BY ROGER PENROSE

The Emperor's New Mind:
Concerning Computers, Minds,
and the Laws of Physics

Shadows of the Mind:
A Search for the Missing Science
of Consciousness

Roger Penrose

THE ROAD TO REALITY

A Complete Guide to the Laws of the Universe

12

ALFRED A. KNOPF · NEW YORK 2006

THIS IS A BORZOI BOOK
PUBLISHED BY ALFRED A. KNOPF

Originally published in Great Britain by Jonathan Cape, London, in 2004.

www.aaknopf.com

Knopf, Borzoi Books, and the colophon are registered trademarks of
Random House, Inc.

Library of Congress Cataloging-in-Publication Data

Penrose, Roger.
The road to reality : a complete guide to the laws of the universe / Roger
Penrose. — 1st American ed.
p. cm.
Includes bibliographical refrences and index.
ISBN 0-679-45443-8 (alk. paper)
1. Mathematical physics. 2. Physical laws. I. Title.
QC20.P366 2005
530.1—dc22
2004061543

Manufactured in the United States of America
Published February 24, 2005
Reprinted Four Times
Sixth Printing, February 2006

Contents

1
The roots of science

1.1 The quest for the forces that shape the world

WHAT laws govern our universe? How shall we know them? How may this knowledge help us to comprehend the world and hence guide its actions to our advantage?

Since the dawn of humanity, people have been deeply concerned by questions like these. At first, they had tried to make sense of those influences that do control the world by referring to the kind of understanding that was available from their own lives. They had imagined that whatever or whoever it was that controlled their surroundings would do so as they would themselves strive to control things: originally they had considered their destiny to be under the influence of beings acting very much in accordance with their own various familiar human drives. Such driving forces might be pride, love, ambition, anger, fear, revenge, passion, retribution, loyalty, or artistry. Accordingly, the course of natural events—such as sunshine, rain, storms, famine, illness, or pestilence—was to be understood in terms of the whims of gods or goddesses motivated by such human urges. And the only action perceived as influencing these events would be appeasement of the god-figures.

But gradually patterns of a different kind began to establish their reliability. The precision of the Sun's motion through the sky and its clear relation to the alternation of day with night provided the most obvious example; but also the Sun's positioning in relation to the heavenly orb of stars was seen to be closely associated with the change and relentless regularity of the seasons, and with the attendant clear-cut influence on the weather, and consequently on vegetation and animal behaviour. The motion of the Moon, also, appeared to be tightly controlled, and its phases determined by its geometrical relation to the Sun. At those locations on Earth where open oceans meet land, the tides were noticed to have a regularity closely governed by the position (and phase) of the Moon. Eventually, even the much more complicated apparent motions of the planets began to yield up their secrets, revealing an immense underlying precision and regularity. If the heavens were indeed controlled by the whims of gods, then these gods themselves seemed under the spell of exact mathematical laws.

Likewise, the laws controlling earthly phenomena—such as the daily and yearly changes in temperature, the ebb and flow of the oceans, and the growth of plants—being seen to be influenced by the heavens in this respect at least, shared the mathematical regularity that appeared to guide the gods. But this kind of relationship between heavenly bodies and earthly behaviour would sometimes be exaggerated or misunderstood and would assume an inappropriate importance, leading to the occult and mystical connotations of astrology. It took many centuries before the rigour of scientific understanding enabled the true influences of the heavens to be disentangled from purely suppositional and mystical ones. Yet it had been clear from the earliest times that such influences did indeed exist and that, accordingly, the mathematical laws of the heavens must have relevance also here on Earth.

Seemingly independently of this, there were perceived to be other regularities in the behaviour of earthly objects. One of these was the tendency for all things in one vicinity to move in the same downward direction, according to the influence that we now call *gravity*. Matter was observed to transform, sometimes, from one form into another, such as with the melting of ice or the dissolving of salt, but the total quantity of that matter appeared never to change, which reflects the law that we now refer to as *conservation of mass*. In addition, it was noticed that there are many material bodies with the important property that they retain their shapes, whence the idea of rigid spatial motion arose; and it became possible to understand spatial relationships in terms of a precise, well-defined geometry—the 3-dimensional geometry that we now call *Euclidean*. Moreover, the notion of a 'straight line' in this geometry turned out to be the same as that provided by rays of light (or lines of sight). There was a remarkable precision and beauty to these ideas, which held a considerable fascination for the ancients, just as it does for us today.

Yet, with regard to our everyday lives, the implications of this mathematical precision for the actions of the world often appeared unexciting and limited, despite the fact that the mathematics itself seemed to represent a deep truth. Accordingly, many people in ancient times would allow their imaginations to be carried away by their fascination with the subject and to take them far beyond the scope of what was appropriate. In astrology, for example, geometrical figures also often engendered mystical and occult connotations, such as with the supposed magical powers of pentagrams and heptagrams. And there was an entirely suppositional attempted association between Platonic solids and the basic elementary states of matter (see Fig. 1.1). It would not be for many centuries that the deeper understanding that we presently have, concerning the actual relationships between mass, gravity, geometry, planetary motion, and the behaviour of light, could come about.

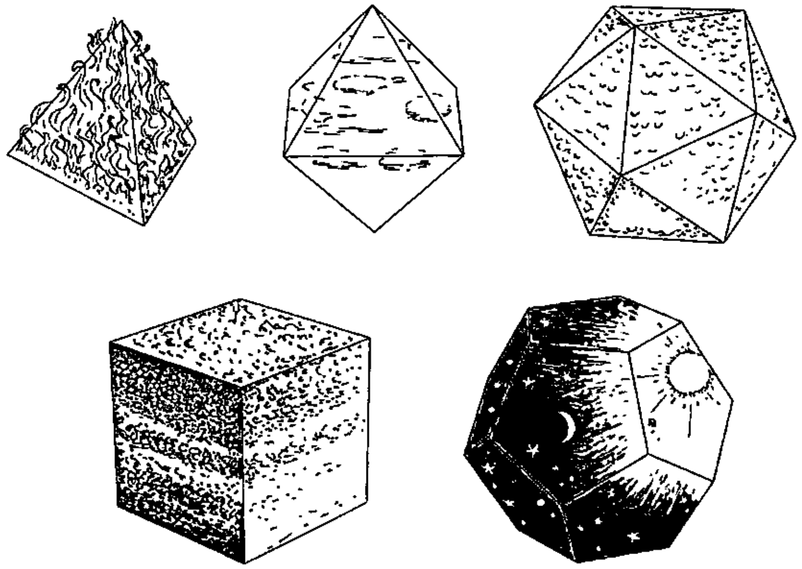

Fig. 1.1 A fanciful association, made by the ancient Greeks, between the five Platonic solids and the four 'elements' (fire, air, water, and earth), together with the heavenly firmament represented by the dodecahedron.

1.2 Mathematical truth

The first steps towards an understanding of the real influences controlling Nature required a disentangling of the true from the purely suppositional. But the ancients needed to achieve something else first, before they would be in any position to do this reliably for their understanding of Nature. What they had to do first was to discover how to disentangle the true from the suppositional in *mathematics*. A procedure was required for telling whether a given mathematical assertion is or is not to be trusted as true. Until that preliminary issue could be settled in a reasonable way, there would be little hope of seriously addressing those more difficult problems concerning forces that control the behaviour of the world and whatever their relations might be to mathematical truth. This realization that the key to the understanding of Nature lay within an unassailable mathematics was perhaps the first major breakthrough in science.

Although mathematical truths of various kinds had been surmised since ancient Egyptian and Babylonian times, it was not until the great Greek philosophers Thales of Miletus (*c*.625–547 BC) and

Pythagoras[1]* of Samos (*c.*572–497 BC) began to introduce the notion of *mathematical proof* that the first firm foundation stone of mathematical understanding—and therefore of science itself—was laid. Thales may have been the first to introduce this notion of proof, but it seems to have been the Pythagoreans who first made important use of it to establish things that were not otherwise obvious. Pythagoras also appeared to have a strong vision of the importance of *number*, and of arithmetical concepts, in governing the actions of the physical world. It is said that a big factor in this realization was his noticing that the most beautiful harmonies produced by lyres or flutes corresponded to the simplest fractional ratios between the lengths of vibrating strings or pipes. He is said to have introduced the 'Pythagorean scale', the numerical ratios of what we now know to be frequencies determining the principal intervals on which Western music is essentially based.[2] The famous *Pythagorean theorem*, asserting that the square on the hypotenuse of a right-angled triangle is equal to the sum of the squares on the other two sides, perhaps more than anything else, showed that indeed there is a precise relationship between the arithmetic of numbers and the geometry of physical space (see Chapter 2).

He had a considerable band of followers—the *Pythagoreans*—situated in the city of Croton, in what is now southern Italy, but their influence on the outside world was hindered by the fact that the members of the Pythagorean brotherhood were all sworn to secrecy. Accordingly, almost all of their detailed conclusions have been lost. Nonetheless, some of these conclusions were leaked out, with unfortunate consequences for the 'moles'—on at least one occasion, death by drowning!

In the long run, the influence of the Pythagoreans on the progress of human thought has been enormous. For the first time, with mathematical proof, it was possible to make significant assertions of an unassailable nature, so that they would hold just as true even today as at the time that they were made, no matter how our knowledge of the world has progressed since then. The truly timeless nature of mathematics was beginning to be revealed.

But what is a mathematical proof? A proof, in mathematics, is an impeccable argument, using only the methods of pure logical reasoning, which enables one to infer the validity of a given mathematical assertion from the pre-established validity of other mathematical assertions, or from some particular primitive assertions—the *axioms*—whose validity is taken to be self-evident. Once such a mathematical assertion has been established in this way, it is referred to as a *theorem*.

Many of the theorems that the Pythagoreans were concerned with were geometrical in nature; others were assertions simply about numbers. Those that were concerned merely with numbers have a perfectly unambiguous validity today, just as they did in the time of Pythagoras. What about the *geometrical* theorems that the Pythagoreans had obtained using their procedures of mathematical proof? They too have a clear validity today, but now there is a complicating issue. It is an issue whose nature is more obvious to us from our modern vantage point than it was at that time of Pythagoras. The ancients knew of only one kind of geometry, namely that which we now refer to as *Euclidean geometry*, but now we know of many other types. Thus, in considering the geometrical theorems of ancient Greek times, it becomes important to specify that the notion of geometry being referred to is indeed Euclid's geometry. (I shall be more explicit about these issues in §2.4, where an important example of non-Euclidean geometry will be given.)

Euclidean geometry is a specific mathematical structure, with its own specific axioms (including some less assured assertions referred to as postulates), which provided an excellent approximation to a particular aspect of the physical world. That was the aspect of reality, well familiar to the ancient Greeks, which referred to the laws governing the geometry of rigid objects and their relations to other rigid objects, as they are moved around in 3-dimensional space. Certain of these properties were so familiar and self-consistent that they tended to become regarded as 'self-evident' mathematical truths and were taken as axioms (or postulates). As we shall be seeing in Chapters 17–19 and §§27.8,11, Einstein's general relativity—and even the Minkowskian spacetime of special relativity—provide geometries for the physical universe that are different from, and yet more accurate than, the geometry of Euclid, despite the fact that the Euclidean geometry of the ancients was already extraordinarily accurate. Thus, we must be careful, when considering geometrical assertions, whether to trust the 'axioms' as being, in any sense, actually *true*.

But what does 'true' mean, in this context? The difficulty was well appreciated by the great ancient Greek philosopher Plato, who lived in Athens from *c.*429 to 347 BC, about a century and a half after Pythagoras. Plato made it clear that the mathematical propositions—the things that could be regarded as unassailably true—referred not to actual physical objects (like the approximate squares, triangles, circles, spheres, and cubes that might be constructed from marks in the sand, or from wood or stone) but to certain idealized entities. He envisaged that these ideal entities inhabited a different world, distinct from the physical world. Today, we might refer to this world as the *Platonic world of mathematical forms*. Physical structures, such as squares, circles, or triangles cut from papyrus, or marked on a flat surface, or perhaps cubes, tetrahedra, or spheres carved from marble, might conform to these ideals very closely, but only approximately. The actual *mathematical* squares, cubes, circles, spheres,

*Notes, indicated in the text by superscript numbers, are gathered at the ends of the chapter (in this case on p. 23).

triangles, etc., would not be part of the physical world, but would be inhabitants of Plato's idealized mathematical world of forms.

1.3 Is Plato's mathematical world 'real'?

This was an extraordinary idea for its time, and it has turned out to be a very powerful one. But does the Platonic mathematical world actually exist, in any meaningful sense? Many people, including philosophers, might regard such a 'world' as a complete fiction—a product merely of our unrestrained imaginations. Yet the Platonic viewpoint is indeed an immensely valuable one. It tells us to be careful to distinguish the precise mathematical entities from the approximations that we see around us in the world of physical things. Moreover, it provides us with the blueprint according to which modern science has proceeded ever since. Scientists will put forward models of the world—or, rather, of certain aspects of the world—and these models may be tested against previous observation and against the results of carefully designed experiment. The models are deemed to be appropriate if they survive such rigorous examination and if, in addition, they are internally consistent structures. The important point about these models, for our present discussion, is that they are basically purely abstract *mathematical* models. The very question of the internal consistency of a scientific model, in particular, is one that requires that the model be precisely specified. The required precision demands that the model be a mathematical one, for otherwise one cannot be sure that these questions have well-defined answers.

If the model itself is to be assigned any kind of 'existence', then this existence is located within the Platonic world of mathematical forms. Of course, one might take a contrary viewpoint: namely that the model is itself to have existence only within our various *minds*, rather than to take Plato's world to be in any sense absolute and 'real'. Yet, there is something important to be gained in regarding mathematical structures as having a reality of their own. For our individual minds are notoriously imprecise, unreliable, and inconsistent in their judgements. The precision, reliability, and consistency that are required by our scientific theories demand something beyond any one of our individual (untrustworthy) minds. In mathematics, we find a far greater robustness than can be located in any particular mind. Does this not point to something outside ourselves, with a reality that lies beyond what each individual can achieve?

Nevertheless, one might still take the alternative view that the mathematical world has no independent existence, and consists merely of certain ideas which have been distilled from our various minds and which have been found to be totally trustworthy and are agreed by all.

Yet even this viewpoint seems to leave us far short of what is required. Do we mean 'agreed by all', for example, or 'agreed by those who are in their right minds', or 'agreed by all those who have a Ph.D. in mathematics' (not much use in Plato's day) and who have a right to venture an 'authoritative' opinion? There seems to be a danger of circularity here; for to judge whether or not someone is 'in his or her right mind' requires some external standard. So also does the meaning of 'authoritative', unless some standard of an unscientific nature such as 'majority opinion' were to be adopted (and it should be made clear that majority opinion, no matter how important it may be for democratic government, should in no way be used as the criterion for scientific acceptability). Mathematics itself indeed seems to have a robustness that goes far beyond what any individual mathematician is capable of perceiving. Those who work in this subject, whether they are actively engaged in mathematical research or just using results that have been obtained by others, usually feel that they are merely explorers in a world that lies far beyond themselves—a world which possesses an objectivity that transcends mere opinion, be that opinion their own or the surmise of others, no matter how expert those others might be.

It may be helpful if I put the case for the actual existence of the Platonic world in a different form. What I mean by this 'existence' is really just the objectivity of mathematical truth. Platonic existence, as I see it, refers to the existence of an objective external standard that is not dependent upon our individual opinions nor upon our particular culture. Such 'existence' could also refer to things other than mathematics, such as to morality or aesthetics (cf. §1.5), but I am here concerned just with mathematical objectivity, which seems to be a much clearer issue.

Let me illustrate this issue by considering one famous example of a mathematical truth, and relate it to the question of 'objectivity'. In 1637, Pierre de Fermat made his famous assertion now known as 'Fermat's Last Theorem' (that no positive *n*th power[3] of an integer, i.e. of a whole number, can be the sum of two other positive *n*th powers if n is an integer greater than 2), which he wrote down in the margin of his copy of the *Arithmetica*, a book written by the 3rd-century Greek mathematician Diophantos. In this margin, Fermat also noted: 'I have discovered a truly marvellous proof of this, which this margin is too narrow to contain.' Fermat's mathematical assertion remained unconfirmed for over 350 years, despite concerted efforts by numerous outstanding mathematicians. A proof was finally published in 1995 by Andrew Wiles (depending on the earlier work of various other mathematicians), and this proof has now been accepted as a valid argument by the mathematical community.

Now, do we take the view that Fermat's assertion was always true, long before Fermat actually made it, or is its validity a purely cultural matter, dependent upon whatever might be the subjective standards of the community of human mathematicians? Let us try to suppose that the validity of the Fermat assertion is in fact a subjective matter. Then it would not be an absurdity for some other mathematician X to have come up with an actual and specific counter-example to the Fermat assertion, so long as X had done this before the date of 1995.[4] In such a circumstance, the mathematical community would have to accept the correctness of X's counter-example. From then on, any effort on the part of Wiles to prove the Fermat assertion would have to be fruitless, for the reason that X had got his argument in first and, as a result, the Fermat assertion would now be false! Moreover, we could ask the further question as to whether, consequent upon the correctness of X's forthcoming counter-example, Fermat himself would necessarily have been mistaken in believing in the soundness of his 'truly marvellous proof', at the time that he wrote his marginal note. On the subjective view of mathematical truth, it could possibly have been the case that Fermat had a valid proof (which would have been accepted as such by his peers at the time, had he revealed it) and that it was Fermat's secretiveness that allowed the possibility of X later obtaining a counter-example! I think that virtually all mathematicians, irrespective of their professed attitudes to 'Platonism', would regard such possibilities as patently absurd.

Of course, it might still be the case that Wiles's argument in fact contains an error and that the Fermat assertion is indeed false. Or there could be a fundamental error in Wiles's argument but the Fermat assertion is true nevertheless. Or it might be that Wiles's argument is correct in its essentials while containing 'non-rigorous steps' that would not be up to the standard of some future rules of mathematical acceptability. But these issues do not address the point that I am getting at here. The issue is the objectivity of the Fermat assertion itself, not whether anyone's particular demonstration of it (or of its negation) might happen to be convincing to the mathematical community of any particular time.

It should perhaps be mentioned that, from the point of view of mathematical logic, the Fermat assertion is actually a mathematical statement of a particularly simple kind,[5] whose objectivity is especially apparent. Only a tiny minority[6] of mathematicians would regard the truth of such assertions as being in any way 'subjective'—although there might be some subjectivity about the types of argument that would be regarded as being convincing. However, there are other kinds of mathematical assertion whose truth could plausibly be regarded as being a 'matter of opinion'. Perhaps the best known of such assertions is the *axiom of choice*. It is not important for us, now, to know what the axiom of choice is. (I shall describe it in §16.3.) It is cited here only as an example. Most mathematicians would probably regard the axiom of choice as 'obviously true', while

others may regard it as a somewhat questionable assertion which might even be false (and I am myself inclined, to some extent, towards this second viewpoint). Still others would take it as an assertion whose 'truth' is a mere matter of opinion or, rather, as something which can be taken one way or the other, depending upon which system of axioms and rules of procedure (a 'formal system'; see §16.6) one chooses to adhere to. Mathematicians who support this final viewpoint (but who accept the objectivity of the truth of particularly clear-cut mathematical statements, like the Fermat assertion discussed above) would be relatively weak Platonists. Those who adhere to objectivity with regard to the truth of the axiom of choice would be stronger Platonists.

