crests coincide, the resulting amplitude is large (constructive interference). Where a crest of one wave lies on top of a trough in the other, the result is zero (destructive interference). The total wave pattern shows a regular alternation of high and low points of the wave envelope, called beats.

of the amplitude), whereas in some other places the crests coincide with the troughs (leading to a cancellation of the motion and hence to a zero amplitude). This pattern generation is called `interference`. When the amplitudes add, we have constructive interference; when they cancel it is destructive.

An important example of interference, which we will use a great deal below, is the one in which a wave falls on to a barrier with two holes in it (Fig. 3.10). We will assume that the holes are small compared with the wavelength, and Huygens's principle shows that the wave beyond the barrier consists of two spherical waves. By superposition of these waves, we see clear interference bands, called `fringes`. If we were to place a row of observers beyond the barrier, they would report regions of double amplitude alternating with zero amplitude.

You should make a transparent photocopy of the spherical wave in Fig. 3.11, and experiment by superposing it on to its original.† You will notice immediately that the spacing of the interference fringes in the two-hole diffraction changes dramatically as the distance between the holes changes: the smaller the distance between the holes, the larger the spacing between

† Technically, this superposition is not quite the same as the one discussed above, because in the troughs the wave does not have a negative value. It is impossible to print with negative ink, so the superposition occurs by adding according to what computer buffs call *logical* AND *mode*: if 1 represents black and 0 white, then $0+1=1+0=1+1=1$ and $0+0=0$. Even so, the basic interference phenomenon persists, in the form of *Moiré fringes.*

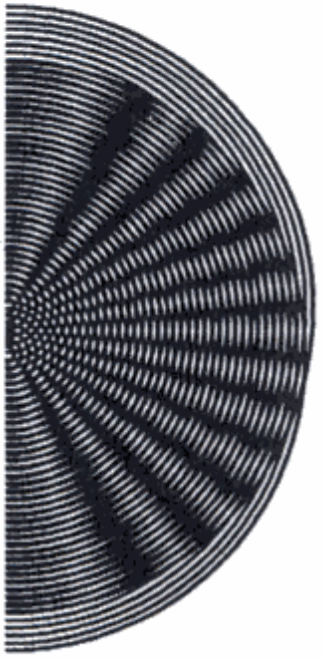

Fig. 3.10 Interference between two spherical waves that emanate from small holes; this is a more detailed version of 3.8.

the fringes, and vice versa (Fig. 3.12). Hence we conclude that interference between two waves depends on the relative position in space of the wave sources.

Further experimentation shows that interference also depends on the relative position in time of the waves. If we arrange things in such a way that the holes in the screen show exactly the same motion (when a wave crest passes one hole, the other one lets through a wave crest also, at exactly the same time), then we say that the waves are *in phase*; the phase indicates the relative position in time of the waves, because it shows which fraction of a cycle the wave has completed. It is often useful to express the difference in phase between two waves in degrees, from 0° to 360°, after which we start again at zero. If the phase difference is 0°, the waves are in phase and they oscillate exactly in step. Thus, crests coincide with crests and troughs with troughs: when the waves are in phase, the interference is fully constructive. If the phase difference is 180°, a crest coincides with a trough and the interference is maximally destructive.

Interference can produce zero amplitude, for example when two waves with the same amplitude arrive 180° out of phase with each other. Furthermore, it is very important to realize here that only the phase *difference* matters in

