John Barrow: The Universe that discovered itself: Oxford 2000

7

Selection effects

> Procrustes, you will remember, stretched or chopped down his guests to fit the bed he had constructed. But perhaps you have not heard the rest of the story. He measured them up before they left next morning, and wrote a learned paper 'On the Uniformity of Stature of Travellers' for the Anthropological Society of Attica.
>
> A. S. Eddington

Patterns in the trees

> Let every student of nature take this as his rule that whatever the mind seizes upon with particular satisfaction is to be held in suspicion.
>
> Francis Bacon

On London underground trains there exists a particular type of advertisement which I can even remember seeing as a child. It informs you that if you supply the next entry in several sequences of numbers, or pick the odd one out of a collection of superficially similar shapes, then you could land a well-paid job in the wonderful company whose telephone number you can find listed below *et cetera*. Advertisements like this illustrate how readily our society has come to equate the ability to spot patterns or relationships, usually those of a mathematical or geometrical nature, with 'intelligence'. Indeed, most IQ tests place a very heavy emphasis upon these mental abilities. Whether or not they measure anything as definable as 'intelligence' is usually irrelevant to those who set these tests. They are interested primarily in that specialized ability of pattern recognition and isolation. In this sense, intelligence is defined as being what intelligence tests measure. It is likely that our very existence as a species owes much to this ability to delineate patterns. Evolution may have made us a little too adept at spotting them. Our propensity to see patterns where none exist at all is witnessed by ancient Man's enthusiastic identification of ploughs and hunters, crabs and scales tracing out the patterns of stars we call the constellations, or by modern Man's enthusiasm for Martian canals. But this disposition to

identify patterns where none exist is a better one to be saddled with than a failure to perceive patterns that really do exist. If you keep telling your family that you see tigers in the trees when there aren't any they will merely think you are a little paranoid, but fail to see tigers in the trees when they really are there and you're dead! Over-sensitive pattern recognition tends to survive.

Today, tigers are not such a problem. But we have inherited an ability to perceive patterns that in some sense are not really there. The fact that this propensity varies from individual to individual is something that psychologists have attempted to exploit by using the Rorschach ink-blot personality test to evaluate mentally disturbed patients. It is interesting to undertake some simple experiments to see how the eye and brain operate. The picture drawn below in Figure 7.1 consists of many concentric circles of dots. Each circle contains the same number of dots, and the dots on one circle lie halfway between those of the circles on either side of it. So the dots all lie along straight lines passing through the centre of the system of circles. Thus the patterns which 'really' exist in the sense of having been built into the picture deliberately, are circles and straight lines. But what does the human eye perceive? Close to the centre of the picture we do see circular rings, but towards the outside the eye picks out crescent-shaped petals. It does this because the brain is adept at drawing imaginary lines from a point to its nearest neighbour. Near the centre of the picture the dots are closely ringed, so that the nearest neighbour to any particular dot is one next-door to it on the same circular ring. As we look farther from the centre the dots on any circle become more spaced-out, and the nearest neighbour is to be found on an adjacent circle. The crescents are the lines the eye most readily draws in your mind between these closest neighbours. As a further experiment you might like to tilt the book, and look at the picture along the page. You will see a new set of apparent patterns because the different perspective creates a new set of nearest neighbours for the points. The patterns change according to your angle of view. Figure 7.2 reveals the confusion that your brain experiences when it cannot decide upon a pattern. It continually changes its fix, and this creates a peculiar dynamic effect. There are many influences at work in producing a percept of Figure 7.2, not least of which is the positioning of prominent visual cues at the centres of circular arrays of points traced by nearest neighbours. The eye identifies the symmetrical patterns most readily.