I shall come back to the axiom of choice in §16.3, since it has some relevance to the mathematics underlying the behaviour of the physical world, despite the fact that it is not addressed much in physical theory. For the moment, it will be appropriate not to worry overly about this issue. If the axiom of choice can be settled one way or the other by some appropriate form of unassailable mathematical reasoning,[7] then its truth is indeed an entirely objective matter, and either it belongs to the Platonic world or its negation does, in the sense that I am interpreting this term 'Platonic world'. If the axiom of choice is, on the other hand, a mere matter of opinion or of arbitrary decision, then the Platonic world of absolute mathematical forms contains neither the axiom of choice nor its negation (although it could contain assertions of the form 'such-and-such follows from the axiom of choice' or 'the axiom of choice is a theorem according to the rules of such-and-such mathematical system').

The mathematical assertions that can belong to Plato's world are precisely those that are objectively true. Indeed, I would regard mathematical objectivity as really what mathematical Platonism is all about. To say that some mathematical assertion has a Platonic existence is merely to say that it is true in an objective sense. A similar comment applies to mathematical *notions*—such as the concept of the number 7, for example, or the rule of multiplication of integers, or the idea that some set contains infinitely many elements—all of which have a Platonic existence because they are objective notions. To my way of thinking, Platonic existence is simply a matter of objectivity and, accordingly, should certainly not be viewed as something 'mystical' or 'unscientific', despite the fact that some people regard it that way.

As with the axiom of choice, however, questions as to whether some particular proposal for a mathematical entity is or is not to be regarded as having objective existence can be delicate and sometimes technical. Despite this, we certainly need not be mathematicians to appreciate the general robustness of many mathematical concepts. In Fig. 1.2, I have depicted various small portions of that famous mathematical entity known

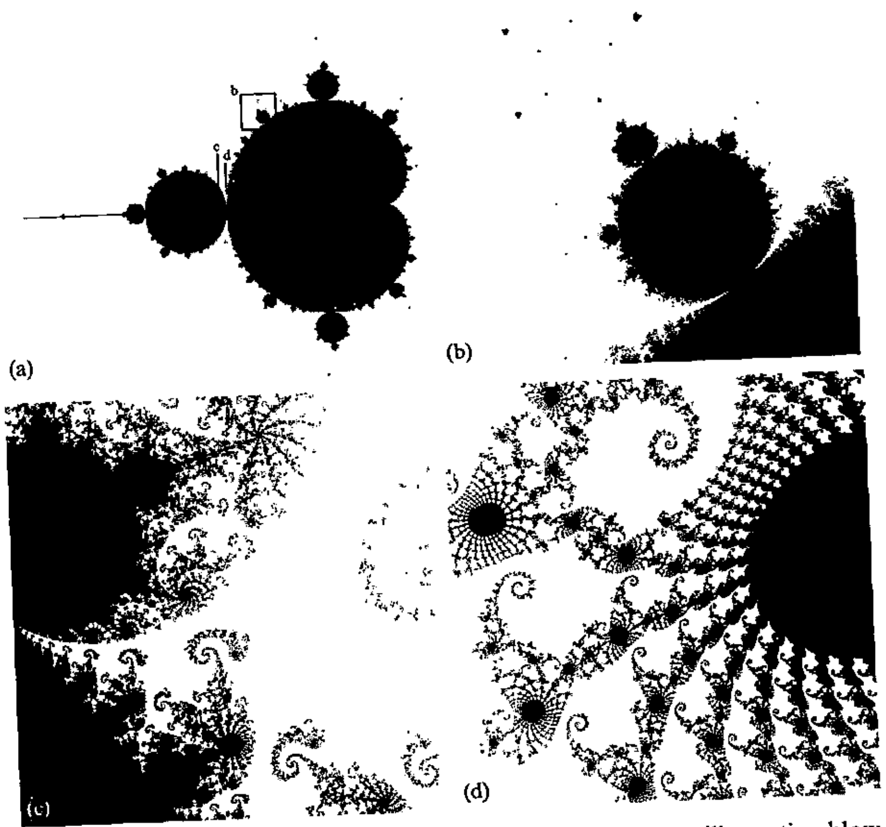

Fig. 1.2 (a) The Mandelbrot set. (b), (c), and (d) Some details, illustrating blow-ups of those regions correspondingly marked in Fig. 1.2a, magnified by respective linear factors 11.6, 168.9, and 1042 (and caps 300, 300, 200, 200; see Note 4.1).

as the *Mandelbrot set*. The set has an extraordinarily elaborate structure, but it is not of any human design. Remarkably, this structure is defined by a mathematical rule of particular simplicity. We shall come to this explicitly in §4.5, but it would distract us from our present purposes if I were to try to provide this rule in detail now.

The point that I wish to make is that no one, not even Benoit Mandelbrot himself when he first caught sight of the incredible complications in the fine details of the set, had any real preconception of the set's extraordinary richness. The Mandelbrot set was certainly no invention of any human mind. The set is just objectively there in the mathematics itself. If it has meaning to assign an actual existence to the Mandelbrot set, then that existence is not within our minds, for no one can fully comprehend the set's

endless variety and unlimited complication. Nor can its existence lie within the multitude of computer printouts that begin to capture some of its incredible sophistication and detail, for at best those printouts capture but a shadow of an approximation to the set itself. Yet it has a robustness that is beyond any doubt; for the same structure is revealed—in all its perceivable details, to greater and greater fineness the more closely it is examined—independently of the mathematician or computer that examines it. Its existence can only be within the Platonic world of mathematical forms.

I am aware that there will still be many readers who find difficulty with assigning any kind of actual existence to mathematical structures. Let me make the request of such readers that they merely broaden their notion of what the term 'existence' can mean to them. The mathematical forms of Plato's world clearly do not have the same kind of existence as do ordinary physical objects such as tables and chairs. They do not have spatial locations; nor do they exist in time. Objective mathematical notions must be thought of as timeless entities and are not to be regarded as being conjured into existence at the moment that they are first humanly perceived. The particular swirls of the Mandelbrot set that are depicted in Fig. 1.2c or 1.2d did not attain their existence at the moment that they were first seen on a computer screen or printout. Nor did they come about when the general idea behind the Mandelbrot set was first humanly put forth—not actually first by Mandelbrot, as it happened, but by R. Brooks and J. P. Matelski, in 1981, or perhaps earlier. For certainly neither Brooks nor Matelski, nor initially even Mandelbrot himself, had any real conception of the elaborate detailed designs that we see in Fig. 1.2c and 1.2d. Those designs were already 'in existence' since the beginning of time, in the potential timeless sense that they would necessarily be revealed precisely in the form that we perceive them today, no matter at what time or in what location some perceiving being might have chosen to examine them.

1.4 Three worlds and three deep mysteries

Thus, mathematical existence is different not only from physical existence but also from an existence that is assigned by our mental perceptions. Yet there is a deep and mysterious connection with each of those other two forms of existence: the physical and the mental. In Fig. 1.3, I have schematically indicated all of these three forms of existence—the physical, the mental, and the Platonic mathematical—as entities belonging to three separate 'worlds', drawn schematically as spheres. The mysterious connections between the worlds are also indicated, where in drawing the diagram

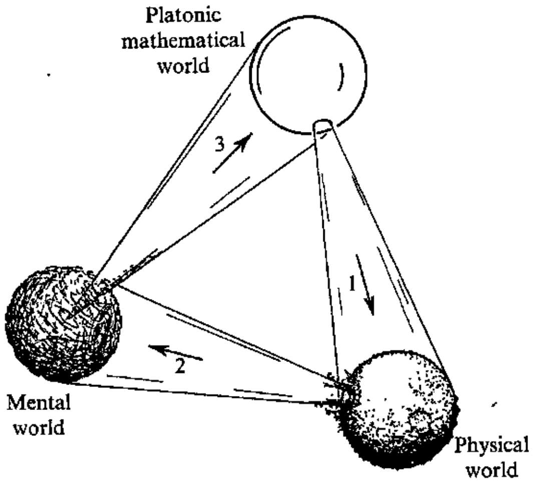

Fig. 1.3 Three 'worlds'—the Platonic mathematical, the physical, and the mental—and the three profound mysteries in the connections between them.

I have imposed upon the reader some of my beliefs, or prejudices, concerning these mysteries.

It may be noted, with regard to the *first* of these mysteries—relating the Platonic mathematical world to the physical world—that I am allowing that only a small part of the world of mathematics need have relevance to the workings of the physical world. It is certainly the case that the vast preponderance of the activities of pure mathematicians today has no obvious connection with physics, nor with any other science (cf. §34.9), although we may be frequently surprised by unexpected important applications. Likewise, in relation to the *second* mystery, whereby mentality comes about in association with certain physical structures (most specifically, healthy, wakeful human brains), I am not insisting that the majority of physical structures need induce mentality. While the brain of a cat may indeed evoke mental qualities, I am not requiring the same for a rock. Finally, for the *third* mystery, I regard it as self-evident that only a small fraction of our mental activity need be concerned with absolute mathematical truth! (More likely we are concerned with the multifarious irritations, pleasures, worries, excitements, and the like, that fill our daily lives.) These three facts are represented in the smallness of the base of the connection of each world with the next, the worlds being taken in a clockwise sense in the diagram. However, it is in the encompassing of each entire world within the scope of its connection with the world preceding it that I am revealing my prejudices.

Thus, according to Fig. 1.3, the entire physical world is depicted as being governed according to mathematical laws. We shall be seeing in later chapters that there is powerful (but incomplete) evidence in support of this contention. On this view, everything in the physical universe is indeed governed in completely precise detail by mathematical principles—perhaps by equations, such as those we shall be learning about in chapters to follow, or perhaps by some future mathematical notions fundamentally different from those which we would today label by the term 'equations'. If this is right, then even our own physical actions would be entirely subject to such ultimate mathematical control, where 'control' might still allow for some random behaviour governed by strict probabilistic principles.

Many people feel uncomfortable with contentions of this kind, and I must confess to having some unease with it myself. Nonetheless, my personal prejudices are indeed to favour a viewpoint of this general nature, since it is hard to see how any line can be drawn to separate physical actions under mathematical control from those which might lie beyond it. In my own view, the unease that many readers may share with me on this issue partly arises from a very limited notion of what 'mathematical control' might entail. Part of the purpose of this book is to touch upon, and to reveal to the reader, some of the extraordinary richness, power, and beauty that can spring forth once the right mathematical notions are hit upon.

In the Mandelbrot set alone, as illustrated in Fig. 1.2, we can begin to catch a glimpse of the scope and beauty inherent in such things. But even these structures inhabit a very limited corner of mathematics as a whole, where behaviour is governed by strict computational control. Beyond this corner is an incredible potential richness. How do I really feel about the possibility that all my actions, and those of my friends, are ultimately governed by mathematical principles of this kind? I can live with that. I would, indeed, prefer to have these actions controlled by something residing in some such aspect of Plato's fabulous mathematical world than to have them be subject to the kind of simplistic base motives, such as pleasure-seeking, personal greed, or aggressive violence, that many would argue to be the implications of a strictly scientific standpoint.

Yet, I can well imagine that a good many readers will still have difficulty in accepting that all actions in the universe could be entirely subject to mathematical laws. Likewise, many might object to two other prejudices of mine that are implicit in Fig. 1.3. They might feel, for example, that I am taking too hard-boiled a scientific attitude by drawing my diagram in a way that implies that all of mentality has its roots in physicality. This is indeed a prejudice, for while it is true that we have no reasonable scientific evidence for the existence of 'minds' that do not have a physical basis, we cannot be completely sure. Moreover, many of a religious persuasion would argue strongly for the possibility of physically independent minds and might appeal to what they regard as powerful evidence of a different kind from that which is revealed by ordinary science.

A further prejudice of mine is reflected in the fact that in Fig. 1.3 I have represented the entire Platonic world to be within the compass of mentality. This is intended to indicate that—at least in principle—there are no mathematical truths that are beyond the scope of reason. Of course, there are mathematical statements (even straightforward arithmetical addition sums) that are so vastly complicated that no one could have the mental fortitude to carry out the necessary reasoning. However, such things would be *potentially* within the scope of (human) mentality and would be consistent with the meaning of Fig. 1.3 as I have intended to represent it. One must, nevertheless, consider that there might be other mathematical statements that lie outside even the potential compass of reason, and these would violate the intention behind Fig. 1.3. (This matter will be considered at greater length in §16.6, where its relation to Gödel's famous incompleteness theorem will be discussed.)[8]

In Fig. 1.4, as a concession to those who do not share all my personal prejudices on these matters, I have redrawn the connections between the three worlds in order to allow for all three of these possible violations of my prejudices. Accordingly, the possibility of physical action beyond the scope of mathematical control is now taken into account. The diagram also allows for the belief that there might be mentality that is not rooted in physical structures. Finally, it permits the existence of true mathematical assertions whose truth is in principle inaccessible to reason and insight.

This extended picture presents further potential mysteries that lie even beyond those which I have allowed for in my own preferred picture of the world, as depicted in Fig. 1.3. In my opinion, the more tightly organized scientific viewpoint of Fig. 1.3 has mysteries enough. These mysteries are not removed by passing to the more relaxed scheme of Fig. 1.4. For it remains a deep puzzle why mathematical laws should apply to the world with such phenomenal precision. (We shall be glimpsing something of the extraordinary accuracy of the basic physical theories in §19.8, §26.7, and §27.13.) Moreover, it is not just the precision but also the subtle sophistication and mathematical beauty of these successful theories that is profoundly mysterious. There is also an undoubted deep mystery in how it can come to pass that appropriately organized physical material—and here I refer specifically to living human (or animal) brains—can somehow conjure up the mental quality of conscious awareness. Finally, there is also a mystery about how it is that we perceive mathematical truth. It is not just that our brains are programmed to 'calculate' in reliable ways. There is something much more profound than that in the insights that even the humblest among us possess when we appreciate, for example, the actual meanings of the terms 'zero', 'one', 'two', 'three', 'four', etc.[9]

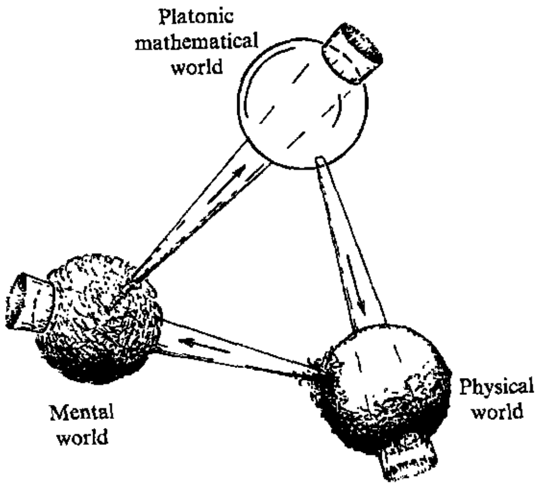

Fig. 1.4 A redrawing of Fig. 1.3 in which violations of three of the prejudices of the author are allowed for.

Some of the issues that arise in connection with this third mystery will be our concern in the next chapter (and more explicitly in §§16.5,6) in relation to the notion of *mathematical proof*. But the main thrust of this book has to do with the first of these mysteries: the remarkable relationship between mathematics and the actual behaviour of the physical world. No proper appreciation of the extraordinary power of modern science can be achieved without at least some acquaintance with these mathematical ideas. No doubt, many readers may find themselves daunted by the prospect of having to come to terms with such mathematics in order to arrive at this appreciation. Yet, I have the optimistic belief that they may not find all these things to be so bad as they fear. Moreover, I hope that I may persuade many readers that, despite what she or he may have previously perceived, mathematics can be fun!

I shall not be especially concerned here with the second of the mysteries depicted in Figs. 1.3 and 1.4, namely the issue of how it is that mentality—most particularly conscious awareness—can come about in association with appropriate physical structures (although I shall touch upon this deep question in §34.7). There will be enough to keep us busy in exploring the physical universe and its associated mathematical laws. In addition, the issues concerning mentality are profoundly contentious, and it would distract from the purpose of this book if we were to get embroiled in them. Perhaps one comment will not be amiss here, however. This is that, in my own opinion, there is little chance that any deep understanding of the nature of the mind can come about without our first learning much more about the very basis of physical reality. As will become clear from the discussions that will be presented in later chapters, I believe that major revolutions are required in our physical understanding. Until these revolutions have come to pass, it is, in my view, greatly optimistic to expect that much real progress can be made in understanding the actual nature of mental processes.[10]

1.5 The Good, the True, and the Beautiful

In relation to this, there is a further set of issues raised by Figs. 1.3 and 1.4. I have taken Plato's notion of a 'world of ideal forms' only in the limited sense of mathematical forms. Mathematics is crucially concerned with the particular ideal of *Truth*. Plato himself would have insisted that there are two other fundamental absolute ideals, namely that of the *Beautiful* and of the *Good*. I am not at all averse to admitting to the existence of such ideals, and to allowing the Platonic world to be extended so as to contain absolutes of this nature.

Indeed, we shall later be encountering some of the remarkable interrelations between truth and beauty that both illuminate and confuse the issues of the discovery and acceptance of physical theories (see §§34.2,5,9 particularly; see also Fig. 34.1). Moreover, quite apart from the undoubted (though often ambiguous) role of beauty for the mathematics underlying the workings of the physical world, aesthetic criteria are fundamental to the development of mathematical ideas for their own sake, providing both the drive towards discovery and a powerful guide to truth. I would even surmise that an important element in the mathematician's common conviction that an external Platonic world actually has an existence independent of ourselves comes from the extraordinary unexpected hidden beauty that the ideas themselves so frequently reveal.

Of less obvious relevance here—but of clear importance in the broader context—is the question of an absolute ideal of morality: what is good and what is bad, and how do our minds perceive these values? Morality has a profound connection with the mental world, since it is so intimately related to the values assigned by conscious beings and, more importantly, to the very presence of consciousness itself. It is hard to see what morality might mean in the absence of sentient beings. As science and technology progress, an understanding of the physical circumstances under which mentality is manifested becomes more and more relevant. I believe that it is more important than ever, in today's technological culture, that scientific questions should not be divorced from their moral implications. But these issues would take us too far afield from the immediate scope of this book. We need to address the question of separating true from false before we can adequately attempt to apply such understanding to separate good from bad.

There is, finally, a further mystery concerning Fig. 1.3, which I have left to the last. I have deliberately drawn the figure so as to illustrate a paradox. How can it be that, in accordance with my own prejudices, each world appears to encompass the next one in its entirety? I do not regard this issue as a reason for abandoning my prejudices, but merely for demonstrating the presence of an even deeper mystery that transcends those which I have been pointing to above. There may be a sense in which the three worlds are not separate at all, but merely reflect, individually, aspects of a deeper truth about the world as a whole of which we have little conception at the present time. We have a long way to go before such matters can be properly illuminated.