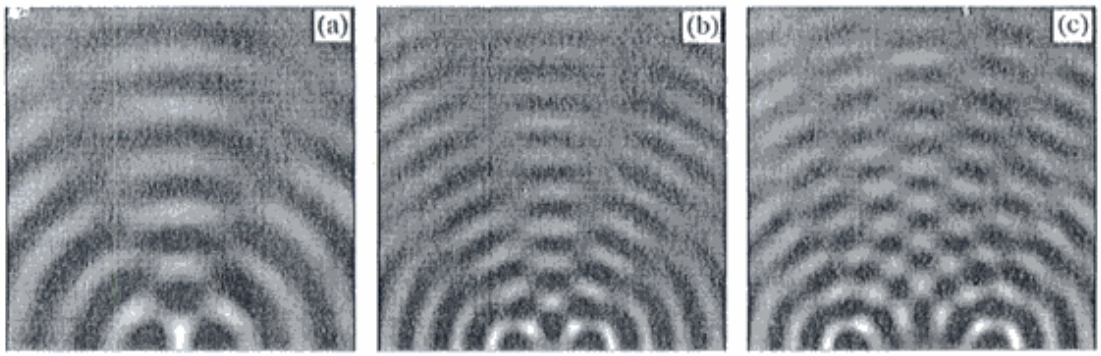

Fig. 3.13 Examples of interference for the double-hole experiment. These greyscale maps show the actual result of letting two spherical waves interfere. Notice the difference between (c), where the holes are far apart and the interference bands are close together, and (a), where the situation is the reverse. Case (b) is in between these extremes.

naturally in standing wave patterns. We already saw something very similar in the discussion of the spin s of particles, which was given by $s = n/2$. We will encounter similar cases throughout this book (e.g. in Chapter 9) as `quantum numbers`.

When a light beam from a point source is thrown on to a screen with two very small holes, the intensity of the light received on a film beyond the screen follows exactly the sequence of interference maxima and minima which we expect on the basis of the above diagrams. It is evident from this and other such experiments that we can describe light as a wave with a definite velocity, wavelength and frequency. Moreover, light waves can be combined by means of linear superposition. Radio waves have low frequency, light is intermediate, X-rays and gamma rays have high frequencies. Our eyes see the frequency of light as *colour*: red at the low frequency end, via orange, yellow, green and blue, to violet at the high frequency end of the spectrum.

3.5 Particles of light

Light behaves like a particle. In the laboratory, we can measure the photoelectric effect that occurs when light particles – called `photons` – give up their energy to catapult electrons to freedom out of the metal in which they were held captive. We can also observe the scattering that occurs when photons collide with free electrons and recoil off them like billiard balls, an unmistakable sign of particle behaviour.