This little game of illusions has a serious side to it. At the moment astronomers are trying to determine whether significant intrinsic patterns of lines and cells really exist in the observed distribution of galaxies in the Universe. Patterns certainly do 'appear' to exist, but it is not clear whether they are just chance effects highlighted by the eye's peculiar fondness for

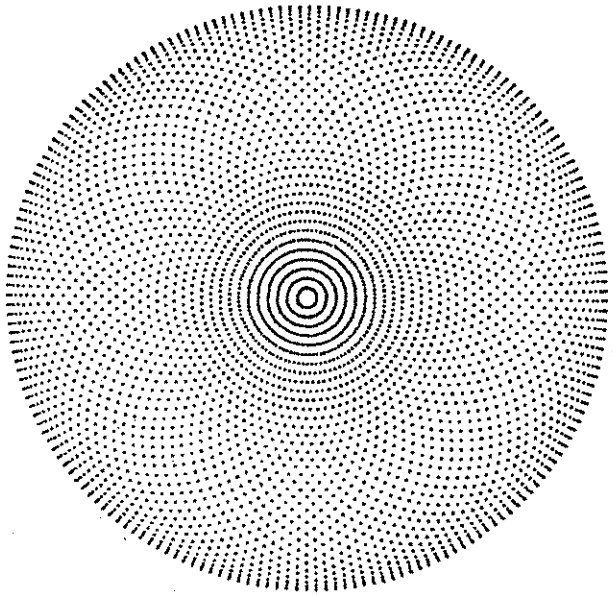

Figure 7.1 This pattern is composed of equally-spaced concentric rings of dots. Each ring contains the same number of dots, and points in alternate rings are radially aligned. Despite these built-in patterns of circles and straight lines, the eye perceives three dominant patterns: concentric rings near the centre, crescent-shaped 'petals' farther out, and radial straight lines near the perimeter. These impressions are dictated by the tendency of the eye to trace out imaginary lines between nearest neighbours. Near the centre of the figure the nearest neighbour of any point lies on the same circle. Farther out the nearest neighbour is to be found on an adjacent circle, until there is crowding near the periphery. If the reader tilts the page and views the Figure at an angle it will be found that the pattern has completely changed. This change reflects the fact that the nearest neighbours are altered by the projection effect of the inclined viewpoint.

patterns, or attributable to features intrinsic to the galaxy formation process.

Psychologically we find pattern, symmetry, and order appealing. Throughout the arts of ancient cultures we find symmetrical patterns of great sophistication developed for purely decorative purposes. Subsequently, some of these patterns and symmetries have been found to possess sophisticated mathematical properties. In modern times many examples of the draughtsmanship of Maurits Escher (see for example Figure 5.1) have turned out to exploit very subtle mathematical symmetries which he

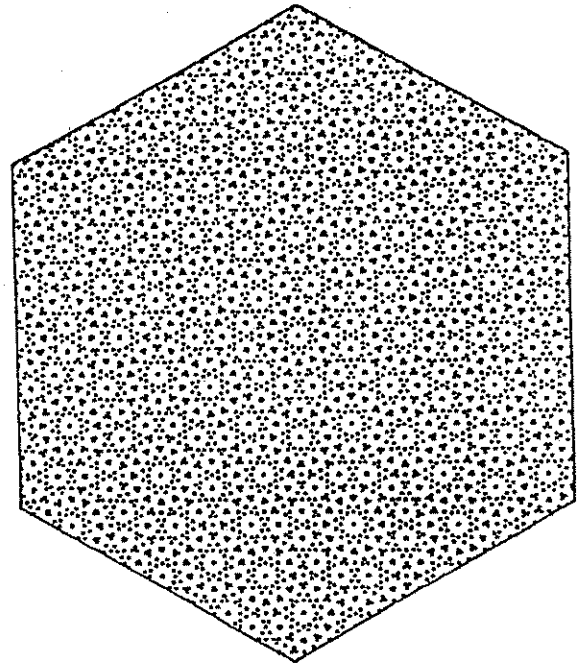

Figure 7.2 The Marroquin Design. In this design a continual flux of different patterns is evident. The eye has a tendency to seek out larger and larger circular patterns which eventually dissolve to be replaced by the distinctive local centres of the pattern. This design also contains a more subtle pattern which many viewers seem unable to discern. Within each of the largest of the circles that the eye easily picks out there is inscribed a twelve-sided 'Swiss' cross and this pattern is repeated periodically throughout the design. [From Marroquin, J. L. (1976). Human visual perception of structure. Master's thesis, MIT by kind permission of MIT.]