I have allowed myself to stray too much from the issues that will concern us here. The main purpose of this chapter has been to emphasize the central importance that mathematics has in science, both ancient and modern. Let us now take a glimpse into Plato's world—at least into a relatively small but important part of that world, of particular relevance to the nature of physical reality.

Notes

Section 1.2

1.1. Unfortunately, almost nothing reliable is known about Pythagoras, his life, his followers, or of their work, apart from their very existence and the recognition by Pythagoras of the role of simple ratios in musical harmony. See Burkert (1972). Yet much of great importance is commonly attributed to the Pythagoreans. Accordingly, I shall use the term 'Pythagorean' simply as a label, with no implication intended as to historical accuracy.

1.2. This is the pure 'diatonic scale' in which the frequencies (in inverse proportion to the lengths of the vibrating elements) are in the ratios 24 : 27 : 30 : 32 : 36 : 40 : 45 : 48, giving many instances of simple ratios, which underlie harmonies that are pleasing to the ear. The 'white notes' of a modern piano are tuned (according to a compromise between Pythagorean purity of harmony and the facility of key changes) as approximations to these Pythagorean ratios, according to the *equal temperament* scale, with relative frequencies $1:\alpha^2:\alpha^4:\alpha^5:\alpha^7:\alpha^9:\alpha^{11}:\alpha^{12}$, where $\alpha = \sqrt[12]{2} = 1.05946\ldots.$ (Note: α^5 means the fifth power of α, i.e. $\alpha \times \alpha \times \alpha \times \alpha \times \alpha$. The quantity $\sqrt[12]{2}$ is the twelfth root of 2, which is the number whose twelfth power is 2, i.e. $2^{1/12}$, so that $\alpha^{12} = 2$. See Note 1.3 and §5.2.)

Section 1.3

1.3. Recall from Note 1.2 that the *n*th power of a number is that number multiplied by itself *n* times. Thus, the third power of 5 is 125, written $5^3 = 125$; the fourth power of 3 is 81, written $3^4 = 81$; etc.

1.4. In fact, while Wiles was trying to fix a 'gap' in his proof of Fermat's Last Theorem which had become apparent after his initial presentation at Cambridge in June 1993, a rumour spread through the mathematical community that the mathematician Noam Elkies had found a counter-example to Fermat's assertion. Earlier, in 1988, Elkies had found a counter-example to Euler's conjecture—that there are no integer solutions to the equation $x^4 + y^4 + z^4 = w^4$—thereby proving it false. It was not implausible, therefore, that he had proved that Fermat's assertion also was false. However, the e-mail that started the rumour was dated 1 April and was revealed to be a spoof perpetrated by Henri Darmon; see Singh (1997), p. 293.

1.5. Technically it is a Π_1-sentence; see §16.6.

1.6. I realize that, in a sense, I am falling into my own trap by making such an assertion. The issue is not really whether the mathematicians taking such an extreme subjective view happen to constitute a tiny minority or not (and I have certainly not conducted a trustworthy survey among mathematicians on this point); the issue is whether such an extreme position is actually to be taken seriously. I leave it to the reader to judge.

1.7. Some readers may be aware of the results of Gödel and Cohen that the axiom of choice is independent of the more basic standard axioms of set theory (the Zermelo–Frankel axiom system). It should be made clear that the Gödel–Cohen argument does not in itself establish that the axiom of choice will never be settled one way or the other. This kind of point is stressed, for example, in the final section of Paul Cohen's book (Cohen 1966, Chap. 14, §13), except that, there, Cohen is more explicitly concerned with the *continuum hypothesis* than the axiom of choice; see §16.5.

Section 1.4

1.8. There is perhaps an irony here that a fully fledged anti-Platonist, who believes that mathematics is 'all in the mind' must also believe—so it seems—that there are no true mathematical statements that are in principle beyond reason. For example, if Fermat's Last Theorem had been inaccessible (in principle) to reason, then this anti-Platonist view would allow no validity either to its truth or to its falsity, such validity coming only through the mental act of perceiving some proof or disproof.

1.9. See e.g. Penrose (1997b).

1.10. My own views on the kind of change in our physical world-view that will be needed in order that conscious mentality may be accommodated are expressed in Penrose (1989, 1994, 1997a,1997b).

2
An ancient theorem and a modern question

2.1 The Pythagorean theorem

LET us consider the issue of geometry. What, indeed, are the different 'kinds of geometry' that were alluded to in the last chapter? To lead up to this issue, we shall return to our encounter with Pythagoras and consider that famous theorem that bears his name:[1] for any right-angled triangle, the square of the length of the hypotenuse (the side opposite the right angle) is equal to the sum of the squares of the lengths of the other two sides (Fig. 2.1). What reasons do we have for believing that this assertion is true? How, indeed, do we 'prove' the Pythagorean theorem? Many arguments are known. I wish to consider two such, chosen for their particular transparency, each of which has a different emphasis.

For the first, consider the pattern illustrated in Fig. 2.2. It is composed entirely of squares of two different sizes. It may be regarded as 'obvious' that this pattern can be continued indefinitely and that the entire plane is thereby covered in this regular repeating way, without gaps or overlaps, by squares of these two sizes. The repeating nature of this pattern is made manifest by the fact that if we mark the centres of the larger squares, they form the vertices of another system of squares, of a somewhat greater size than either, but tilted at an angle to the original ones (Fig. 2.3) and which alone will cover the entire plane. Each of these tilted squares is marked in exactly the same way, so that the markings on these squares fit together to

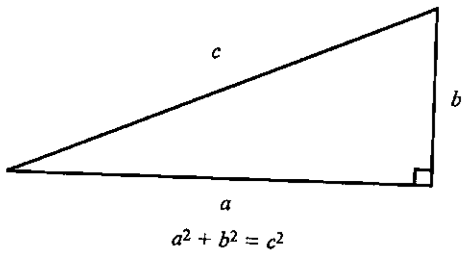

Fig. 2.1 The Pythagorean theorem: for any right-angled triangle, the squared length of the hypotenuse c is the sum of the squared lengths of the other two sides a and b.

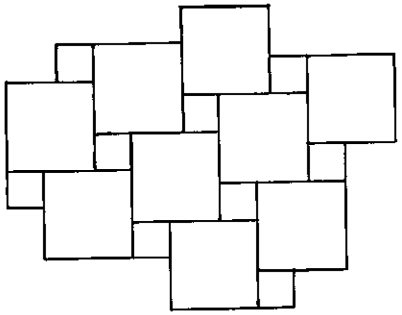

Fig. 2.2 A tessellation of the plane by squares of two different sizes.

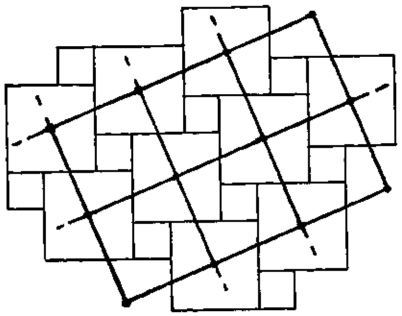

Fig. 2.3 The centres of the (say) larger squares form the vertices of a lattice of still larger squares, tilted at an angle.

form the original two-square pattern. The same would apply if, instead of taking the centres of the larger of the two squares of the original pattern, we chose any other point, together with its set of corresponding points throughout the pattern. The new pattern of tilted squares is just the same as before but moved along without rotation—i.e. by means of a motion referred to as a *translation*. For simplicity, we can now choose our starting point to be one of the corners in the original pattern (see Fig. 2.4).

It should be clear that the area of the tilted square must be equal to the sum of the areas of the two smaller squares—indeed the pieces into which the markings would subdivide this larger square can, for any starting point for the tilted squares, be moved around, without rotation, until they fit together to make the two smaller squares (e.g. Fig. 2.5). Moreover, it is evident from Fig. 2.4 that the edge-length of the large tilted square is the hypotenuse of a right-angled triangle whose two other sides have lengths equal to those of the two smaller squares. We have thus established the Pythagorean theorem: the square on the hypotenuse is equal to the sum of the squares on the other two sides.

The above argument does indeed provide the essentials of a simple proof of this theorem, and, moreover, it gives us some 'reason' for believing that the theorem has to be true, which might not be so obviously the case with some more formal argument given by a succession of logical steps without clear motivation. It should be pointed out, however, that there are several implicit assumptions that have gone into this argument. Not the least of these is the assumption that the seemingly obvious pattern of repeating squares shown in Fig. 2.2 or even in Fig. 2.6 is actually geometrically possible—or even, more critically, that a *square* is something geometrically possible! What do we mean by a 'square' after all? We normally think of a square as a plane figure, all of whose sides are equal and all of whose angles are right angles. What is a right angle? Well, we can imagine two

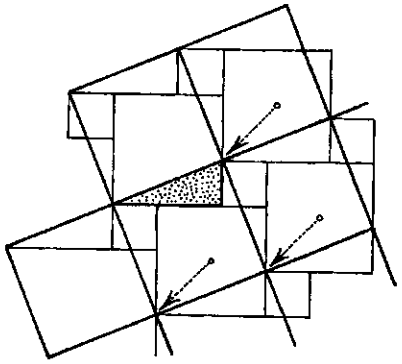

Fig. 2.4 The lattice of tilted squares can be shifted by a translation, here so that the vertices of the tilted lattice lie on vertices of the original two-square lattice, showing that the side-length of a tilted square is the hypotenuse of a right-angled triangle (shown shaded) whose other two side-lengths are those of the original two squares.

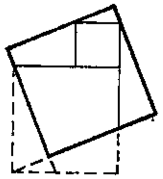

Fig. 2.5 For any particular starting point for the tilted square, such as that depicted, the tilted square is divided into pieces that fit together to make the two smaller squares.

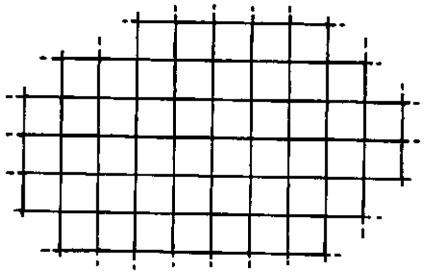

Fig. 2.6 The familiar lattice of equal squares. How do we know it exists?

straight lines crossing each other at some point, making four angles that are all equal. Each of these equal angles is then a right angle.

Let us now try to construct a square. Take three equal line segments AB, BC, and CD, where ABC and BCD are right angles, D and A being on the same side of the line BC, as in Fig. 2.7. The question arises: is AD the same length as the other three segments? Moreover, are the angles DAB and CDA also right angles? These angles should be equal to one another by a left–right symmetry in the figure, but are they actually right angles? This only seems obvious because of our familiarity with squares, or perhaps because we can recall from our schooldays some statement of Euclid that can be used to tell us that the sides BA and CD would have to be 'parallel' to each other, and some statement that any 'transversal' to a pair of parallels has to have corresponding angles equal, where it meets the two

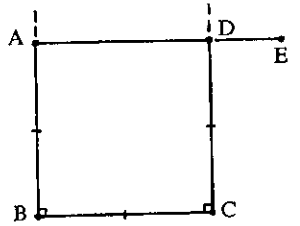

Fig. 2.7 Try to construct a square. Take ABC and BCD as right angles, with AB = BC = CD. Does it follow that DA is also equal to these lengths and that DAB and CDA are also right angles?

parallels. From this, it follows that the angle DAB would have to be equal to the angle complementary to ADC (i.e. to the angle EDC, in Fig. 2.7, ADE being straight) as well as being, as noted above, equal to the angle ADC. An angle (ADC) can only be equal to its complementary angle (EDC) if it is a right angle. We must also prove that the side AD has the same length as BC, but this now also follows, for example, from properties of transversals to the parallels BA and CD. So, it is indeed true that we can prove from this kind of Euclidean argument that squares, made up of right angles, actually do exist. But there is a deep issue hiding here.

2.2 Euclid's postulates

In building up his notion of geometry, Euclid took considerable care to see what assumptions his demonstrations depended upon.[2] In particular, he was careful to distinguish certain assertions called *axioms*—which were taken as self-evidently true, these being basically definitions of what he meant by points, lines, etc.—from the five *postulates*, which were assumptions whose validity seemed less certain, yet which appeared to be true of the geometry of our world. The final one of these assumptions, referred to as Euclid's fifth postulate, was considered to be less obvious than the others, and it was felt, for many centuries, that it ought to be possible to find a way of proving it from the other more evident postulates. Euclid's fifth postulate is commonly referred to as the *parallel postulate* and I shall follow this practice here.

Before discussing the parallel postulate, it is worth pointing out the nature of the other four of Euclid's postulates. The postulates are concerned with the geometry of the (Euclidean) plane, though Euclid also considered three-dimensional space later in his works. The basic elements of his plane geometry are points, straight lines, and circles. Here, I shall consider a 'straight line' (or simply a 'line') to be indefinitely extended in both directions; otherwise I refer to a 'line segment'. Euclid's *first* postulate effectively asserts that there is a (unique) straight line segment connecting any two points. His *second* postulate asserts the unlimited (continuous) extendibility of any straight line segment. His *third* postulate asserts the existence of a circle with any centre and with any value for its radius. Finally, his *fourth* postulate asserts the equality of all right angles.[3]

From a modern perspective, some of these postulates appear a little strange, particularly the fourth, but we must bear in mind the origin of the ideas underlying Euclid's geometry. Basically, he was concerned with the movement of idealized rigid bodies and the notion of *congruence* which was signalled when one such idealized rigid body was moved into coincidence with another. The equality of a right angle on one body with that on another had to do with the possibility of moving the one so that the lines forming its right angle would lie along the lines forming the right angle of the other. In effect, the fourth postulate is asserting the isotropy and homogeneity of space, so that a figure in one place could have the 'same' (i.e. congruent) geometrical shape as a figure in some other place. The second and third postulates express the idea that space is indefinitely extendible and without 'gaps' in it, whereas the first expresses the basic nature of a straight line segment. Although Euclid's way of looking at geometry was rather different from the way that we look at it today, his first four postulates basically encapsulated our present-day notion of a (two-dimensional) metric space with complete homogeneity and isotropy, and infinite in extent. In fact, such a picture seems to be in close accordance with the very large-scale spatial nature of the actual universe, according to modern cosmology, as we shall be coming to in §27.11 and §28.10.

What, then, is the nature of Euclid's fifth postulate, the parallel postulate? As Euclid essentially formulated this postulate, it asserts that if two straight line segments *a* and *b* in a plane both intersect another straight line *c* (so that *c* is what is called a *transversal* of *a* and *b*) such that the sum of the interior angles on the same side of *c* is less than two right angles, then *a* and *b*, when extended far enough on that side of *c*, will intersect somewhere (see Fig. 2.8a). An equivalent form of this postulate (sometimes referred to as *Playfair's axiom*) asserts that, for any straight line and for any point not on the line, there is a unique straight line through the point which is parallel to the line (see Fig. 2.8b). Here, 'parallel' lines would be two straight lines in the same plane that do not intersect each other (and recall that *my* 'lines' are fully extended entities, rather than Euclid's 'segments of lines').[2.1]

[2.1] Show that if Euclid's form of the parallel postulate holds, then Playfair's conclusion of the uniqueness of parallels must follow.

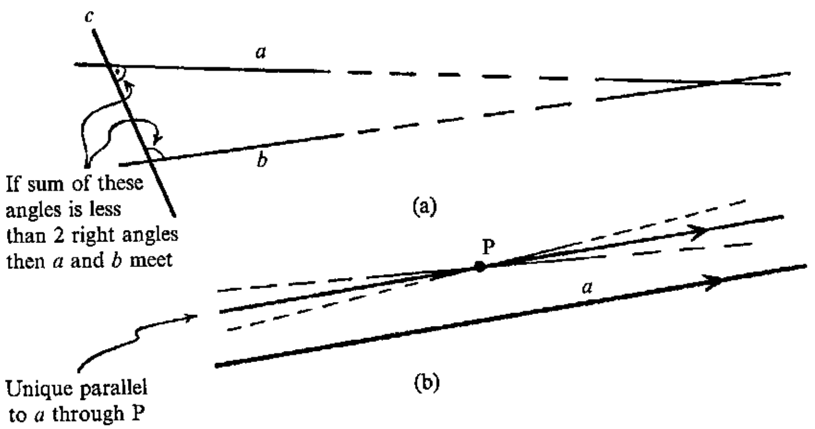
c
a
b
If sum of these angles is less than 2 right angles then a and b meet
(a)
P
a
Unique parallel to a through P
(b)

Fig. 2.8 (a) Euclid's parallel postulate. Lines a and b are transversals to a third line c, such that the interior angles where a and b meet c add to less than two right angles. Then a and b (assumed extended far enough) will ultimately intersect each other. (b) Playfair's (equivalent) axiom: if a is a line in a plane and P a point of the plane not on a, then there is just one line parallel to a through P, in the plane.

Once we have the parallel postulate, we can proceed to establish the property needed for the existence of a square. If a transversal to a pair of straight lines meets them so that the sum of the interior angles on one side of the transversal is two right angles, then one can show that the lines of the pair are indeed parallel. Moreover, it immediately follows that any other transversal of the pair has just the same angle property. This is basically just what we needed for the argument given above for the construction of our square. We see, indeed, that it is just the parallel postulate that we must use to show that our construction actually yields a square, with all its angles right angles and all its sides the same. Without the parallel postulate, we cannot establish that squares (in the normal sense where all their angles are right angles) actually exist.

It may seem to be merely a matter of mathematical pedantry to worry about precisely which assumptions are needed in order to provide a 'rigorous proof' of the existence of such an obvious thing as a square. Why should we really be concerned with such pedantic issues, when a 'square' is just that familiar figure that we all know about? Well, we shall be seeing shortly that Euclid actually showed some extraordinary perspicacity in worrying about such matters. Euclid's pedantry is related to a deep issue that has a great deal to say about the actual geometry of the universe, and in more than one way. In particular, it is not at all an obvious matter whether physical 'squares' exist on a cosmological scale

in the actual universe. This is a matter for observation, and the evidence at the moment appears to be conflicting (see §2.7 and §28.10).

2.3 Similar-areas proof of the Pythagorean theorem

I shall return to the mathematical significance of *not* assuming the parallel postulate in the next section. The relevant physical issues will be re-examined in §18.4, §27.11, §28.10, and §34.4. But, before discussing such matters, it will be instructive to turn to the other proof of the Pythagorean theorem that I had promised above.

One of the simplest ways to see that the Pythagorean assertion is indeed true in Euclidean geometry is to consider the configuration consisting of the given right-angled triangle subdivided into two smaller triangles by dropping a perpendicular from the right angle to the hypotenuse (Fig. 2.9). There are now three triangles depicted: the original one and the two into which it has now been subdivided. Clearly the area of the original triangle is the sum of the areas of the two smaller ones.