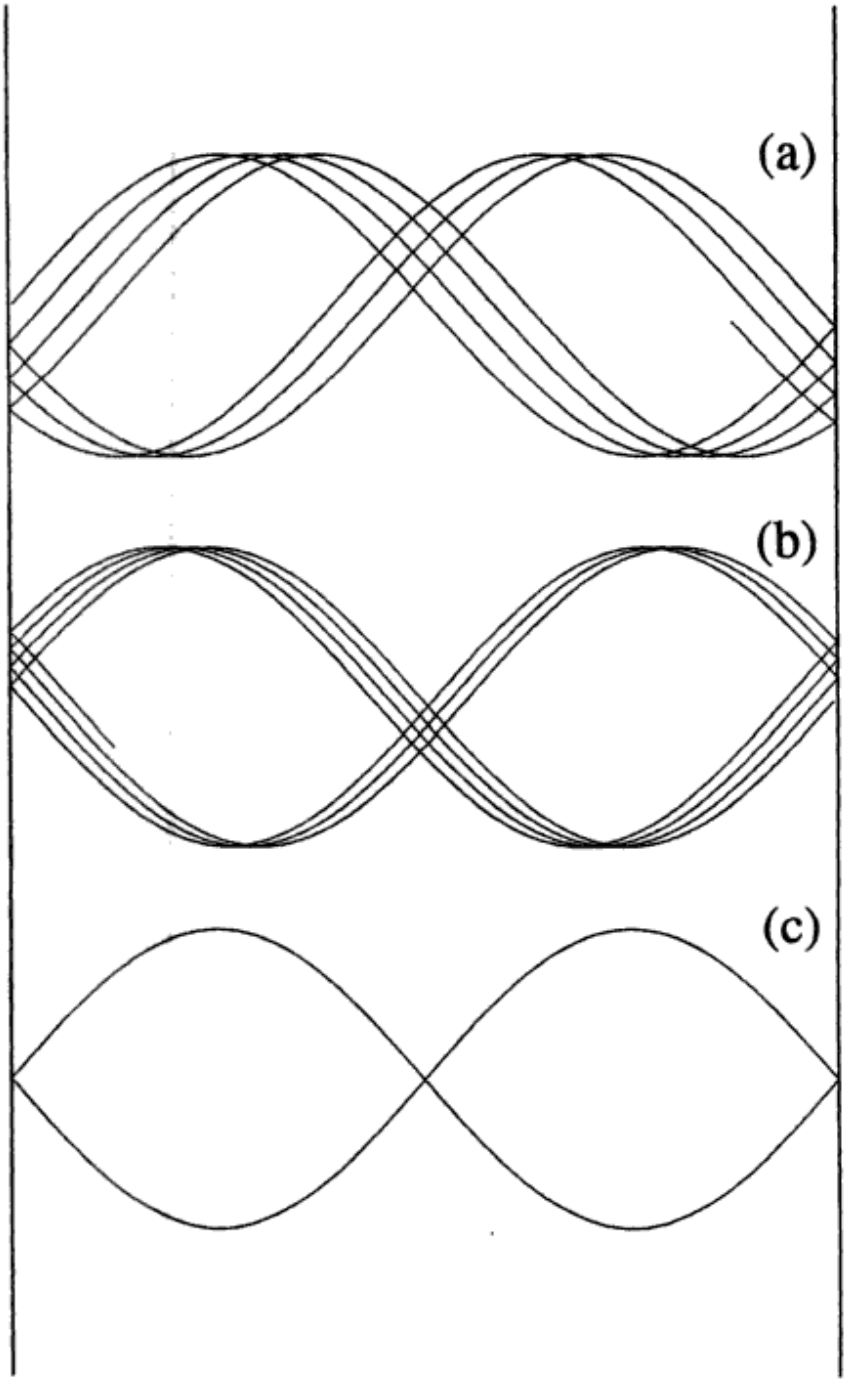

Fig. 3.14 Construction of a standing wave between two walls. In (a), the wavelength is a little too short, and the reflected wave does not return exactly to its point of departure. In (b), the wavelength is a little longer but still not long enough. Only in (c) does the wavelength match the distance between the walls. In that case, constructive interference occurs, resulting in standing wave.

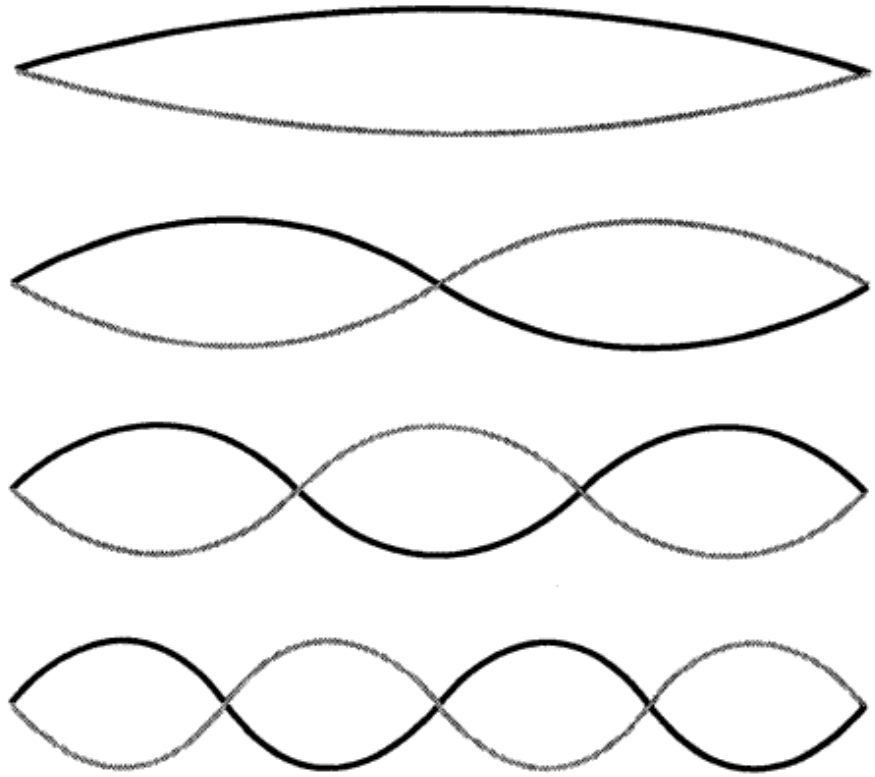

Fig. 3.15 Examples of modes on a string. The fundamental is shown at the top; below it are the successive overtones. These correspond to the harmonics shown in the sound spectrum of Fig. 2.6.