perceived visually with no knowledge of mathematics at all. Consequently Escher seems to have been adopted by mathematicians as their cultural attaché for the arts. All this must trouble the scientific realist very deeply. It appears that the human mind has evolved an ability to recognize geometrical patterns where none exist. What else might it be recognizing that does not really exist?

One of the realist's axioms of faith which we listed in Chapter 1 held our separation of natural phenomena from the perception of them to be a harmless simplification. Maybe this is not true. Our perception of Nature

8

as governed by particular geometrical factors and predictable regularities may be an illusion. Rather, such orderly trends may be the only aspects of Nature we are any good at discovering (look again at Figure 7.2, and try to see the network of twelve-sided Swiss crosses that cover it). This is not to deny, as some anti-realists might, the reality of the laws of Nature we perceive and employ, but simply leaves open the possibility that this mathematical watch-world we have mapped out may fall short of the whole. Our cross-examination of Nature may have elicited only a particular type of evidence because we have hit upon a successful line of enquiry, and by the direction of our questioning we are able to determine that part of the whole truth that comes out.

This type of subjectivity has worried modern philosophers and historians of science. There is a traditional way of writing the history of science that has died out amongst the historians, but it can still be found in the writings of many scientists interested in the history of their subject. It is also the perception of the ordinary person in the street as to what history is about. Indeed, it is how we were taught it at school: dates, people, events; how we arrived at the present. In this story of the rise of science we might trace the forerunners of the 'right' answer we know today. Who can we find in the history books who thought that the Earth went round the sun before Copernicus wrote about it? These individuals from cultures widely separated in space and time we link together as forerunners of the heliocentric view. We draw an imaginary thread through the ages to chart the course that we judge to be the 'correct' one. All wrong views are ignored. This approach was dubbed the 'Whig' theory of history by Herbert Butterfield. The name derived from those past historians who treated history as a record of events that culminated in the political system dear to their own hearts: the liberal democracy. Scientists' pictures of their subject's history usually suffer from this delusion. It is rather like taking their realist approach to Nature out of its useful context and applying it to the course of science, believing that there is one true history which is either right or wrong. There is a true record of events, of course, but only one that includes all of them. Erroneous ideas and incorrect measurements also played a role in the unpredictable course of scientific discovery.

The professional historian takes a more anti-realist view. He wants to concentrate on how science is done, and this has led to the operationalist or instrumentalist (or even structuralist) view becoming very popular. A seductive view, usually associated with the name of Thomas Kuhn, lays stress upon the fact that science is a *human* activity, and hence tries to develop a sort of sociology of science based on the activity of scientists. Kuhn is interested in the human bias towards particular scientific ideas, not on the scale of a particular concrete example as in our pictures above, but

within an entire area of science. He believes that for most of the time scientists beaver away in a routine manner measuring and calibrating things, calculating details, filling in the gaps of knowledge. Gradually a critical state is set up as fundamental problems are identified which cannot be resolved within the existing picture. But then something revolutionary comes about. When the climate is right, a new idea or 'paradigm' emerges, perhaps because an old idea is indubitably disproved, or because someone comes up with a brand-new idea with far-reaching consequences. The direction of that subject area then undergoes a period of dramatic reorientation in which the new paradigm becomes the focal point for speculation, but this period is then superseded by a return to normal activity. Normal activity is distinguished by what Kuhn calls 'puzzle solving'. This term is judiciously chosen, for puzzles are problems that have assured solutions. Likewise, the activities of normal science are imagined to concentrate upon issues which must admit of solutions within the framework of ideas and puzzle-solving methods which define the current paradigm. Occasionally, difficulties will be encountered, and the paradigm will only be able to encompass certain 'anomalous' results if inconsistencies are ignored by being labelled 'problems' or 'paradoxes', or if *ad hoc* methods of analysis are employed. Eventually, the number of such unnatural acts will become more than the paradigm can bear, and, so Kuhn claims, a revolution will ensue in which a new paradigm will arise that is able to accommodate both the successes and the difficulties of the former. The paradigm is dead; long live the paradigm.