Now, it is a simple matter to see that these three triangles are all *similar* to one another. This means that they are all the same *shape* (though of different sizes), i.e. obtained from one another by a uniform expansion or contraction, together with a rigid motion. This follows because each of the three triangles possesses exactly the same angles, in some order. Each of the two smaller triangles has an angle in common with the largest one and one of the angles of each triangle is a right angle. The third angle must also agree because the sum of the angles in any triangle is always the same. Now, it is a general property of similar plane figures that their areas are in proportion to the squares of their corresponding linear dimensions. For each triangle, we can take this linear dimension to be its longest side, i.e. its hypotenuse. We note that the hypotenuse of each of the smaller triangles is

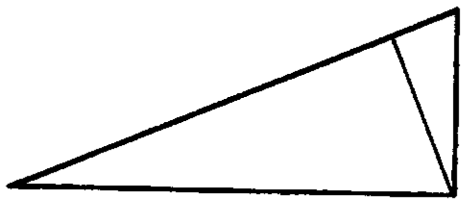

Fig. 2.9 Proof of the Pythagorean theorem using similar triangles. Take a right-angled triangle and drop a perpendicular from its right angle to its hypotenuse. The two triangles into which the original triangle is now divided have areas which sum to that of the original triangle. All three triangles are similar, so their areas are in proportion to the squares of their respective hypotenuses. The Pythagorean theorem follows.

the same as one of the (non-hypotenuse) sides of the original triangle. Thus, it follows at once (from the fact that the area of the original triangle is the sum of the areas of the other two) that the square on the hypotenuse on the original triangle is indeed the sum of the squares on the other two sides: *the Pythagorean theorem*!

There are, again, some particular assumptions in this argument that we shall need to examine. One important ingredient of the argument is the fact that the angles of a triangle always add up to the same value. (This value of this sum is of course 180°, but Euclid would have referred to it as 'two right angles'. The more modern 'natural' mathematical description is to say that the angles of a triangle, in Euclid's geometry, add up to π. This is to use radians for the absolute measure of angle, where the degree sign '°' counts as $\pi/180$, so we can write $180° = \pi$.) The usual proof is depicted in Fig. 2.10. We extend CA to E and draw a line AD, through A, which is parallel to CB. Then (as follows from the parallel postulate) the angles EAD and ACB are equal, and also DAB and CBA are equal. Since the angles EAD, DAB, and BAC add up to π (or to 180°, or to two right angles), so also must the three angles ACB, CBA, and BAC of the triangle—as was required to prove. But notice that the parallel postulate was used here.

This proof of the Pythagorean theorem also makes use of the fact that the areas of similar figures are in proportion to the squares of any linear measure of their sizes. (Here we chose the hypotenuse of each triangle to represent this linear measure.) This fact not only depends on the very existence of similar figures of different sizes—which for the triangles of Fig. 2.9 we established using the parallel postulate—but also on some more sophisticated issues that relate to how we actually define 'area' for non-rectangular shapes. These general matters are addressed in terms of the carrying out of limiting procedures, and I do not want to enter into this kind of discussion just for the moment. It will take us into some deeper issues related to the kind of numbers that are used in geometry. The question will be returned to in §§3.1–3.

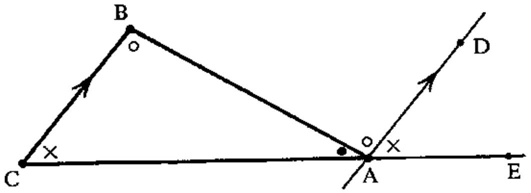

Fig. 2.10 Proof that the sum of the angles of a triangle ABC sums to π (= 180° = two right angles). Extend CA to E; draw AD parallel to CB. It follows from the parallel postulate that the angles EAD and ACB are equal and the angles DAB and CBA are equal. Since the angles EAD, DAB, and BAC sum to π, so also do the angles ACB, CBA, and BAC.

An important message of the discussion in the preceding sections is that the Pythagorean theorem seems to depend on the parallel postulate. Is this really so? Suppose the parallel postulate were false? Does that mean that the Pythagorean theorem might itself actually be false? Does such a possibility make any sense? Let us try to address the question of what would happen if the parallel postulate is indeed allowed to be taken to be false. We shall seem to be entering a mysterious make-belief world, where the geometry that we learned at school is turned all topsy-turvy. Indeed, but we shall find that there is also a deeper purpose here.

2.4 Hyperbolic geometry: conformal picture

Have a look at the picture in Fig. 2.11. It is a reproduction of one of M. C. Escher's woodcuts, called *Circle Limit I*. It actually provides us with a very accurate representation of a kind of geometry—called *hyperbolic* (or sometimes *Lobachevskian*) geometry—in which the parallel postulate is false, the Pythagorean theorem fails to hold, and the angles of a triangle do not add to π. Moreover, for a shape of a given size, there does not, in general, exist a similar shape of a larger size.

In Fig. 2.11, Escher has used a particular representation of hyperbolic geometry in which the entire 'universe' of the hyperbolic plane is 'squashed' into the interior of a circle in an ordinary Euclidean plane. The bounding circle represents 'infinity' for this hyperbolic universe. We can see that, in Escher's picture, the fish appear to get very crowded as they get close to this bounding circle. But we must think of this as an illusion. Imagine that you happened to be one of the fish. Then whether you are situated close to the rim of Escher's picture or close to its centre, the entire (hyperbolic) universe will look the same to you. The notion of 'distance' in this geometry does not agree with that of the Euclidean plane in terms of which it has been represented. As we look down upon Escher's picture from our Euclidean perspective, the fish near the bounding circle appear to us to be getting very tiny. But from the 'hyperbolic' perspective of the white or the black fish themselves, they think that they are exactly the same size and shape as those near the centre. Moreover, although from our outside Euclidean perspective they appear to get closer and closer to the bounding circle itself, from their own hyperbolic perspective that boundary always remains infinitely far away. Neither the bounding circle nor any of the 'Euclidean' space outside it has any existence for them. Their entire universe consists of what to us seems to lie strictly within the circle.

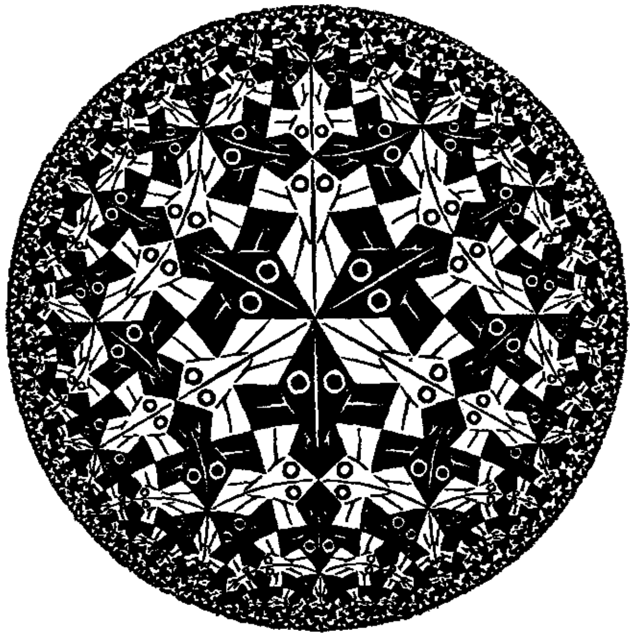

Fig. 2.11 M. C. Escher's woodcut *Circle Limit I*, illustrating the conformal representation of the hyperbolic plane.

In more mathematical terms, how is this picture of hyperbolic geometry constructed? Think of any circle in a Euclidean plane. The set of points lying in the interior of this circle is to represent the set of points in the entire hyperbolic plane. Straight lines, according to the hyperbolic geometry are to be represented as segments of Euclidean circles which meet the bounding circle *orthogonally*—which means at right angles. Now, it turns out that the hyperbolic notion of an *angle* between any two curves, at their point of intersection, is precisely the same as the Euclidean measure of angle between the two curves at the intersection point. A representation of this nature is called *conformal*. For this reason, the particular representation of hyperbolic geometry that Escher used is sometimes referred to as the *conformal model* of the hyperbolic plane. (It is also frequently referred to as the *Poincaré disc*. The dubious historical justification of this terminology will be discussed in §2.6.)

We are now in a position to see whether the angles of a triangle in hyperbolic geometry add up to π or not. A quick glance at Fig. 2.12 leads us to suspect that they do not and that they add up to something less. In fact, the sum of the angles of a triangle in hyperbolic geometry always falls short of π. We might regard that as a somewhat unpleasant feature of hyperbolic geometry, since we do not appear to get a 'neat' answer for the

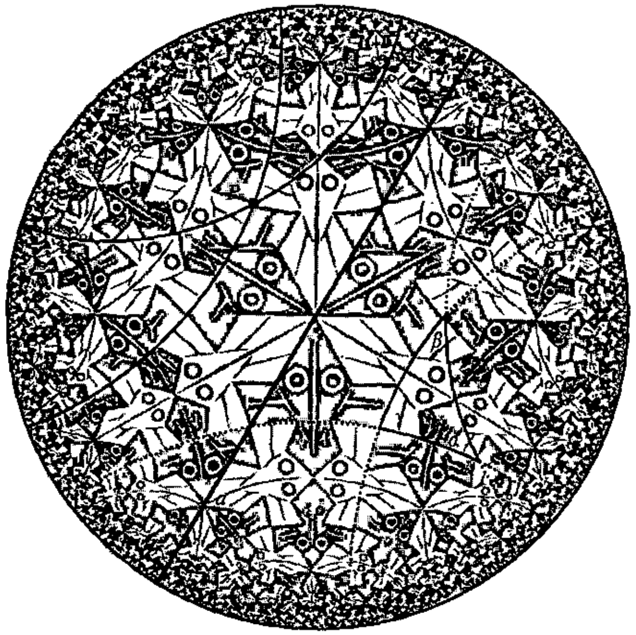

Fig. 2.12 The same Escher picture as Fig. 2.11, but with hyperbolic straight lines (Euclidean circles or lines meeting the bounding circle orthogonally) and a hyperbolic triangle illustrated. Hyperbolic angles agree with the Euclidean ones. The parallel postulate is evidently violated (lettering as in Fig. 2.8b) and the angles of a triangle sum to less than π.

sum of the angles of a triangle. However, there is actually something particularly elegant and remarkable about what does happen when we add up the angles of a hyperbolic triangle: the shortfall is always proportional to the area of the triangle. More explicitly, if the three angles of the triangle are α, β, and γ, then we have the formula (found by Johann Heinrich Lambert 1728–1777)

$$\pi - (\alpha + \beta + \gamma) = C\Delta,$$

where Δ is the area of the triangle and C is some constant. This constant depends on the 'units' that are chosen in which lengths and areas are to be measured. We can always scale things so that $C = 1$. It is, indeed, a remarkable fact that the area of a triangle can be so simply expressed in hyperbolic geometry. In Euclidean geometry, there is no way to express the area of a triangle simply in terms of its angles, and the expression for the area of a triangle in terms of its side-lengths is considerably more complicated.

In fact, I have not quite finished my description of hyperbolic geometry in terms of this conformal representation, since I have not yet described how the hyperbolic *distance* between two points is to be defined (and it would be appropriate to know what 'distance' is before we can really talk about areas). Let me give you an expression for the hyperbolic distance between two points A and B inside the circle. This is

$$\log \frac{\mathrm{QA} \cdot \mathrm{PB}}{\mathrm{QB} \cdot \mathrm{PA}},$$

where P and Q are the points where the Euclidean circle (i.e. hyperbolic straight line) through A and B orthogonal to the bounding circle *meets* this bounding circle and where 'QA', etc., refer to Euclidean distances (see Fig. 2.13). If you want to include the C of Lambert's area formula (with $C \neq 1$), just multiply the above distance expression by $C^{-1/2}$ (the reciprocal of the square root of C)^4.[2.2] For reasons that I hope may become clearer later, I shall refer to the quantity $C^{-1/2}$ as the *pseudo-radius* of the geometry.

If mathematical expressions like the above 'log' formula seem daunting, please do not worry. I am only providing it for those who like to see things explicitly. In any case, I am not going to explain why the expression works (e.g. why the shortest hyperbolic distance between two points, defined in this way, is actually measured along a hyperbolic straight line, or why the distances along a hyperbolic straight line 'add up' appropriately).[2.3] Also, I apologize for the 'log' (logarithm), but that is the way things are. In fact, this is a *natural logarithm* ('log to the base e') and I shall be having a good deal to say about it in §§5.2,3. We shall find that logarithms are really very beautiful and mysterious entities (as is the number e), as well as being important in many different contexts.

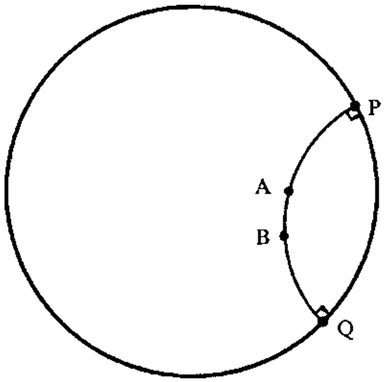

Fig. 2.13 In the conformal representation, the hyperbolic distance between A and B is log {QA.PB/QB.PA} where QA, etc. are Euclidean distances, P and Q being where the Euclidean circle through A and B, orthogonal to the bounding circle (hyperbolic line), meets this circle.

Hyperbolic geometry, with this definition of distance, turns out to have all the properties of Euclidean geometry apart from those which need the parallel postulate. We can construct triangles and other plane figures of different shapes and sizes, and we can move them around 'rigidly' (keeping their hyperbolic shapes and sizes from changing) with as much freedom as we can in Euclidean geometry, so that a natural notion of when two shapes are 'congruent' arises, just as in Euclidean geometry, where 'congruent' means 'can be moved around rigidly until they come into coincidence'. All the white fish in Escher's woodcut are indeed congruent to each other, according to this hyperbolic geometry, and so also are all the black fish.

2.5 Other representations of hyperbolic geometry

Of course, the white fish do not all look the same shape and size, but that is because we are viewing them from a Euclidean rather than a hyperbolic perspective. Escher's picture merely makes use of one particular Euclidean *representation* of hyperbolic geometry. Hyperbolic geometry itself is a more abstract thing which does not depend upon any particular Euclidean representation. However, such representations are indeed very helpful to us in that they provide ways of visualizing hyperbolic geometry by referring it to something that is more familiar and seemingly more 'concrete' to us, namely Euclidean geometry. Moreover, such representations make it clear that hyperbolic geometry is a consistent structure and that, consequently, the parallel postulate cannot be proved from the other laws of Euclidean geometry.

There are indeed other representations of hyperbolic geometry in terms of Euclidean geometry, which are distinct from the conformal one that Escher employed. One of these is that known as the *projective* model. Here, the entire hyperbolic plane is again depicted as the interior of a circle in a Euclidean plane, but the hyperbolic straight lines are now represented as straight Euclidean lines (rather than as circular arcs). There is, however, a price to pay for this apparent simplification, because the hyperbolic angles are now not the same as the Euclidean angles, and many people would regard this price as too high. For those readers who are interested, the hyperbolic distance between two points A and B in this representation is given by the expression (see Fig. 2.14)

$$\frac{1}{2} \log \frac{\mathrm{RA} \cdot \mathrm{SB}}{\mathrm{RB} \cdot \mathrm{SA}}$$

[2.2] Can you see a simple reason why?

[2.3] See if you can prove that, according to this formula, if A, B, and C are three successive points on a hyperbolic straight line, then the hyperbolic distances 'AB', etc. satisfy 'AB' + 'BC' = 'AC'. You may assume the general property of logarithms, $\log(ab) = \log a + \log b$ as described in §§5.2, 3.

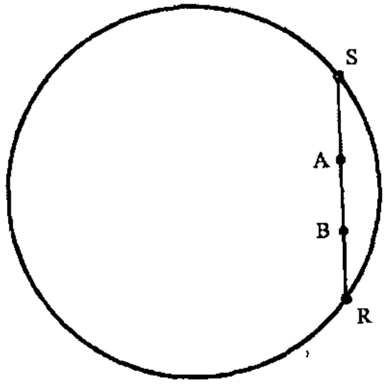

Fig. 2.14 In the projective representation, the formula for hyperbolic distance is now $\frac{1}{2}\log\{RA.SB/RB.SA\}$, where R and S are the intersections of the Euclidean (i.e. hyperbolic) straight line AB with the bounding circle.

(taking $C = 1$, this being almost the same as the expression we had before, for the conformal representation), where R and S are the intersections of the extended straight line AB with the bounding circle. This representation of hyperbolic geometry, can be obtained from the conformal one by means of an expansion radially out from the centre by an amount given by

$$\frac{2R^2}{R^2 + r_c^2},$$

where R is the radius of the bounding circle and r_c is the Euclidean distance out from the centre of the bounding circle of a point in the conformal representation (see Fig. 2.15).[2.4] In Fig. 2.16, Escher's picture of Fig. 2.11 has been transformed from the conformal to the projective model using this formula. (Despite lost detail, Escher's precise artistry is still evident.)

There is a more directly geometrical way of relating the conformal and projective representations, via yet another clever representation of this same geometry. All three of these representations are due to the ingenious Italian geometer Eugenio Beltrami (1835–1900). Consider a sphere S, whose equator coincides with the bounding circle of the projective representation of hyperbolic geometry given above. We are now going to find a representation of hyperbolic geometry on the *northern hemisphere* S^+ of S, which I shall call the *hemispheric* representation. See Fig. 2.17. To pass from the projective representation in the plane (considered as horizontal) to the new one on the sphere, we simply project vertically upwards (Fig. 2.17a). The straight lines in the plane, representing hyperbolic straight lines, are represented on S^+ by semicircles meeting the equator orthogonally. Now, to get from the representation on S^+ to the conformal representation on the plane, we project from the *south pole* (Fig. 2.17b). This is what is called *stereographic projection*, and it will play important roles later on in this book (see §8.3, §18.4, §22.9, §33.6). Two important properties of stereographic projection that we shall come to in §8.3 are that it is *conformal*, so that it preserves angles, and that it sends circles on the sphere to circles (or, exceptionally, to straight lines) on the plane.[2.5],[2.6]

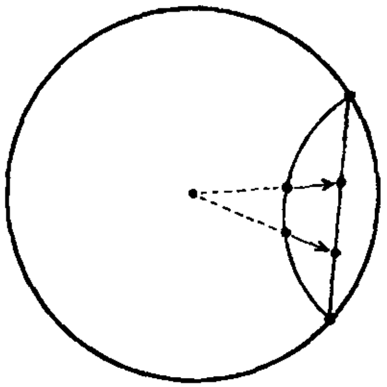

Fig. 2.15 To get from the conformal to the projective representation, expand out from the centre by a factor $2R^2/(R^2 + r_c^2)$, where R is the radius of the bounding circle and r_c is the Euclidean distance out of the point in the conformal representation.