We need not make a deep excursion into particle motion. Most people are, from everyday experience, familiar with particle behaviour. But a brief reminder of some of the concepts of particle mechanics (often called 'classical mechanics') will assist understanding of many of the concepts in the book.

The basic question of mechanics is: *where is what when?* To find answers to this question, we must first find precise quantities to describe the 'where', 'what' and 'when'. In the following we will not yet talk about the effects of relativity, so we must make a mental note that changes will have to be made later.

The 'what' is a tough one to formulate; we will use the word `particle`, roughly meaning something that is an idealization of everyday objects. A particle in mechanics, sometimes called a point mass, is everything that an ordinary object like a tennis ball is, except that it does not have spatial extent. Hence, properties that require size in order to have physical consequences – shape, rotation and so forth – are suppressed.

With 'where' things are a bit easier. The quantity describing it is the position: an arrow pointing to the particle, starting from a suitable reference point. If we try to find a mathematical object corresponding to this distance-with-direction, it turns out that an ordinary number is not good enough. Position is a more complex beast, which doesn't obey the arithmetic laws of single numbers. As it happens, a good mathematical object with which to describe position is a row of three numbers; accordingly, we say that space has three dimensions.

There is a lot of leeway in the manner used to express these three numbers, called coordinates. For example, they might be 'forward, sideways, up', or 'latitude, longitude, height', or 'azimuth, altitude, range'. Various forms of coordinate systems have been invented to make calculations easier. Thus, we have rectangular coordinates (for Bauhaus architects), spherical coordinates (for sailors), elliptical coordinates (for navigators using Loran beacons) and so forth. But we always need three numbers to specify a position in space.

The row of numbers is called a vector (a vector of dimension 1, which is a single number, is called a scalar). We can visualize the position vector as an arrow pointing in the direction of the particle; the length of the vector corresponds to the distance between the particle and the reference point. Two vectors can be added to make a third by pairwise adding their coordinates by ordinary arithmetic addition. Thus, if $\vec{A}$ and $\vec{P}$ are vectors, consisting of a row of numbers $(a, b, c, \cdots)$ and $(p, q, r, \cdots)$, then $\vec{A}+\vec{P}$ is also a vector, corresponding to the row $(a+p, b+q, c+r, \cdots)$.

About the 'when', people have speculated and philosophized for aeons. The necessity for a 'when' is not obvious, and in an imagined universe one may perhaps be able to do without it. In the actual Universe we find that measurements of the position vector are not unique without the specification of another quantity, called time. It turns out that time is a scalar: a single number, obeying the customary $2+2=4$, suffices to describe it.

3.6 The equation of motion

In order to find a quantitative answer to 'where is what when', we must find a prescription that tells us uniquely what position corresponds to what time. Let us take the difference between the position at an arbitrary time and the position a very small amount of time later (in Chapter 4 we will see what is meant by 'small', both in space and in time). We take this difference and divide it by the time elapsed between positions. Thus we obtain the rate of change of the position, called the velocity. By definition, (velocity)

The force of symmetry gives an elementary introduction to the spectacular interplay between the three great themes of contemporary physics: quantum behaviour, relativity and symmetry. In clear, non-technical language, though without oversimplification, it explores many fascinating aspects of modern physics, discussing the nature and interaction of force and matter.

Through the examination of relevant physical effects, and analogies from daily experience, the book presents in some detail the workings and implications of special relativity, quantum mechanics and symmetries. In so doing, the importance of these fields and their influence on the everyday world is highlighted. Towards the end of the book, its major themes are drawn together to describe the most successful physics theory in history, the 'standard model' of subatomic particles. The strange, counter-intuitive world of the very fast and the very small provides an excellent illustration of many of the topics discussed in earlier chapters.

The lively and non-technical approach of this book will make it suitable for first-year undergraduates in the physical sciences and mathematics, or those just about to embark on such courses, and for anyone with a general interest in these topics. It will also be a valuable accompaniment to more advanced texts on quantum mechanics and particle physics.

The force of symmetry

CAMBRIDGE
UNIVERSITY PRESS

ISBN 0-521-45591-X
90000
9 780521 455916