Kuhn's vocabulary, with its paradigms and puzzles, has been absorbed into the lingua franca of the scientist. Yet most scientists would instinctively repudiate the assumptions that lie at the heart of Kuhn's analysis. For Kuhn, science is neither right nor wrong. Problems are not resolved: merely dissolved. The laws of Nature as perceived by scientists are neither true nor false. Paradigmata are transient fashions, rather like artistic styles. According to Kuhn, every scientist views the Universe through the rose-tinted spectacles of some paradigm or other. Needless to say this view is anathema to the disciple of Karl Popper, whose realist base leads him to maintain that there are definite statements that can be universally agreed upon by different scientists, because they have recourse to the experimental test of falsification against the facts of reality.

There are fashions in the philosophy of science as well, and Kuhn's picture is just one of them. Eventually, if Kuhn himself is to be believed, his picture of the progress of science will become encumbered by contradictions and ambiguities, and it will be replaced by a new and better one. Indeed, this process has already begun. The popular view of the Kuhnian theory is invariably that painted in the first edition of his famous work *The Structure of Scientific Revolutions*, published in 1962. But this work was

subjected to considerable criticism because of the vagueness and inconsistency of Kuhn's terminology (one critic, who Kuhn himself cited as being particularly cogent, identified twenty-one distinct usages of the term 'paradigm'!), and subsequent editions of the book were considerably revised. Finally, in 1974 Kuhn retreated from his former view of paradigms as underlying world-views driving the course of science through revolutionary discontinuities. Rather than aim to describe the sociology of science in terms of some global dynamic, Kuhn turned to the micro-sociological view of paradigms dictating the direction of research inside small but influential research groups. This view is less striking and considerably less controversial than its predecessor. None the less, it is still a philosophy that is ultimately self-refuting.

The crux of the Kuhnian view is that the transient paradigms which are periodically adopted and discarded are neither right nor wrong; they are just tools which have a temporary expediency determined by the existing climate of opinion. It is also an opinion concerning what scientists do. It takes the view that scientists do not discover anything that is really true. All our existing scientific viewpoints will eventually spawn anomalies, and will therefore be replaced by a new viewpoint, or paradigm, until it too is found wanting. Furthermore, the decision as to which of two competing theories is to be adopted is not taken solely on objective grounds. Additional subjective criteria like scope, simplicity, or symmetry are taken into account. Supporters of different paradigms may have different ideas as to what would constitute a decisive test of which one is superior.

There has been much learned argument as to whether this picture of normal activity assuming a particular world model, followed by mounting crisis resulting in a 'scientific revolution' and the adoption of a new set of assumptions, and so on, *ad nauseam*, is adequate. It attributes to science a sort of mob psychology that is really extremely vague. The reason why it is possible to squeeze the practice of science into its confines is probably because the same can be done with just about any activity: organized crime in Chicago, styles in high-jumping, terrorism, football tactics, car design, Paris fashions: you name it and Kuhn's theory applies. By seeking to treat the history of science like the history of art—as the story of the coming and going of styles—it seems to ignore the existence of an underlying core of fact, and its role in dictating the attitude of scientists and the direction of their interests. It is most often a change in the storehouse of facts that leads to a fundamental change within one area of science. It is also debatable whether there have really ever been any scientific revolutions of the Kuhnian variety. New theories usually contain the old ones within them as particular examples of a phenomenon that is more general than previously suspected. The march of progress resembles the redrafting and modification of a story rather than its succession by some stylistic revolution, like the making of it into a movie.