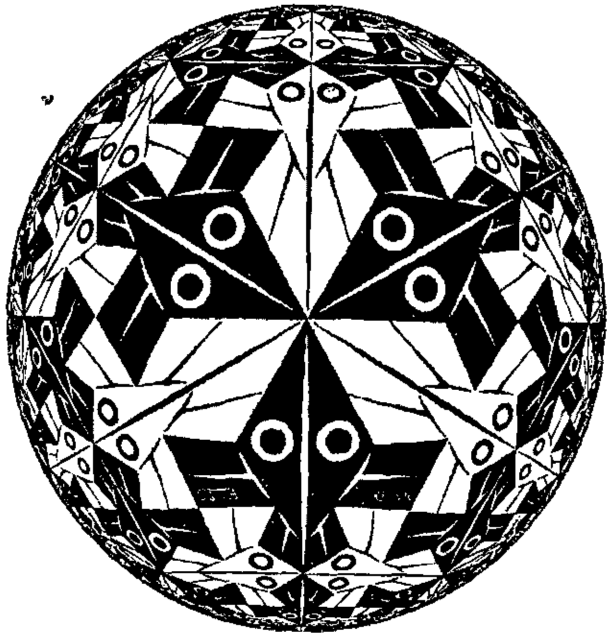

Fig. 2.16 Escher's picture of Fig. 2.11 transformed from the conformal to the projective representation.

[2.4] Show this. (*Hint*: You can use Beltrami's geometry, as illustrated in Fig. 2.17, if you wish.)

[2.5] Assuming these two stated properties of stereographic projection, the conformal representation of hyperbolic geometry being as stated in §2.4, show that Beltami's hemispheric representation is conformal, with hyperbolic 'straight lines' as vertical semicircles.

[2.6] Can you see how to prove these two properties? (*Hint*: Show, in the case of circles, that the cone of projection is intersected by two planes of exactly opposite tilt.)

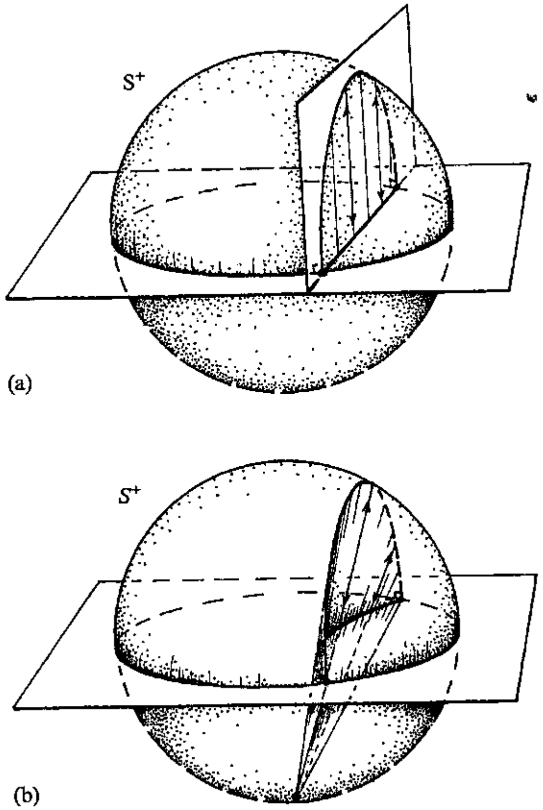

Fig. 2.17 Beltrami's geometry, relating three of his representations of hyperbolic geometry. (a) The hemispheric representation (conformal on the *northern hemisphere* S^+) projects vertically to the projective representation on the equatorial disc. (b) The hemispheric representation projects stereographically, from the *south pole* to the conformal representation on the equatorial disc.

The existence of various different models of hyperbolic geometry, expressed in terms of Euclidean space, serves to emphasize the fact that these are, indeed, merely 'Euclidean models' of hyperbolic geometry and are not to be taken as telling us what hyperbolic geometry actually *is*. Hyperbolic geometry has its own 'Platonic existence', just as does Euclidean geometry (see §1.3 and the Preface). No one of the models is to be taken as the 'correct' picturing of hyperbolic geometry at the expense of the others. The representations of it that we have been considering are very valuable as aids to our understanding, but only because the Euclidean framework is the one which we are more used to. For a sentient creature brought up with a direct experience of hyperbolic (rather than Euclidean) geometry, a

model of Euclidean geometry in hyperbolic terms might seem the more natural way around. In §18.4, we shall encounter yet another model of hyperbolic geometry, this time in terms of the Minkowskian geometry of special relativity.

To end this section, let us return to the question of the existence of squares in hyperbolic geometry. Although squares whose angles are right angles do not exist in hyperbolic geometry, there are 'squares' of a more general type, whose angles are less than right angles. The easiest way to construct a square of this kind is to draw two straight lines intersecting at right angles at a point O. Our 'square' is now the quadrilateral whose four vertices are the intersections A, B, C, D (taken cyclicly) of these two lines with some circle with centre O. See Fig. 2.18. Because of the symmetry of the figure, the four sides of the resulting quadrilateral ABCD are all equal and all of its four angles must also be equal. But are these angles right angles? Not in hyperbolic geometry. In fact they can be any (positive) angle we like which is less than a right angle, but not equal to a right angle. The bigger the (hyperbolic) square (i.e. the larger the circle, in the above construction), the smaller will be its angles. In Fig. 2.19a, I have depicted a lattice of hyperbolic squares, using the conformal model, where there are five squares at each vertex point (instead of the Euclidean four), so the angle is $\frac{2}{5}\pi$, or 72°. In Fig. 2.19b, I have depicted the same lattice using the projective model. It will be seen that this does not allow the modifications that would be needed for the two-square lattice of Fig. 2.2.[2.7]

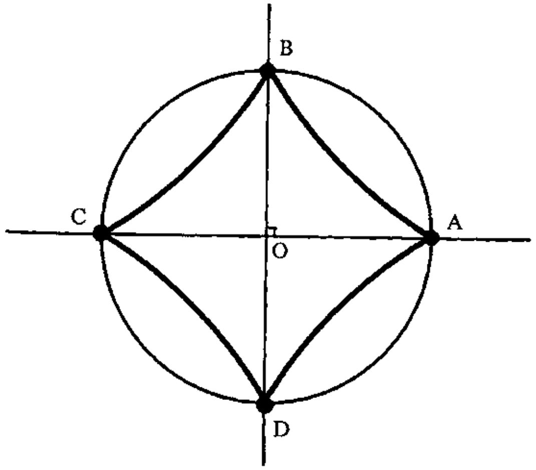

Fig. 2.18 A hyperbolic 'square' is a hyperbolic quadrilateral, whose vertices are the intersections A, B, C, D (taken cyclically) of two perpendicular hyperbolic straight lines through some point O with some circle centred at O. Because of symmetry, the four sides of ABCD as well as all the four angles are equal. These angles are not right angles, but can be equal to any given positive angle less than $\frac{1}{2}\pi$.

[2.7] See if you can do something similar, but with hyperbolic regular pentagons and squares.

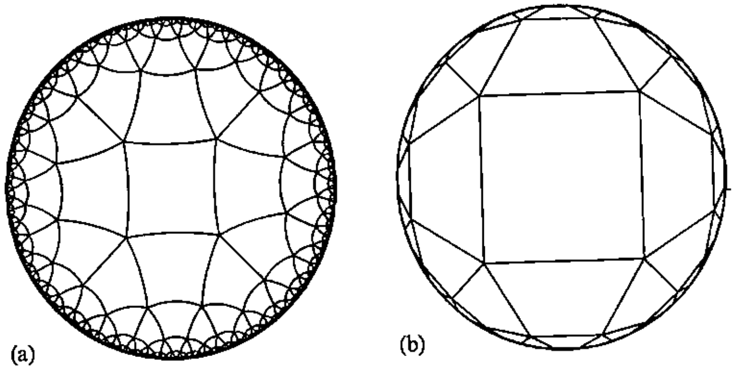

Fig. 2.19 A lattice of squares, in hyperbolic space, in which five squares meet at each vertex, so the angles of the square are $\frac{2\pi}{5}$, or 72°. (a) Conformal representation. (b) Projective representation.

2.6 Historical aspects of hyperbolic geometry

A few historical comments concerning the discovery of hyperbolic geometry are appropriate here. For centuries following the publication of Euclid's elements, in about 300 BC, various mathematicians attempted to prove the fifth postulate from the other axioms and postulates. These efforts reached their greatest heights with the heroic work by the Jesuit Girolamo Saccheri in 1733. It would seem that Saccheri himself must ultimately have thought his life's work a failure, constituting merely an unfulfilled attempt to *prove* the parallel postulate by showing that the hypothesis that the angle sum of every triangle is less than two right angles led to a contradiction. Unable to do this logically after momentous struggles, he concluded, rather weakly:

> The hypothesis of acute angle is absolutely false; because repugnant to the nature of the straight line.[5]

The hypothesis of 'acute angle' asserts that the lines a and b of Fig. 2.8. sometimes do not meet. It is, in fact, viable and actually yields hyperbolic geometry!

How did it come about that Saccheri effectively discovered something that he was trying to show was impossible? Saccheri's proposal for proving Euclid's fifth postulate was to make the assumption that the fifth postulate was false and then derive a contradiction from this assumption. In this way he proposed to make use of one of the most time-honoured and fruitful principles ever to be put forward in mathematics—very possibly first introduced by the Pythagoreans—called *proof by contradiction* (or *reductio ad absurdum*, to give it its Latin name). According to this procedure, in order to prove that some assertion is true, one first makes the supposition that the assertion in question is *false*, and one then argues from this that some contradiction ensues. Having found such a contradiction, one deduces that the assertion must be true after all.[6] Proof by contradiction provides a very powerful method of reasoning in mathematics, frequently applied today. A quotation from the distinguished mathematician G. H. Hardy is apposite here:

> *Reductio ad absurdum*, which Euclid loved so much, is one of a mathematician's finest weapons. It is a far finer gambit than any chess gambit: a chess player may offer the sacrifice of a pawn or even a piece, but a mathematician offers *the game*.[7]

We shall be seeing other uses of this important principle later (see §3.1 and §§16.4,6).

However, Saccheri failed in his attempt to find a contradiction. He was therefore not able to obtain a proof of the fifth postulate. But in striving for it he, in effect, found something far greater: a new geometry, different from that of Euclid—the geometry, discussed in §§2.4,5, that we now call *hyperbolic geometry*. From the assumption that Euclid's fifth postulate was false, he derived, instead of an actual contradiction, a host of strange-looking, barely believable, but interesting theorems. However, strange as these results appeared to be, none of them was actually a contradiction. As we now know, there was no chance that Saccheri would find a genuine contradiction in this way, for the reason that hyperbolic geometry *does* actually exist, in the mathematical sense that there is such a consistent structure. In the terminology of §1.3, hyperbolic geometry inhabits Plato's world of mathematical forms. (The issue of hyperbolic geometry's *physical* reality will be touched upon in §2.7 and §28.10.)

A little after Saccheri, the highly insightful mathematician Johann Heinrich Lambert (1728–1777) also derived a host of fascinating geometrical results from the assumption that Euclid's fifth postulate is false, including the beautiful result mentioned in §2.4 that gives the area of a hyperbolic triangle in terms of the sum of its angles. It appears that Lambert may well have formed the opinion, at least at some stage of his life, that a consistent geometry perhaps could be obtained from the denial of Euclid's fifth postulate. Lambert's tentative reason seems to have been that he could contemplate the theoretical possibility of the geometry on a 'sphere of imaginary radius', i.e. one for which the 'squared radius' is negative. Lambert's formula $\pi - (\alpha + \beta + \gamma) = C\Delta$ gives the area, Δ, of a hyperbolic triangle, where α, β, and γ are the angles of the triangle and where C is a constant ($-C$ being what we would now call the 'Gaussian curvature' of the hyperbolic plane). This formula looks basically the same as a previously known one due, in 1603, to Thomas Hariot (1560–1621), $\varDelta = R^2(\alpha + \beta + \gamma - \pi)$, for the area $\varDelta$ of a *spherical triangle*, drawn with great circle arcs[8] on a sphere of radius R (see Fig. 2.20).[2.8] To retrieve Lambert's formula, we have to put

$$C = -\frac{1}{R^2}.$$

But, in order to give the *positive* value of C, as would be needed for hyperbolic geometry, we require the sphere's radius to be 'imaginary' (i.e. to be the square root of a negative number). Note that the radius R is given by the imaginary quantity $(-C)^{-1/2}$. This explains the term 'pseudo-radius', introduced in §2.4, for the real quantity $C^{-1/2}$. In fact Lambert's procedure is perfectly justified from our more modern perspectives (see Chapter 4 and §18.4, Fig. 18.9), and it indicates great insight on his part to have foreseen this.

It is, however, the conventional standpoint (somewhat unfair, in my opinion) to deny Lambert the honour of having first constructed non-Euclidean geometry, and to consider that (about half a century later) the first person to have come to a clear acceptance of a fully consistent geometry, distinct from that of Euclid, in which the parallel postulate is false, was the great mathematician Carl Friedrich Gauss. Being an exceptionally cautious man, and being fearful of the controversy that such a revelation might cause, Gauss did not publish his findings, and kept them to himself.[9] Some 30 years after Gauss had begun working on it, hyperbolic

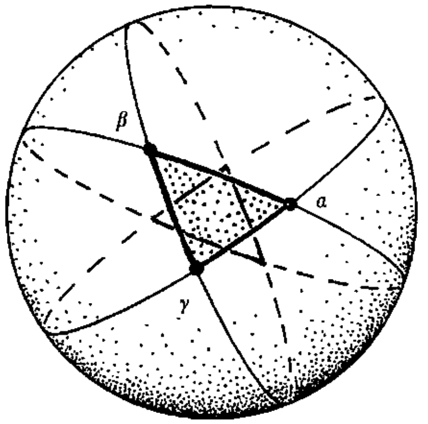

Fig. 2.20 Hariot's formula for the area of a *spherical triangle*, on a sphere of radius R, with angles α, β, γ, is $\Delta = R^2(\alpha + \beta + \gamma - \pi)$. Lambert's formula, for a hyperbolic triangle, has $C = -1/R^2$.

[2.8] Try to prove this spherical triangle formula, basically using only symmetry arguments and the fact that the total area of the sphere is $4\pi R^2$. *Hint*: Start with finding the area of a segment of a sphere bounded by two great circle arcs connecting a pair of antipodal points on the sphere; then cut and paste and use symmetry arguments. Keep Fig. 2.20 in mind.

geometry was independently rediscovered by various others, including the Hungarian János Bolyai (by 1829) and, most particularly, the Russian geometer Nicolai Ivanovich Lobachevsky in about 1826 (whence hyperbolic geometry is frequently called *Lobachevskian* geometry).

The specific projective and conformal realizations of hyperbolic geometry that I have described above were both found by Eugenio Beltrami, and published in 1868, together with some other elegant representations including the hemispherical one mentioned in §2.5. The conformal representation is, however, commonly referred to as the 'Poincaré model', because Poincaré's rediscovery of this representation in 1882 is better known than the original work of Beltrami (largely because of the important use that Poincaré made of this model).[10] Likewise, poor old Beltrami's projective representation is sometimes called the 'Klein representation'. It is not uncommon in mathematics that the name normally attached to a mathematical concept is not that of the original discoverer. At least, in this case, Poincaré did *re*discover the conformal representation (as did Klein the projective one in 1871). There are other instances in mathematics where the mathematician(s) whose name(s) are attached to a result did not even know of the result in question![11]

The representation of hyperbolic geometry that Beltrami is best known for is yet another one, which he found also in 1868. This represents the geometry on a certain surface known as a *pseudo-sphere* (see Fig. 2.21). This surface is obtained by rotating a *tractrix*, a curve first investigated by Isaac Newton in 1676, about its 'asymptote'. The asymptote is a straight line which the curve approaches, becoming asymptotically tangent to it as the curve recedes to infinity. Here, we are to imagine the asymptote to be drawn on a horizontal plane of rough texture. We are to think of a light, straight, stiff rod, at one end P of which is attached a heavy point-like weight, and the other end R moves along the asymptote. The point P then traces out a tractrix. Ferdinand Minding found, in 1839, that the pseudo-sphere has a constant

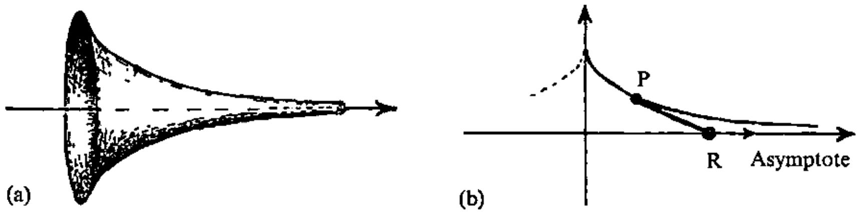

Fig. 2.21 (a) A *pseudo-sphere*. This is obtained by rotating, about its asymptote (b) a *tractrix*. To construct a tractrix, imagine its plane to be horizontal, over which is dragged a light, frictionless straight, stiff rod. One end of the rod is a point-like weight P with friction, and the other end R moves along the (straight) asymptote.

negative intrinsic geometry, and Beltrami used this fact to construct the first model of hyperbolic geometry. Beltrami's pseudo-sphere model seems to be the one that persuaded mathematicians of the consistency of plane hyperbolic geometry, since the measure of hyperbolic distance agrees with the Euclidean distance along the surface. However, it is a somewhat awkward model, because it represents hyperbolic geometry only locally, rather than presenting the entire geometry all at once, as do Beltrami's other models.

2.7 Relation to physical space

Hyperbolic geometry also works perfectly well in higher dimensions. Moreover, there are higher-dimensional versions of both the conformal and projective models. For three-dimensional hyperbolic geometry, instead of a bounding circle, we have a bounding sphere. The entire infinite three-dimensional hyperbolic geometry is represented by the interior of this finite Euclidean sphere. The rest is basically just as we had it before. In the conformal model, straight lines in this three-dimensional hyperbolic geometry are represented as Euclidean circles which meet the bounding sphere orthogonally; angles are given by the Euclidean measures, and distances are given by the same formula as in the two-dimensional case. In the projective model, the hyperbolic straight lines are Euclidean straight lines, and distances are again given by the same formula as in the two-dimensional case.

What about our actual universe on cosmological scales? Do we expect that its spatial geometry is Euclidean, or might it accord more closely with some other geometry, such as the remarkable hyperbolic geometry (but in three dimensions) that we have been examining in §§2.4–6. This is indeed a serious question. We know from Einstein's general relativity (which we shall come to in §17.9 and §19.6) that Euclid's geometry is only an (extraordinarily accurate) approximation to the actual geometry of physical space. This physical geometry is not even exactly uniform, having small ripples of irregularity owing to the presence of matter density. Yet, strikingly, according to the best observational evidence available to cosmologists today, these ripples appear to average out, on cosmological scales, to a remarkably exact degree (see §27.13 and §§28.4–10), and the spatial geometry of the actual universe seems to accord with a uniform (homogeneous and isotropic—see §27.11) geometry extraordinarily closely. Euclid's first four postulates, at least, would seem to have stood the test of time impressively well.

A remark of clarification is needed here. Basically, there are three types of geometry that would satisfy the conditions of homogeneity (every point the same) and isotropy (every direction the same), referred to as Euclidean, hyperbolic, and elliptic. Euclidean geometry is familiar to us (and has been for some 23 centuries). Hyperbolic geometry has been our main concern in this chapter. But what is elliptic geometry? Essentially, elliptic plane geometry is that satisfied by figures drawn on the surface of a sphere. It figured in the discussion of Lambert's approach to hyperbolic geometry in §2.6. See Fig. 2.22a,b,c,

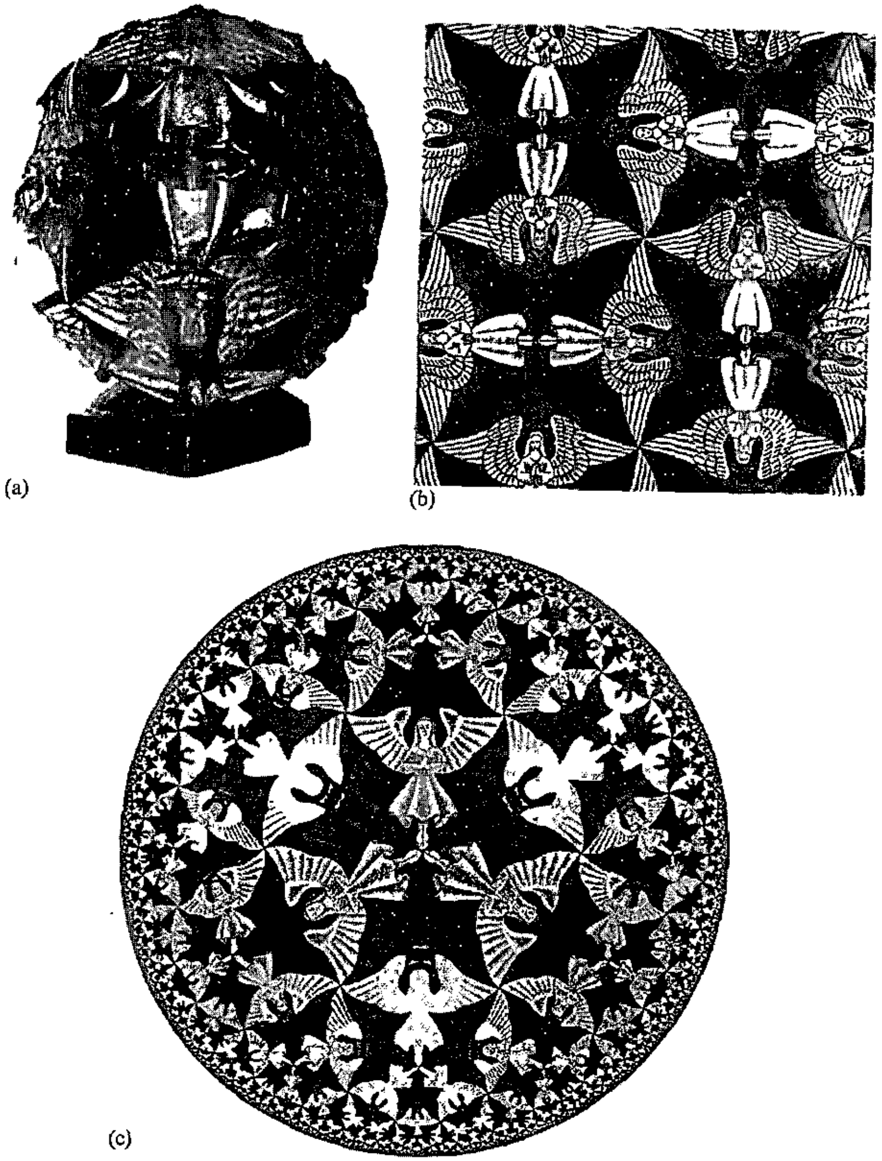

Fig. 2.22 The three basic kinds of uniform plane geometry, as illustrated by Escher using tessellations of angels and devils. (a) Elliptic case (positive curvature), (b) Euclidean case (zero curvature), and (c) Hyperbolic case (negative curvature)—in the conformal representation (Escher's *Circle Limit IV*, to be compared with Fig. 2.11).

for Escher's rendering of the elliptic, Euclidean, and hyperbolic cases, respectively, using a similar tessellation of angels and devils in all three cases, the third one providing an interesting alternative to Fig. 2.11. (There is also a three-dimensional version of elliptic geometry, and there are versions in which diametrically opposite points of the sphere are considered to represent the same point. These issues will be discussed a little more fully in §27.11.) However, the elliptic case could be said to violate Euclid's second and third postulates (as well as the first). For it is a geometry that is finite in extent (and for which more than one line segment joins a pair of points).

What, then, is the observational status of the large-scale spatial geometry of the universe? It is only fair to say that we do not yet know, although there have been recent widely publicized claims that Euclid was right all along, and his fifth postulate holds true also, so the averaged spatial geometry is indeed what we call 'Euclidean'.[12] On the other hand, there is also evidence (some of it coming from the same experiments) that seems to point fairly firmly to a *hyperbolic* overall geometry for the spatial universe.[13] Moreover, some theoreticians have long argued for the elliptic case, and this is certainly not ruled out by that same evidence that is argued to support the Euclidean case (see the later parts of §34.4). As the reader will perceive, the issue is still fraught with controversy and, as might be expected, often heated argument. In later chapters in this book, I shall try to present a good many of the considerations that have been put forward in this connection (and I do not attempt to hide my own opinion in favour of the hyperbolic case, while trying to be as fair to the others as I can).

Fortunately for those, such as myself, who are attracted to the beauties of hyperbolic geometry, and also to the magnificence of modern physics, there is another role for this superb geometry that is undisputedly fundamental to our modern understanding of the physical universe. For the space of *velocities*, according to modern relativity theory, is certainly a three-dimensional hyperbolic geometry (see §18.4), rather than the Euclidean one that would hold in the older Newtonian theory. This helps us to understand some of the puzzles of relativity. For example, imagine a projectile hurled forward, with near light speed, from a vehicle that also moves forwards with comparable speed past a building. Yet, relative to that building, the projectile can never exceed light speed. Though this seems impossible, we shall see in §18.4 that it finds a direct explanation in terms of hyperbolic geometry. But these fascinating matters must wait until later chapters.

What about the Pythagorean theorem, which we have seen to fail in hyperbolic geometry? Must we abandon this greatest of the specific Pythagorean gifts to posterity? Not at all, for hyperbolic geometry—and, indeed, all the 'Riemannian' geometries that generalize hyperbolic geometry in an irregularly curved way (forming the essential framework for Einstein's general theory of relativity; see §13.8, §14.7, §18.1, and §19.6)—depends vitally upon the Pythagorean theorem holding in the limit of small distances. Moreover, its enormous influence permeates other vast areas of mathematics and physics (e.g. the 'unitary' metric structure of quantum mechanics, see §22.3). Despite the fact that this theorem is, in a sense, superseded for 'large' distances, it remains central to the small-scale structure of geometry, finding a range of application that enormously exceeds that for which it was originally put forward.

Notes

Section 2.1

2.1. It is historically very unclear who actually first proved what we now refer to as the 'Pythagorean theorem', see Note 1.1. The ancient Egyptians and Babylonians seem to have known at least many instances of this theorem. The true role played by Pythagoras or his followers is largely surmise.

Section 2.2

2.2. Even with this amount of care, however, various hidden assumptions remained in Euclid's work, mainly to do with what we would now call 'topological' issues that would have seemed to be 'intuitively obvious' to Euclid and his contemporaries. These unmentioned assumptions were pointed out only centuries later, particularly by Hilbert at the end of the 19th century. I shall ignore these in what follows.

2.3. See e.g. Thomas (1939). Compare also Schutz (1997), who gives a nice axiomatic account of Minkowski's 4-dimensional spacetime geometry (§17.8, §18.1).

Section 2.4

2.4. The 'exponent' notation, such as $C^{-1/2}$, is frequently used in this book. As already referred to in Note 1.1, a^5 means $a \times a \times a \times a \times a$; correspondingly, for a positive integer n, the product of a with itself a total of n times is written a^n. This notation extends to negative exponents, so that a^{-1} is the reciprocal $1/a$ of a, and a^{-n} is the reciprocal $1/a^n$ of a^n, or equivalently $(a^{-1})^n$. In accordance with the more general discussion of §5.2, $a^{1/n}$, for a positive number a, is the 'nth root of a', which is the (positive) number satisfying $(a^{1/n})^n = a$ (see Note 1.2). Moreover, $a^{m/n}$ is the mth power of $a^{1/n}$.

Section 2.6

2.5. Saccheri (1733), Prop. XXXIII.

2.6. There is a standpoint known as *intuitionism*, which is held to by a (rather small) minority of mathematicians, in which the principle of 'proof by contradiction' is not accepted. The objection is that this principle can be *non-constructive* in that it sometimes leads to an assertion of the existence of some mathematical entity, without any actual construction for it having been provided. This has some relevance to the issues discussed in §16.6. See Heyting (1956).

2.7. Hardy (1940), p. 34.

2.8. Great circle arcs are the 'shortest' curves (geodesics) on the surface of a sphere; they lie on planes through the sphere's centre.

2.9. It is a matter of some dispute whether Gauss, who was professionally concerned with matters of geodesy, might actually have tried to ascertain whether there are measurable deviations from Euclidean geometry in physical space. Owing to his well-known reticence in matters of non-Euclidean geometry, it is unlikely that he would let it be known if he were in fact trying to do this, particularly since (as we now know) he would be bound to fail, owing to the smallness of the effect, according to modern theory. The present consensus seems to be that he was 'just doing geodesy', being concerned with the curvature of the Earth, and not of space. But I find it a little hard to believe that he would not also have been on the lookout for any significant discrepancy with Euclidean geometry; see Fauvel and Gray (1987), Gray (1979).

2.10. The so-called 'Poincaré half-plane' representation (with metric form $(dx^2 + dy^2)/y^2$; see §14.7) is also due to Beltrami; see Beltrami (1868). The constant negative curvature of the 'Poincaré metric' $4(dx^2 + dy^2)/(1 - x^2 - y^2)^2$ of Figs. 2.11–13 was actually noted by Riemann (1854)!

2.11. This appears to have applied even to the great Gauss himself (who had, on the other hand, very frequently anticipated other mathematicians' work). There is an important topological mathematical theorem now referred to as the 'Gauss–Bonnet theorem', which can be elegantly proved by use of the so-called 'Gauss map', but the theorem itself appears actually to be due to Blaschke and the elegant proof procedure just referred to was found by Olinde Rodrigues. It appears that neither the result nor the proof procedure were even known to Gauss or to Bonnet. There is a more elemental 'Gauss–Bonnet' theorem, correctly cited in several texts, see Willmore (1959), also Rindler (2001).

Section 2.7

2.12. The main evidence for the overall structure of the universe, as a whole comes from a detailed analysis of the *cosmic microwave background radiation* (CMB) that will be discussed in §§27.7,10,11,13, §§28.5,10, and §30.14. A basic reference is de Bernardis *et al.* (2000); for more accurate, more recent data, see Netterfield *et al.* (2001) (concerning BOOMERanG). See also Hanany *et al.* (2000) (concerning MAXIMA), Halverson *et al.* (2001) (concerning DASI), and Bennett *et al.* (2003) (concerning WMAP).

2.13. See Gurzadyan and Torres (1997) and Gurzadyan and Kocharyan (1994) for the theoretical underpinnings, and Gurzadyan and Kocharyan (1992) (for COBE data) and Gurzadyan *et al.* (2002, 2003) (for BOOMERanG data and (2004) for WMAP data) for the corresponding analysis of the actual CMB data.

3 Kinds of number in the physical world

3.1 A Pythagorean catastrophe?

Let us now return to the issue of proof by contradiction, the very principle that Saccheri tried hard to use in his attempted proof of Euclid's fifth postulate. There are many instances in classical mathematics where the principle *has* been successfully applied. One of the most famous of these dates back to the Pythagoreans, and it settled a mathematical issue in a way which greatly troubled them. This was the following. Can one find a rational number (i.e. a fraction) whose square is precisely the number 2? The answer turns out to be no, and the mathematical assertion that I shall demonstrate shortly is, indeed, that there is no such rational number.

Why were the Pythagoreans so troubled by this discovery? Recall that a fraction—that is, a rational number—is something that can be expressed as the ratio *a/b* of two integers (or whole numbers) *a* and *b*, with *b* non-zero. (See the Preface for a discussion of the definition of a fraction.) The Pythagoreans had originally hoped that all their geometry could be expressed in terms of lengths that could be measured in terms of rational numbers. Rational numbers are rather simple quantities, being describable and understood in simple finite terms; yet they can be used to specify distances that are as small as we please or as large as we please. If all geometry could be done with rationals, then this would make things relatively simple and easily comprehensible. The notion of an 'irrational' number, on the other hand, requires infinite processes, and this had presented considerable difficulties for the ancients (and with good reason). Why is there a difficulty in the fact that there is no rational number that squares to 2? This comes from the Pythagorean theorem itself. If, in Euclidean geometry, we have a square whose side length is unity, then its diagonal length is a number whose square is $1^2 + 1^2 = 2$ (see Fig. 3.1). It would indeed be catastrophic for geometry if there were no actual number that could describe the length of the diagonal of a square. The Pythagoreans tried, at first, to make do with a notion of 'actual number' that could be described simply in terms of ratios of whole numbers. Let us see why this will not work.

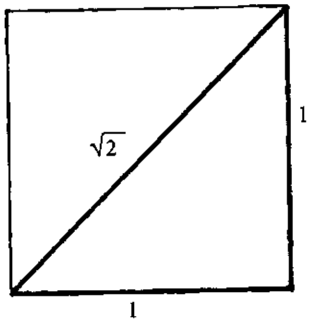

Fig. 3.1 A square of unit side-length has diagonal $\sqrt{2}$, by the Pythagorean theorem.

The issue is to see why the equation

$$\left(\frac{a}{b}\right)^2 = 2$$

has no solution for integers a and b, where we take these integers to be positive. We shall use proof by contradiction to prove that no such a and b can exist. We therefore try to suppose, on the contrary, that such an a and b *do* exist. Multiplying the above equation by b^2 on both sides, we find that it becomes

$$a^2 = 2b^2$$

and we clearly conclude[1] that $a^2 > b^2 > 0$. Now the right-hand side, $2b^2$, of the above equation is even, whence a must be even (not odd, since the square of any odd number is odd). Hence $a = 2c$, for some positive integer c. Substituting $2c$ for a in the above equation, and squaring it out, we obtain

$$4c^2 = 2b^2,$$

that is, dividing both sides by 2,

$$b^2 = 2c^2,$$

and we conclude $b^2 > c^2 > 0$. Now, this is precisely the same equation that we had displayed before, except that b now replaces a, and c replaces b. Note that the corresponding integers are now smaller than they were before. We can now repeat the argument again and again, obtaining an unending sequence of equations

$$a^2 = 2b^2,\ b^2 = 2c^2,\ c^2 = 2d^2,\ d^2 = 2e^2,\ \dots,$$

where

$$a^2 > b^2 > c^2 > d^2 > e^2 > \dots,$$

all of these integers being positive. But any decreasing sequence of positive integers must come to an end, contradicting the fact that this sequence is unending. This provides us with a contradiction to what has been supposed, namely that there is a rational number which squares to 2. It follows that there is no such rational number—as was required to prove.[2]

Certain points should be remarked upon in the above argument. In the first place, in accordance with the normal procedures of mathematical proof, certain properties of numbers have been appealed to in the argument that were taken as either 'obvious' or having been previously established. For example, we made use of the fact that the square of an odd number is always odd and, moreover, that if an integer is not odd then it is even. We also used the fundamental fact that every strictly decreasing sequence of positive integers must come to an end.

One reason that it can be important to identify the precise assumptions that go into a proof—even though some of these assumptions could be perfectly 'obvious' things—is that mathematicians are frequently interested in other kinds of entity than those with which the proof might be originally concerned. If these other entities satisfy the same assumptions, then the proof will still go through and the assertion that had been proved will be seen to have a greater generality than originally perceived, since it will apply to these other entities also. On the other hand, if some of the needed assumptions fail to hold for these alternative entities, then the assertion that may turn out to be false for these entities. (For example, it is important to realize that the parallel postulate was used in the proofs of the Pythagorean theorem given in §2.2, for the theorem is actually false for hyperbolic geometry.)

In the above argument, the original entities are integers and we are concerned with those numbers—the rational numbers—that are constructed as quotients of integers. With such numbers it is indeed the case that none of them squares to 2. But there are other kinds of number than merely integers and rationals. Indeed, the need for a square root of 2 forced the ancient Greeks, very much against their wills at the time, to proceed outside the confines of integers and rational numbers—the only kinds of number that they had previously been prepared to accept. The kind of number that they found themselves driven to was what we now call a 'real number': a number that we now express in terms of an unending decimal expansion (although such a representation was not available to the ancient Greeks). In fact, 2 does indeed have a real-number square root, namely (as we would now write it)

$$\sqrt{2} = 1.414\,213\,562\,373\,095\,048\,801\,688\,72\ldots.$$

We shall consider the *physical* status of such 'real' numbers more closely in the next section.

As a curiosity, we may ask why the above proof of the non-existence of a square root of 2 fails for real numbers (or for real-number ratios, which amounts to the same thing). What happens if we replace 'integer' by 'real number' throughout the argument? The basic difference is that it is not true that any strictly decreasing sequence of positive reals (or even of fractions) must come to an end, and the argument breaks down at that point.[3] (Consider the unending sequence $1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}, \ldots,$ for example.) One might worry what an 'odd' and 'even' real number would be in this context. In fact the argument encounters no difficulty at that stage because *all* real numbers would have to count as 'even', since for any real a there is always a real c such that $a = 2c$, division by 2 being always possible for reals.

3.2 The real-number system

Thus it was that the Greeks were forced into the realization that rational numbers are not enough, if the ideas of (Euclid's) geometry are to be properly developed. Nowadays, we do not worry unduly if a certain geometrical quantity cannot be measured simply in terms of rational numbers alone. This is because the notion of a 'real number' is very familiar to us. Although our pocket calculators express numbers in terms of only a finite number of digits, we readily accept that this is an approximation forced upon us by the fact that the calculator is a finite object. We are prepared to allow that the ideal (Platonic) mathematical number could certainly require that the decimal expansion continues indefinitely. This applies, of course, even to the decimal representation of most fractions, such as

$$\tfrac{1}{3} = 0.333\,333\,333\ldots,$$
$$\tfrac{29}{12} = 2.416\,666\,666\ldots,$$
$$\tfrac{9}{7} = 1.285\,714\,285\,714\,285,$$
$$\tfrac{237}{148} = 1.601\,351\,351\,35\ldots.$$

For a fraction, the decimal expanson is always *ultimately periodic*, which is to say that after a certain point the infinite sequence of digits consists of some finite sequence repeated indefinitely. In the above examples the repeated sequences are, respectively, 3, 6, 285714, and 135.