If Kuhn is correct, then it is largely irrelevant whether or not there exist laws of Nature, and what forms they take if they do, for science is an entirely human activity that cannot find them out. Kuhnian science is the scientist looking in a partially reflecting mirror. Whereas Popper would be willing to concede that we will almost certainly never discover *the* laws of Nature, because they are buried so deep in reality, none the less, unique and universal laws do exist. Kuhn, by contrast, regards laws of Nature as an ever-changing creation of the scientist's mind, part of the symbiotic psychological relationship between the observed and the observer. This is the most radical and general view that one could take about the subjectivity that is introduced into our study of Nature by our human intellectual tendencies: having recognized that there is a sociology of science it concludes there is nothing more to science than its sociology.

Before leaving Kuhn and his paradigms, we should mention one significant feature of his work that reveals an important aspect of modern physical science. Kuhn was a historian who had studied the Copernican revolution in great detail. This was a true revolution in the sense that the Ptolemaic idea of the solar system and its motions was overthrown and ultimately replaced by the heliocentric Copernican picture. Kuhn was perhaps misled into thinking that all scientific revolutions were of this 'overthrow' variety. Nothing could be further from the truth. The development of physics follows a different pattern. When Einstein's theory of gravitation improves upon Newton's theory it does not sweep it into the dustbin and replace it. Newton's theory of gravity will be used as long as there are scientists on this planet. We find that Newton's theory is a limiting case (when speeds are much less than that of light and gravity is weak) of Einstein's theory. Similarly, Newton's mechanics can be recovered from quantum mechanics in a suitable limit. Most recently, we have found likewise that Einstein's theory of gravity arises as a limiting (low energy, large distance) case of superstring theory. Thus the pattern of modern scientific revolutions is quite different from the Copernican model that impressed Kuhn. They are subsuming revolutions not overthrowing revolutions. This pattern is an important one for non-scientists to appreciate. A failure to do so leads to ridiculous applications of post-modernism to the scientific process, treating it merely as a sequence of opinions and points of view, like forms of literary criticism. This is a temptation that should be resisted.

10

The phantoms of the laboratory

> Science cannot solve the ultimate mystery of Nature. And it is because in the last analysis we ourselves are part of the mystery we are trying to solve.
>
> Max Planck

When Francis Bacon abandoned the deductive logical reasoning of Aristotle and his medieval disciples, he acted out of a conviction that our knowledge must be founded upon the things we learn from Nature rather than the phantoms and prejudices we find nurtured within our minds by the philosophical systems of the past. He warned against four specific 'phantoms' which he believed adversely bias our thinking about Nature, and divert our search for the true laws of her operation down culs-de-sac and blind alleys. They are influences that stand between us and the raw truth about the world we seek to observe and understand.

First, Bacon warns us that we are inhabited by the *Phantoms of the Tribe*: those tendencies which are by-products of our human nature. Over the centuries we like to think that we have minimized these influences which had once made it so natural to believe that Man was the centre of the physical Universe and the focal point of all Nature's workings, and that everything else was specially designed for our convenience and benefit. This prejudice is not easy to overcome, as evolutionary biologists in the United States have recently discovered. In the last section we saw how our perception of visual patterns has been over-sensitized by millennia of natural selection. Phantoms of the Tribe need not be confined to 'tribal' prejudices; they can also include the necessary physiological properties which allow the survival of psychological prejudices.

Next, Bacon points to those personal prejudices each of us possesses: he calls them the *Phantoms of the Cave*. These are still with us. Some scientists just do not like certain lines of development in their disciplines, and even cease to contribute to their subject in the hope that certain unsavoury new ideas will pass out of fashion. Sometimes this individual bias will have a sound basis—the new research direction may simply involve the use of advanced mathematical or experimental techniques that the scientist may not have the experience or expertise to cope with; but often it has a much more human origin—the new development may have been initiated by an individual who our scientist just cannot abide, and the thought of working on that person's ideas is more than he can stomach!