Decimal expansions were not available to the ancient Greeks, but they had their own ways of coming to terms with irrational numbers. In effect, what they adopted was a system of representing numbers in terms of what are now called *continued fractions*. There is no need to go into this in full detail here, but some brief comments are appropriate. A continued fraction[4] is a finite or infinite expression $a + (b + (c + (d + \cdots)^{-1})^{-1})^{-1}$, where $a, b, c, d, \ldots$ are positive integers:

$$a + \cfrac{1}{b + \cfrac{1}{c + \cfrac{1}{d + \cdots}}}$$

Any rational number larger than 1 can be written as a *terminating* such expression (where to avoid ambiguity we normally require the final integer to be greater than 1), e.g. $52/9 = 5 + (1 + (3 + (2)^{-1})^{-1})^{-1}$:

$$\frac{52}{9} = 5 + \cfrac{1}{1 + \cfrac{1}{3 + \cfrac{1}{2}}}$$

and, to represent a positive rational less than 1, we just allow the first integer in the expression to be zero. To express a real number, which is not rational, we simply[3.1] allow the continued-fraction expression to run on forever, some examples being[5]

$$\sqrt{2} = 1 + (2 + (2 + (2 + (2 + \cdots)^{-1})^{-1})^{-1})^{-1},$$

$$7 - \sqrt{3} = 5 + (3 + (1 + (2 + (1 + (2 + (1 + (2 + \cdots)^{-1})^{-1})^{-1})^{-1})^{-1})^{-1})^{-1},$$

$$\pi = 3 + (7 + (15 + (1 + (292 + (1 + (1 + (1 + (2 + \cdots)^{-1})^{-1})^{-1})^{-1})^{-1})^{-1})^{-1})^{-1}.$$

In the first two of these infinite examples, the sequences of natural numbers that appear—namely 1, 2, 2, 2, 2, ... in the first case and 5, 3, 1, 2, 1, 2, 1, 2, ... in the second—have the property that they are ultimately periodic (the 2 repeating indefinitely in the first case and the sequence 1, 2 repeating indefinitely in the second).[3.2] Recall that, as

[3.1] Experiment with your pocket calculator (assuming you have '$\surd$' and 'x^{-1}' keys) to obtain these expansions to the accuracy available. Take $\pi = 3.141\,592\,653\,589\,793\ldots$ (*Hint*: Keep taking note of the integer part of each number, subtracting it off, and then forming the reciprocal of the remainder.)

[3.2] Assuming this eventual periodicity of these two continued-fraction expressions, show that the numbers they represent must be the quantities on the left. (*Hint*: Find a quadratic equation that must be satisfied by this quantity, and refer to Note 3.6.)

already noted above, in the familiar decimal notation, it is the *rational* numbers that have (finite or) ultimately periodic expressions. We may regard it as a strength of the Greek 'continued-fraction' representation, on the other hand, that the rational numbers now always have a finite description. A natural question to ask, in this context, is: which numbers have an *ultimately periodic* continued-fraction representation? It is a remarkable theorem, first proved, to our knowledge, by the great 18th-century mathematician Joseph C. Lagrange (whose most important other ideas we shall encounter later, particularly in Chapter 20) that the numbers whose representation in terms of continued fractions are ultimately periodic are what are called *quadratic irrationals*.[6]

What is a quadratic irrational and what is its importance for Greek geometry? It is a number that can be written in the form

$$a + \sqrt{b},$$

where a and b are fractions, and where b is not a perfect square. Such numbers are important in Euclidean geometry because they are the most immediate irrational numbers that are encountered in ruler-and-compass constructions. (Recall the Pythagorean theorem, which in §3.1 first led us to consider the problem of $\sqrt{2}$, and other simple constructions of Euclidean lengths directly lead us to other numbers of the above form.)

Particular examples of quadratic irrationals are those cases where $a = 0$ and b is a (non-square) natural number, or rational greater than 1, e.g.

$$\sqrt{2}, \sqrt{3}, \sqrt{5}, \sqrt{6}, \sqrt{7}, \sqrt{8}, \sqrt{10}, \sqrt{11}, \ldots.$$

The continued-fraction representation of such a number is particularly striking. The sequence of natural numbers that defines it as a continued fraction has a curious characteristic property. It starts with some number A, then it is immediately followed by a 'palindromic' sequence (i.e. one which reads the same backwards), $B, C, D, \ldots, D, C, B$, followed by $2A$, after which the sequence $B, C, D, \ldots, D, C, B, 2A$ repeats itself indefinitely. The number $\sqrt{14}$ is a good example, for which the sequence is

$$3, 1, 2, 1, 6, 1, 2, 1, 6, 1, 2, 1, 6, 1, 2, 1, 6, \ldots.$$

Here $A = 3$ and the palindromic sequence $B, C, D, \ldots, D, C, B$ is just the three-term sequence 1, 2, 1.

How much of this was known to the ancient Greeks? It seems very likely that they knew quite a lot—very possibly *all* the things that I have described above (including Lagrange's theorem), although they may well have lacked rigorous proofs for everything. Plato's contemporary Theaetetos seems to have established much of this. There appears even to be some evidence of this knowledge (including the repeating palindromic sequences referred to above) revealed in Plato's dialectics.[7]

Although incorporating the quadratic irrationals gets us some way towards numbers adequate for Euclidean geometry, it does not do all that is needed. In the tenth (and most difficult) book of Euclid, numbers like $\sqrt{a + \sqrt{b}}$ are considered (with a and b positive rationals). These are *not* generally quadratic irrationals, but they occur, nevertheless, in ruler-and-compass constructions. Numbers sufficient for such geometric constructions would be those that can be built up from natural numbers by repeated use of the operations of addition, subtraction, multiplication, division, and the taking of square roots. But operating exclusively with such numbers gets extremely complicated, and these numbers are still too limited for considerations of Euclidean geometry that go beyond ruler-and-compass constructions. It is much more satisfactory to take the bold step—and how bold a step this actually is will be indicated in §§16.3–5—of allowing infinite continued-fraction expressions that are completely general. This provided the Greeks with a way of describing numbers that do turn out to be adequate for Euclidean geometry.

These numbers are indeed, in modern terminology, the so-called 'real numbers'. Although a fully satisfactory definition of such numbers is not regarded as having been found until the 19th century (with the work of Dedekind, Cantor, and others), the great ancient Greek mathematician and astronomer Eudoxos, who had been one of Plato's students, had obtained the essential ideas already in the 4th century BC. A few words about Eudoxos's ideas are appropriate here.

First, we note that the numbers in Euclidean geometry can be expressed in terms of *ratios* of lengths, rather than directly in terms of lengths. In this way, no specific unit of length (such as 'inch' or Greek 'dactylos') was needed. Moreover, with ratios of lengths, there would be no restriction as to how many such ratios might be multiplied together (obviating the apparent need for higher-dimensional 'hypervolumes' when more than three lengths are multiplied together). The first step in the Eudoxan theory was to supply a criterion as to when a length ratio $a : b$ would be *greater* than another such ratio $c : d$. This criterion is that some positive integers M and N exist such that the length a added to itself M times exceeds b added to itself N times, while also d added to itself N times exceeds c added to itself M times.[3.3] A corresponding criterion holds expressing the condition that the ratio $a : b$ be *less* than the ratio $c : d$. The condition for equality of these ratios would be that neither of these criteria hold. With this ingenious notion of 'equality' of such ratios, Eudoxos had, in effect, an

[3.3] Can you see why this works?

abstract concept of a 'real number' in terms of length ratios. He also provided rules for the sum and product of such real numbers.[3.4]

There was a basic difference in viewpoint, however, between the Greek notion of a real number and the modern one, because the Greeks regarded the number system as basically 'given' to us, in terms of the notion of *distance* in physical space, so the problem was to try to ascertain how these 'distance' measures actually behaved. For 'space' may well have had the appearance of being itself a Platonic absolute even though actual physical objects existing in this space would inevitably fall short of the Platonic ideal.[8] (However, we shall be seeing in §17.9 and §§19.6,8 how Einstein's general theory of relativity has now changed this perspective on space and matter in a fundamental way.)

A physical object such as a square drawn in the sand or a cube hewn from marble might have been regarded by the ancient Greeks as a reasonable or sometimes an excellent approximation to the Platonic geometrical ideal. Yet any such object would nevertheless provide a mere approximation. Lying behind such approximations to the Platonic forms—so it would have appeared—would be space itself: an entity of such abstract or notional existence that it could well have been regarded as a direct realization of a Platonic reality. The measure of distance in this ideal geometry would be something to *ascertain*; accordingly, it would be appropriate to try to extract this ideal notion of real number from a geometry of a Euclidean space that was assumed to be *given*. In effect, this is what Eudoxos succeeded in doing.

By the 19th and 20th centuries, however, the view had emerged that the mathematical notion of number should stand separately from the nature of physical space. Since mathematically consistent geometries other than that of Euclid had been shown to exist, this rendered it inappropriate to insist that the mathematical notion of 'geometry' should be necessarily extracted from the supposed nature of 'actual' physical space. Moreover, it could be very difficult, if not impossible, to ascertain the detailed nature of this supposed underlying 'Platonic physical geometry' in terms of the behaviour of imperfect physical objects. In order to know the nature of the numbers according to which 'geometrical distance' is to be defined, for example, it would be necessary to know what happens both at indefinitely tiny and indefinitely large distances. Even today, these questions are without clear-cut resolution (and I shall be addressing them again in later chapters). Thus, it was far more appropriate to develop the nature of number in a way that does not directly refer to physical measures. Accordingly, Richard Dedekind and Georg Cantor developed their ideas of what real numbers 'are' by use of notions that do not directly refer to geometry.

[3.4] Can you see how to formulate these?

Dedekind's definition of a real number is in terms of infinite sets of rational numbers. Basically, we think of the rational numbers, both positive and negative (and zero), to be arranged in order of size. We can imagine that this ordering takes place from left to right, where we think of the negative rationals as being displayed going off indefinitely to the left, with 0 in the middle, and the positive rationals displayed going off indefinitely to the right. (This is just for visualization purposes; in fact Dedekind's procedure is entirely abstract.) Dedekind imagines a 'cut' which divides this display neatly in two, with those to the left of the cut being all smaller than those to the right. When the 'knife-edge' of the cut does not 'hit' an actual rational number but falls between them, we say that it defines an *irrational* real number. More correctly, this occurs when those to the left have no actual largest member and those to the right, no actual smallest one. When the system of 'irrationals', as defined in terms of such cuts, is adjoined to the system of rational numbers that we already have, then the complete family of *real numbers* is obtained.

Dedekind's procedure leads, by means of simple definitions, directly to the laws of addition, subtraction, multiplication, and division for real numbers. Moreover, it enables one to go further and define *limits*, whereby such things as the infinite continued fraction that we saw before

$$1+(2+(2+(2+(2+\cdots)^{-1})^{-1})^{-1})^{-1}$$

or the infinite sum

$$1-\frac{1}{3}+\frac{1}{5}-\frac{1}{7}+\frac{1}{9}-\cdots$$

may be assigned real-number meanings. In fact, the first gives us the irrational number $\sqrt{2}$, and the second, $\frac{1}{4}\pi$. The ability to take limits is fundamental for many mathematical notions, and it is this that gives the real numbers their particular strengths.[9] (The reader may recall that the need for 'limiting procedures' was a requirement for the general definition of areas, as was indicated in §2.3.)

3.3 Real numbers in the physical world

There is a profound issue that is being touched upon here. In the development of mathematical ideas, one important initial driving force has always been to find mathematical structures that accurately mirror the behaviour of the physical world. But it is normally not possible to examine the physical world itself in such precise detail that appropriately clear-cut mathematical notions can be abstracted directly from it. Instead, progress is made because mathematical notions tend to have a 'momentum' of their own that appears to spring almost entirely from within the subject itself. Mathematical ideas develop, and various kinds of problem seem to arise naturally. Some of these (as was the case with the problem of finding the length of the diagonal of a square) can lead to an essential extension of the original mathematical concepts in terms of which the problem had been formulated. Such extensions may seem to be forced upon us, or they may arise in ways that appear to be matters of convenience, consistency, or mathematical elegance. Accordingly, the development of mathematics may seem to diverge from what it had been set up to achieve, namely simply to reflect physical behaviour. Yet, in many instances, this drive for mathematical consistency and elegance takes us to mathematical structures and concepts which turn out to mirror the physical world in a much deeper and more broad-ranging way than those that we started with. It is as though Nature herself is guided by the same kind of criteria of consistency and elegance as those that guide human mathematical thought.

An example of this is the real-number system itself. We have no direct evidence from Nature that there is a physical notion of 'distance' that extends to arbitrarily large scales; still less is there evidence that such a notion can be applied on the indefinitely tiny level. Indeed, there is no evidence that 'points in space' actually exist in accordance with a geometry that precisely makes use of real-number distances. In Euclid's day, there was scant evidence to support even the contention that such Euclidean 'distances' extended outwards beyond, say, about 10^{12} metres,[10] or inwards to as little as 10^{-5} metres. Yet, having been driven mathematically by the consistency and elegance of the real-number system, all of our broad-ranging and successful physical theories to date have, without exception, still clung to this ancient notion of 'real number'. Although there might appear to have been little justification for doing this from the evidence that was available in Euclid's day, our faith in the real-number system appears to have been rewarded. For our successful modern theories of cosmology now allow us to extend the range of our real-number distances out to about 10^{26} metres or more, while the accuracy of our theories of particle physics extends this range inwards to 10^{-17} metres or less. (The only scale at which it has been seriously proposed that a change might come about is some 18 orders of magnitude smaller even than that, namely 10^{-35} metres, which is the 'Planck scale' of quantum gravity that will feature strongly in some of our later discussions; e.g. §§31.1,6–12,14 and §32.7.) It may be regarded as a remarkable justification of our use of mathematical idealizations that the range of validity of the real-number system has extended from the total of about 10^{17}, from the smallest to the largest, that seemed appropriate in Euclid's day to at least the 10^{43} that our theories directly employ today, this representing a stupendous increase by a factor of some 10^{26}.

There is a good deal more to the physical validity of the real-number system than this. In the first place, we must consider that areas and volumes are also quantities for which real-number measures are accurately appropriate. A volume measure is the cube of a distance measure (and an area is the square of a distance). Accordingly, in the case of volumes, we may consider that it is the cube of the above range that is relevant. For Euclid's time, this would give us a range of about $(10^{17})^3 = 10^{51}$; for today's theories, at least $(10^{43})^3 = 10^{129}$. Moreover, there are other physical measures that require real-number descriptions, according to our presently successful theories. The most noteworthy of these is time. According to relativity theory, this needs to be adjoined to space to provide us with *spacetime* (which is the subject of our deliberations in Chapter 17). Spacetime volumes are four-dimensional, and it might well be considered that the temporal range (of again about 10^{43} or more in total range, in our well-tested theories) should also be incorporated into our considerations, giving a total of something like at least 10^{172}. We shall see some far larger real numbers even than this coming into our later considerations (see §27.13 and §28.7), although it is not really clear in some cases that the use of real numbers (rather than, say, integers) is essential.

More importantly for physical theory, from Archimedes, through Galileo and Newton, to Maxwell, Einstein, Schrödinger, Dirac, and the rest, a crucial role for the real-number system has been that it provides a necessary framework for the standard formulation of the *calculus* (see Chapter 6). All successful dynamical theories have required notions of the calculus for their formulations. Now, the conventional approach to calculus requires the *infinitesimal* nature of the reals to be what it is. That is to say, on the small end of the scale, it is the entire range of the real numbers that is in principle being made use of. The ideas of calculus underlie other physical notions, such as velocity, momentum, and energy. Consequently, the real-number system enters our successful physical theories in a fundamental way for our description of all these quantities also. Here, as mentioned earlier in connection with areas, in §2.3 and §3.2, the infinitesimal limit of small-scale structure of the real-number system is being called upon.

Yet we may still ask whether the real-number system is really 'correct' for the description of physical reality at its deepest levels. When quantum-mechanical ideas were beginning to be introduced early in the 20th century, there was the feeling that perhaps we were now beginning to witness a discrete or granular nature to the physical world at its smallest scales.[11] Energy could apparently exist only in discrete bundles—or 'quanta'—and the physical quantities of 'action' and 'spin' seemed to occur only in discrete multiples of a fundamental unit (see §§20.1,5 for the classical concept of *action* and §26.6 for its quantum counterpart; see §§22.8–12 for *spin*). Accordingly, various physicists attempted to build up an alternative picture of the world in which discrete processes governed all actions at the tiniest levels.

However, as we now understand quantum mechanics, that theory does not force us (nor even lead us) to the view that there is a discrete or granular nature to space, time, or energy at its tiniest levels (see Chapters 21 and 22, particularly the last sentence of §22.13). Nevertheless, the idea has remained with us that there may indeed be, at root, such a fundamental discreteness to Nature, despite the fact that quantum mechanics, in its standard formulation, certainly does not imply this. For example, the great quantum physicist Erwin Schrödinger was among the first to propose that a change to some form of fundamental spatial discreteness might actually be necessary:[12]

> The idea of a *continuous range*, so familiar to mathematicians in our days, is something quite exorbitant, an enormous extrapolation of what is accessible to us.

He related this proposal to some early Greek thinking concerning the discreteness of Nature. Einstein, also, suggested, in his last published words, that a discretely based ('algebraic') theory might be the way forward for the future physics:[13]

> One can give good reasons why reality cannot be represented as a continuous field....Quantum phenomena...must lead to an attempt to find a purely algebraic theory for the description of reality. But nobody knows how to obtain the basis of such a theory.[14]

Others[15] also have pursued ideas of this kind; see §33.1. In the late 1950s, I myself tried this sort of thing, coming up with a scheme that I referred to as the theory of 'spin networks', in which the discrete nature of quantum-mechanical *spin* is taken as the fundamental building block for a *combinatorial* (i.e. discrete rather than real-number-based) approach to physics. (This scheme will be briefly described in §32.6.) Although my own ideas along this particular direction did not develop to a comprehensive theory (but, to some extent, became later transmogrified into 'twistor theory'; see §33.2), the theory of spin networks has now been imported, by others, into one of the major programmes for attacking the fundamental problem of *quantum gravity*.[16] I shall give brief descriptions of these various ideas in Chapter 32. Nevertheless, as tried and tested physical theory stands today—as it has for the past 24 centuries—real numbers still form a fundamental ingredient of our understanding of the physical world.

3.4 Do natural numbers need the physical world?

In the above description, in §3.2, of the Dedekind approach to the real-number system, I have presupposed that the *rational numbers* are already taken as 'understood'. In fact, it is not a difficult step from the integers to the rationals; rationals are just ratios of integers (see the Preface). What about the integers themselves, then? Are these rooted in physical ideas? The discrete approaches to physics that were referred to in the previous two paragraphs certainly depend upon our notion of *natural number* (i.e. 'counting number') and its extension, by the inclusion of the negative numbers, to the integers. Negative numbers were not considered, by the Greeks, to be actual 'numbers', so let us continue our considerations by first asking about the physical status of the natural numbers themselves.

The *natural numbers* are the quantities that we now denote by 0, 1, 2, 3, 4, etc., i.e. they are the non-negative whole numbers. (The modern procedure is to include 0 in this list, which is an appropriate thing to do from the mathematical point of view, although the ancient Greeks appear not to have recognized 'zero' as an actual number. This had to wait for the Hindu mathematicians of India, starting with Brahmagupta in 7th century and followed up by Mahavira and Bhaskara in the 9th and 12th century, respectively.) The role of the natural numbers is clear and unambiguous. They are indeed the most elementary 'counting numbers', which have a basic role whatever the laws of geometry or physics might be. Natural numbers are subject to certain familiar operations, most particularly the operations of *addition* (such as $37 + 79 = 116$) and *multiplication* (e.g. $37 \times 79 = 2923$), which enable pairs of natural numbers to be combined together to produce new natural numbers. These operations are independent of the nature of the geometry of the world.

We can, however, raise the question of whether the natural numbers themselves have a meaning or indeed existence independent of the actual nature of the physical world. Perhaps our notion of natural numbers depends upon there being, in our universe, reasonably well-defined discrete objects that persist in time. Natural numbers initially arise when we wish to count things, after all. But this seems to depend upon there actually being persistent distinguishable 'things' in the universe which are available to be 'counted'. Suppose, on the other hand, our universe were such that numbers of objects had a tendency to keep changing. Would natural numbers actually be 'natural' concepts in such a universe? Moreover, perhaps the universe actually contains only a finite number of 'things', in which case the 'natural' numbers might themselves come to an end at some point! We can even envisage a universe which consists only of an amorphous featureless substance, for which the very notion of numerical quantification might seem intrinsically inappropriate. Would the notion of 'natural number' be at all relevant for the description of universes of this kind?

Even though it might well be the case that inhabitants of such a universe would find our present mathematical concept of a 'natural number' difficult to come upon, it is hard to imagine that there would not still be an important role for such fundamental entities. There are various ways in which natural numbers can be introduced in pure mathematics, and these do not seem to depend upon the actual nature of the physical universe at all. Basically, it is the notion of a 'set' which needs to be brought into play, this being an abstraction that does not appear to be concerned, in any essential way, with the specific structure of the physical universe. In fact, there are certain definite subtleties concerning this question, and I shall return to that issue later (in §16.5). For the moment, it will be convenient to ignore such subtleties.

Let us consider one way (developed by Cantor from ideas of Giuseppe Peano, and promoted by the distinguished mathematician John von Neumann) that natural numbers can be introduced merely using the abstract notion of set. It also leads on to what are called 'ordinal numbers'. The simplest set of all is referred to as the 'null set' or the 'empty set', and it is characterized by the fact that it contains no members whatever! The empty set is usually denoted by the symbol $\varnothing$, and we can write this definition

$$\varnothing = \{\ \},$$

where the curly brackets delineate a *set*, the specific set under consideration having, as its members, the quantities indicated within the brackets. In this case, there is nothing within the brackets, so the set being described is indeed the empty set. Let us associate $\varnothing$ with the natural number 0. We can now proceed further and define the set whose only member is $\varnothing$; i.e. the set $\{\varnothing\}$. It is important to realize that $\{\varnothing\}$ is not the same set as the empty set $\varnothing$. The set $\{\varnothing\}$ has *one* member (namely $\varnothing$), whereas $\varnothing$ itself has none at all. Let us associate $\{\varnothing\}$ with the natural number 1. We next define the set whose two members are the two sets that we just encountered, namely $\varnothing$ and $\{\varnothing\}$, so this new set is $\{\varnothing, \{\varnothing\}\}$, which is to be associated with the natural number 2. Then we associate with 3 the collection of all the three entities that we have encountered up to this point, namely the set $\{\varnothing, \{\varnothing\}, \{\varnothing, \{\varnothing\}\}\}$, and with 4 the set $\{\varnothing, \{\varnothing\}, \{\varnothing, \{\varnothing\}\}, \{\varnothing, \{\varnothing\}, \{\varnothing, \{\varnothing\}\}\}\}$, whose members are again the sets that we have encountered previously, and so on. This may not be how we usually think of natural numbers, as a matter of definition, but it is one of the ways that mathematicians can come to the concept. (Compare this with the discussion in the Preface.) Moreover, it shows us, at least, that things like the natural numbers[17] can be conjured literally out of nothing, merely by employing the abstract notion of 'set'. We get an infinite sequence of abstract (Platonic) mathematical entities—sets containing, respectively, zero, one, two, three, etc., elements, one set for each of the natural numbers, quite independently of the actual physical nature of the universe. In Fig.1.3 we envisaged a kind of independent 'existence' for Platonic mathematical notions—in this case, the natural numbers themselves—yet this 'existence' can seemingly be conjured up by, and certainly accessed by, the mere exercise of our mental imaginations, without any reference to the details of the nature of the physical universe. Dedekind's construction, moreover, shows how this 'purely mental' kind of procedure can be carried further, enabling us to 'construct' the entire system of real numbers,[18] still without any reference to the actual physical nature of the world. Yet, as indicated above, 'real numbers' indeed seem to have a direct relevance to the real structure of the world—illustrating the very mysterious nature of the 'first mystery' depicted in Fig.1.3.

3.5 Discrete numbers in the physical world

But I am getting slightly ahead of myself. We may recall that Dedekind's construction really made use of sets of *rational* numbers, not of natural numbers directly. As indicated above, it is not hard to 'define' what we mean by a rational number once we have the notion of natural number. But, as an intermediate step, it is appropriate to define the notion of an *integer*, which is a natural number or the *negative* of a natural number (the number zero being its own negative). In a formal sense, there is no difficulty in giving a mathematical definition of 'negative': roughly speaking we just attach a 'sign', written as '−', to each natural number (except 0) and define all the arithmetical rules of addition, subtraction, multiplication, and division (except by 0) consistently. This does not address the question of the 'physical meaning' of a negative number, however. What might it mean to say that there are minus three cows in a field, for example?

I think that it is clear that, unlike the natural numbers themselves, there is no evident physical content to the notion of a negative number of physical objects. Negative integers certainly have an extremely valuable organizational role, such as with bank balances and other financial transactions. But do they have direct relevance to the *physical* world? When I say 'direct relevance' here, I am not referring to circumstances where it would appear that it is negative real numbers that are the relevant measures, such as when a distance measured in one direction counts as positive while that measured in the opposite direction would count as negative (or the same thing with regard to time, in which times extending into the past might count as negative). I am referring, instead, to numbers that are *scalar* quantities, in the sense that there is no directional (or temporal) aspect to the quantity in question. In these circumstances it appears to be the case that it is the system of integers, both positive and negative, that has direct physical relevance.

It is a remarkable fact that only in about the last hundred years has it become apparent that the system of integers does indeed seem to have such direct physical relevance. The first example of a physical quantity which seems to be appropriately quantified by integers is *electric charge*.[19] As far as is known (although there is as yet no complete theoretical justification of this fact), the electric charge of any discrete isolated body is indeed quantified in terms of integral multiples, positive, negative, or zero, of one particular value, namely the charge on the proton (or on the electron, which is the negative of that of the proton).[20] It is now believed that protons are composite objects built up, in a sense, from smaller entities referred to as 'quarks' (and additional chargeless entities called 'gluons'). There are three quarks to each proton, the quarks having electric charges with respective values $\frac{2}{3}$, $\frac{2}{3}$, $-\frac{1}{3}$. These constituent charges add up to give the total value 1 for the proton. If quarks are fundamental entities, then the basic charge unit is one third of that which we seemed to have before. Nevertheless, it is still true that electric charge is measured in terms of integers, but now it is integer multiples of one third of a proton charge. (The role of quarks and gluons in modern particle physics will be discussed in §§25.3–7.)

Electric charge is just one instance of what is called an *additive quantum number*. Quantum numbers are quantities that serve to characterize the particles of Nature. Such a quantum number, which I shall here take to be a real number of some kind, is 'additive' if, in order to derive its value for a composite entity, we simply add up the individual values for the constituent particles—taking due account of the signs, of course, as with the above-mentioned case of the proton and its constituent quarks. It is a very striking fact, according to the state of our present physical knowledge, that all known additive quantum numbers[21] are indeed quantified in terms of the system of integers, not general real numbers, and not simply natural numbers—so that the negative values actually do occur.

In fact, according to 20th-century physics, there is now a certain sense in which it *is* meaningful to refer to a negative number of physical entities. The great physicist Paul Dirac put forward, in 1929–31, his theory of antiparticles, according to which (as it was later understood), for each type of particle, there is also a corresponding *antiparticle* for which each additive quantum number has precisely the negative of the value that it has for the original particle; see §§24.2,8. Thus, the system of integers (with negatives included) does indeed appear to have a clear relevance to the physical universe—a physical relevance that has become apparent only in

the 20th century, despite those many centuries for which integers have found great value in mathematics, commerce, and many other human activities.

One important qualification should be made at this juncture, however. Although it is true that, in a sense, an antiproton is a negative proton, it is not really 'minus one proton'. The reason is that the sign reversal refers only to *additive* quantum numbers, whereas the notion of *mass* is not additive in modern physical theory. This issue will be explained in a bit more detail in §18.7. 'Minus one proton' would have to be an antiproton whose mass is the negative of the mass value of an ordinary proton. But the mass of an actual physical particle is not allowed to be negative. An antiproton has the same mass as an ordinary proton, which is a positive mass. We shall be seeing later that, according to the ideas of quantum field theory, there are things called 'virtual' particles for which the mass (or, more correctly, energy) can be negative. 'Minus one proton' would really be a virtual antiproton. But a virtual particle does not have an independent existence as an 'actual particle'.

Let us now ask the corresponding question about the rational numbers. Has this system of numbers found any direct relevance to the physical universe? As far as is known, this does not appear to be the case, at least as far as conventional theory is concerned. There are some physical curiosities[22] in which the family of rational numbers does play its part, but it would be hard to maintain that these reveal any fundamental physical role for rational numbers. On the other hand, it may be that there is a particular role for the rationals in fundamental quantum-mechanical probabilities (a rational probability possibly representing a choice between alternatives, each of which involves just a finite number of possibilities). This kind of thing plays a role in the theory of spin networks, as will be briefly described in §32.6. As of now, the proper status of these ideas is unclear.

Yet, there are other kinds of number which, according to accepted theory, do appear to play a fundamental role in the workings of the universe. The most important and striking of these are the *complex numbers*, in which the seemingly mystical quantity $\sqrt{-1}$, usually denoted by 'i', is introduced and adjoined to the real-number system. First encountered in the 16th century, but treated for hundreds of years with distrust, the mathematical utility of complex numbers gradually impressed the mathematical community to a greater and greater degree, until complex numbers became an indispensable, even magical, ingredient of our mathematical thinking. Yet we now find that they are fundamental not just to mathematics: these strange numbers also play an extraordinary and very basic role in the operation of the physical universe at its tiniest scales. This is a cause for wonder, and it is an even more striking instance of the convergence between mathematical ideas and the deeper workings of the physical universe than is the system of real numbers that we have been considering in this section. Let us come to these remarkable numbers next.

Notes

Section 3.1

3.1. The notations $>$, $<$, $\geqslant$, $\leqslant$, frequently used in this book, respectively stand for 'is greater than', 'is less than', 'is greater than or equal to', and 'is less than or equal to' (made appropriately grammatical).

3.2. Some readers might be aware of an apparently shorter argument which starts by demanding that a/b be 'in its lowest terms' (i.e. that a and b have no common factor). However, this assumes that such a lowest-terms expression always exists, which, though perfectly true, needs to be shown. Finding a lowest-term expression for a given fraction A/B (implicitly or explicitly—say using the procedure known as Euclid's algorithm; see, for example, Hardy and Wright 1945, p. 134; Davenport 1952, p. 26; Littlewood 1949, Chap. 4; and Penrose 1989, Chap. 2) involves reasoning similar to that given in the text, but more complicated.

3.3. One might well object that it is somewhat curious to use real numbers in the above proof, since the 'real rationals' (i.e. quotients of reals) would simply be real numbers all over again. This does not invalidate what has just been said, however. It may be remarked that it is as well that a and b were taken to be integers, in the original argument, and not themselves taken to be rationals. For, if a and b were merely rational, then the argument would fail at the 'decreasing sequence' part, even though the result itself would still be true.

Section 3.2

3.4. At a casual glance, expressions like $a+(b+(c+(d+\cdots)^{-1})^{-1})^{-1}$ may look rather odd. However, they are very natural in the context of ancient Greek thinking (although the Greeks did not use this particular notation). The procedure of *Euclid's algorithm* was referred to in Note 3.2 in the context of finding the lowest-term form of a fraction. Euclid's algorithm (when unravelled) leads precisely to such a continued fraction expression. The Greeks would apply this same procedure to the ratio of two geometrical lengths. In the most general case, the result would be an *infinite* continued fraction, of the kind considered here.

3.5. For more information (with proofs) concerning continued fractions, see the elegant account given in Chapter 4 of Davenport (1952). It may be remarked that in certain respects the continued-fraction representation of real numbers is deeper and more interesting than the normal one in terms of decimal expansions, finding applications in many different areas of modern mathematics, including the hyperbolic geometry discussed in §§2.4,5. On the other hand, continued fractions are not at all well suited for (most) practical calculation, the conventional decimal representation being far easier to use.

3.6. Quadratic irrationals are so called because they arise in the solution of a general quadratic equation

$$Ax^2+Bx+C=0,$$

with A non-zero, the solutions being

$$-\frac{B}{2A}+\sqrt{\left(\frac{B}{2A}\right)^2-\frac{C}{A}} \text{ and } -\frac{B}{2A}-\sqrt{\left(\frac{B}{2A}\right)^2-\frac{C}{A}}$$

where, to keep within the realm of real numbers, we must have B^2 greater than $4AC$. When A, B, and C are integers or rational numbers, and where there is no rational solution to the equation, the solutions are indeed quadratic irrationals.

3.7. Professor Stelios Negrepontis informs me that this evidence is to be found in the Platonic dialogue the Statesman (= Politikos), the third in the 'trilogy' the Theaetetos-the Sophist-the Politikos. See Negrepontis (2000).

3.8. See Sorabji (1984, 1988) for an account of ancient Greek thinking on the nature of space.

3.9. See Hardy (1914); Conway (1976); Burkill (1962).

Section 3.3

3.10. The scientific notation '10^{12}' for a 'million million' makes use of *exponents*, as described in Notes 1.2 and 2.4. In this book, I shall tend to avoid verbal terms such as 'million', and especially 'billion', in preference to this much clearer scientific notation. The word 'billion' is particularly confusing, as in American usage—now commonly adopted also in the UK—'billion' refers to 10^9, whereas, in the older (more logical) UK usage, in agreement with most other European languages, it refers to 10^{12}. Negative exponents, such as in 10^{-6} (which refers to 'one millionth'), are also used here in accordance with the normal scientific notation.

The distance 10^{12} metres is about 7 times the Earth–Sun separation.This is roughly the distance of the planet Jupiter, although that distance was not known in Euclid's day and would have been guessed to be rather smaller.

3.11. See, for example, Russell (1903), Chap. 4.

3.12. Schrödinger (1952), pp. 30–1.

3.13. See Stachel (1995).

3.14. Einstein (1955), p. 166.

3.15. See e.g. Snyder (1947); Schild (1949); and Ahmavaara (1965).

3.16. See Ashtekar (1986); Ashtekar and Lewandowski (2004); Smolin (1998, 2001); Rovelli (1998, 2003).

Section 3.4

3.17. The notion of 'ordinal number', that is implied here in the finite case, extends also to *infinite* ordinal numbers, the smallest being Cantor's 'ω', which is the ordered collection of *all* finite ordinals.

3.18. This notion of 'construct' should not be taken in too strong a sense, however. We shall be finding in §16.6 that there are certain real numbers (in fact most of them) that are inaccessible by any computational procedure.

Section 3.5

3.19. The Irish physicist George Johnstone Stoney was the first, in 1874, to give a (crude) estimate of the basic electric charge, and, in 1891, coined the term 'electron' for this fundamental unit. In 1909, the American physicist Robert Andrews Millikan designed his famous 'oil-drop' experiment, which precisely showed that the charge on electrically charged bodies (the oil drops, in his experiment) came in integer multiples of a well-defined value—the electron charge.

3.20. In 1959, R. A. Lyttleton and H. Bondi proposed that a slight difference in the proton and (minus) the electron charges, of the order of one part in 10^{18} might account for the expansion of the universe, (for which, see §§27.11,13, and Chapter 28). See Lyttleton and Bondi (1959). Unfortunately, for this theory, such a discrepancy was soon disproved in several experiments. Nevertheless, this idea provided an excellent example of creative thinking.

3.21. I am here distinguishing the 'additive' quantum numbers from the numbers that physicists call 'multiplicative', which we shall come to in §5.5.

3.22. For example, in the 'fractional quantum Hall effect', one finds rational numbers playing a key role; see, for example, Fröhlich and Pedrini (2000).

4
Magical complex numbers

4.1 The magic number 'i'

How is it that -1 can have a square root? The square of a positive number is always positive, and the square of a negative number is again positive (and the square of 0 is just 0 again, so that is hardly of use to us here). It seems impossible that we can find a number whose square is actually negative. Yet, this is the kind of situation that we have seen before, when we ascertained that 2 has no square root within the system of rational numbers. In that case we resolved the situation by extending our system of numbers from the rationals to a larger system, and we settled on the system of reals. Perhaps the same trick will work again.

Indeed it will. In fact what we have to do is something much easier and far less drastic than the passage from the rationals to the reals. (Raphael Bombelli introduced the procedure in 1572 in his work *L'Algebra*, following Gerolamo Cardano's original encounters with complex numbers in his *Ars Magna* of 1545.) All we need do is introduce a single quantity, called 'i', which is to square to -1, and adjoin it to the system of reals, allowing combinations of i with real numbers to form expressions such as

$$a + \mathrm{i}b,$$

where a and b are arbitrary real numbers. Any such combination is called a *complex number*. It is easy to see how to add complex numbers:

$$(a + \mathrm{i}b) + (c + \mathrm{i}d) = (a + c) + \mathrm{i}(b + d)$$

which is of the same form as before (with the real numbers $a + c$ and $b + d$ taking the place of the a and b that we had in our original expression). What about multiplication? This is almost as easy. Let us find the product of $a + \mathrm{i}b$ with $c + \mathrm{i}d$. We first simply multiply these factors, expanding the expression using the ordinary rules of algebra:[1]

$$\begin{aligned}(a + \mathrm{i}b)(c + \mathrm{i}d) &= ac + \mathrm{i}bc + a\mathrm{i}d + \mathrm{i}b\mathrm{i}d\\ &= ac + \mathrm{i}(bc + ad) + \mathrm{i}^2bd.\end{aligned}$$