Bacon calls the third source of distortion *Phantoms of the Market-Place*, and they are rather more subtle than the others. Perhaps they should now be called the *Phantoms of the University*. They emerge from the association

of scientists with each other, and we witness them as the collective result of using a particular language, sharing particular concepts, and the common use of mathematics. In particular, this Phantom might animate the contemporary consensus of philosophers concerning the meaning and method of science. Kuhn seems to regard it as an overwhelming and unavoidable bias; paradigms are rather close to being Phantoms. Moreover, the later Kuhnian association of paradigms as arising through the influence of microcosms of scientists working within influential research groups also fits into this category of Phantom.

Last, we encounter what Bacon perceived to be the most pernicious influence upon our ability to think objectively: the *Phantoms of the Theatre*. The grand philosophical systems of the past and present can be seen as plays, in the sense that they create a world within the world: a world with its own scenery and characters. These philosophical models of the real world are, Bacon warns, just models. If we forget their divorce from the reality they seek to explain then we commit the error of a theatre-goer who confuses the scene on the stage before him with the real life it seeks to represent. The play may well be based upon reality; it may even be richer and bolder than reality, but it is a man-made world none the less. Before twentieth-century science began its dramatic rise, philosophical views were more explicit and important in the work of scientists. Today they are no less evident to the careful observer, but they are implicit or even deliberately obscured. Thus, for example, the belief that a unified Theory of Everything will explain the structure of the Universe uniquely and completely will appear unashamedly in scientific papers, but it is essentially a religious or metaphysical view, in the sense that it rests only upon an unstated axiom of faith.

What Bacon first so acutely recognized about science is that, by virtue of it being a human activity, its results are necessarily subject to human biases. Nevertheless, many subsequent generations of experimenters do not seem to have taken his warnings to mean very much more than 'Beware of Aristotle and the Scholastics'.

Errors

> It ain't what you don't know, that counts, it's what you know that ain't so.
>
> Will Rogers

Men like Copernicus and Newton did not have a clear concept of what we would call 'scientific error'. To the experimental scientist the term 'error' has a wider meaning than the word conveys to the ordinary person in the

street, for whom it means simply a 'mistake'—a blunder like misreading a thermometer, mixing the wrong chemicals, or allowing a nuclear reactor to overheat. But this is not the primary meaning of 'scientific error'. To the scientist the term means two other things.

The first is straightforward: the limiting accuracy to which a quantity can be measured. A simple example of this *experimental error* is illustrated by a statement of somebody's age. If we know only the year in which they were born, then we can only state their present age to lie somewhere within a *range* of twelve months. If we are told that a piece of wood is 20 feet long to the nearest foot we know only that its length lies between 19 feet 6 inches and 20 feet 6 inches. This type of 'error' specifies the limitations in accuracy of our measuring devices or the data we are given: it tells us the biggest mistake we could make even if we have carried out our measurements competently, and not done something stupid like misread the ruler. This type of error is not terribly interesting, but it is obviously one of the goals of experimental scientists to make it as small as possible. No observation or experimental measurement is of any use unless the possible measurement error is also given. To be told that the Conservative Party is five percentage points ahead of Labour in an opinion poll is a meaningless statistic unless the uncertainty, or experimental error, in the poll is also quoted.

We have already seen that the Heisenberg Uncertainty Principle of quantum mechanics ensures that there are inevitable errors associated with the measurement of all quantities even if the measuring instruments are perfect (and in practice, of course, they never are). This strange limitation arises because the very process of observation is inseparable from the state being measured. Perfect knowledge of the Universe is impossible because the act of knowing influences the Universe in an unknowable way. It is as if, by the time we record its state, it has changed slightly. Therefore every possible observation of the physical world must possess some finite measurement error. However, in practice, the measurement errors that limit the accuracy of scientific measurements, even in elementary particle physics, are considerably larger than the irreducible minimum imposed by the quantum theory. Only in the study of quantum liquids at temperatures close to absolute zero does experimental accuracy approach Heisenberg's limit.

The existence of practical limits to our measurement accuracy was not readily appreciated in the past. Copernicus knew that if the planets traced elliptical paths as they orbited the sun, then the observation of two points of a planet's orbit is sufficient to fix uniquely the particular elliptical path that is taken. And so he could not understand why the pairs of points he observed in each planet's orbit did not fix the actual elliptical path taken by the planet in the future. The reason for the disagreement was that the measurements made of the positions were not precisely accurate. Several

slightly different ellipses are compatible with the range within which the real positions could have lain. Likewise, even Newton was puzzled as to why the measured motions of the celestial bodies did not quite fit his mathematical predictions: the answer was, again, that the measurements contained small errors due to the limiting accuracy of the instruments used to record them, but this idea seems to have been quite foreign to these early scientists.

The second form of error that besets experimental science is more subtle in its origins, and more serious in its consequences because one can never be certain that it has even been identified, let alone minimized or eradicated. This species of error we call a '*selection effect*' or '*systematic error*', and its identification requires careful thought and a wide-ranging appreciation of the phenomenon under study. Its existence refutes the naïve idea that all that is required to test theories of science is the experimental method. Experimental arrangements and observational procedures have in-built propensities to gather certain types of facts more readily than others. In order to be sure that you are observing what you think you are observing, it is always necessary to have some theoretical understanding of the wider spectrum of phenomena that could be biasing your observations.

Imagine that you are embarked upon a project to learn all you can about the sizes of rats. Accordingly, you design lots of little rat-traps to catch specimens. Each trap is a small box four inches long, and has a cross-section two inches square at the ends. At one end there is a door, hinged at the top, poised to drop down and shut tight as soon as any rat venturing inside touches a little pile of food at the far end of the box. This stratagem is very successful. After a few weeks of baiting dozens of these rat-traps you find you have collected hundreds of live rats. You weigh them; you measure them; you observe them closely. But other pressing matters intervene. You must leave for an important conference in Tahiti, and are unable to finish the analysis of your measurements. Not having enough time to explain how the experiment was carried out you simply leave the list of the animals' sizes on an assistant's desk and ask him to do a preliminary analysis for you. When you return he is keen to see you, and claims to have discovered something very important about rats. He reports there to be overwhelming evidence for a definite maximum size of rat: rats of up to four inches in length were found with equal frequencies, and not a single rat was found to be greater than two inches in height or width.

A little thought convinces us that such a claim tells us nothing about rats. Our experiment was subject to a trivial example of a 'selection effect': namely, that no rats greater than two inches in height or width could get into the traps. The non-observation of any three-inch-fat rats is telling you something about your experiment, but nothing about rats. The overall

12

picture of the range of rat sizes has been biased by the inability of the experiment to collect large rats.

In practice, such experimental biases are usually a little more subtle than this, and may even be unavoidable. For instance, an astronomer might be interested in discovering how the population of stars or galaxies is distributed according to their brightness. Any results that are obtained will be biased towards finding a disproportionate fraction of the brighter objects because they are easier to see. The nasty thing about such biases is that, no matter how careful you are, you can never be sure that they have all been completely eliminated. Indeed, the art of being a good experimentalist is to a large extent the art of sensing and eliminating sources of systematic error. In terrestrial scientific experiments or data collection these biases are usually determined by the particular apparatus and method used to measure something, and so if the measurement can be made by another independent experiment, preferably using a different method, and the same result is obtained, then the presence of significant systematic error is very unlikely. This is the main reason why scientists tend to be sceptical of dramatic experimental discoveries until they have been confirmed by other different experiments: they have no independent fix on the magnitude of possible systematic biases.

On occasion, the bias of an experiment may be coupled to the inevitability of measurement error. A famous example, again from astronomy, arose during the eighteenth and nineteenth centuries. Friedrich Bessel, a notable mathematician and astronomer who introduced one of the most useful equations of applied mathematics, and still had time to discover the star Sirius B, was a colleague and associate of the great (some would say the greatest) mathematician, Johann Karl Friedrich Gauss. Gauss was one of the first to understand and work out a mathematical treatment of measurement errors. This awareness he conveyed to Bessel, and it enabled him to resolve an awkward astronomical dilemma.

Towards the end of the nineteenth century, astronomers at the Royal Greenwich Observatory had found small but irritating systematic differences in their observations of star motions. Observers differed concerning the times they recorded for stars to pass between two lines drawn on the telescope's field of view. The Astronomer Royal of the day, Neville Maskelyne, had assumed that the intolerable differences between the transit times he measured and those recorded by his assistant were simply the result of incompetence on the part of his assistant. Bessel realized what the source of the discrepancies in these measurements really was, and more important—how to rectify it. The reaction speed of different astronomers was different. Some would always signal the star crossing the first reference line a fraction of a second earlier than others. Bessel calibrated each

astronomer with a so-called 'Personal Equation' to compensate for their individual biases in starting and stopping the clock when a star entered or left the cross-wires spanning the telescope's field of view. This procedure compensates for the problem of 'selection effects' introduced by individual idiosyncracies, and was used until the introduction of modern techniques of automatic electronic measurement.

The 'Groucho Marx Effect'

> I wouldn't want to belong to any club that would accept me as a member.
>
> Groucho Marx

Selection effects do not only arise in the experimental sciences: they also influence our theoretical and mathematical investigations. They are especially rife in theoretical physics. The situation in the fundamental areas of physics which seek to further our understanding of gravitation, quantum theory, and the micro-world of elementary particles is typically like this: we have an elegant system of non-linear mathematical equations in many quantities that have been derived by the application of some powerful invariance principle, and which encapsulate the essence of what is already known about the aspect of Nature under study. However, the outstanding problem is how one *solves* these equations, and so be in a position to work out what our grandiose theory has to say about unknown situations so that these predictions can be checked against what is seen to occur in those circumstances. For example, Einstein's theory of general relativity is a theory of gravitation which is equivalent to ten complicated partial differential equations that must be solved simultaneously to find out how the geometry of space and time responds to the presence of mass and energy within it, and how that geometry dictates where masses will move. In order to test the correctness of the equations of the theory, their solutions need to be found. We can find a particular solution that describes the gravitational field exerted upon the planets by the Sun if we idealize the Sun to be a sphere. It predicts that the motion of the planets should proceed in slightly different orbits than are predicted by Newton's theory of gravity. The difference is most marked for the motion of the planet Mercury, the closest to the Sun, for which the influence of the Sun's gravity is greatest. Observations confirm the existence of the small new effect predicted by Einstein's theory, and show the Sun's shape to be close enough to a perfect sphere for this idealization not to affect the conclusion significantly.

This procedure is clearly not perfect. The equations of physical theories like general relativity are too complicated to be completely solved. Hence,

13

one must resort to significant idealizations and approximations in order to learn something about the features that are latent within the theory, and which could be present in the complete solution. A typical strategy is to try to solve the equations in a special situation; for instance, as with the idealized spherical Sun, when there is a simplifying symmetry. We might, for instance, find a solution that gives the gravitational field exerted by a *spherical* object, or by one that is not changing its shape with time. Idealizations of this sort forbid certain types of variation to arise in the equations, and simplify them dramatically with the result that they are usually soluble. In the case of Einstein's theory of general relativity the first solutions of the equations to be found were for unchanging spherical configurations of material. Since Einstein's equations were first written down in 1915 there has been a continual search for solutions to them. A few years ago four mathematicians co-authored a four-hundred page book devoted to displaying, classifying, and organizing the most important solutions that have been found to date. All the solutions in this book possess special simplifying properties. This is the reason why it has been possible for us to find them.

This is where the 'Groucho Marx Effect' comes in: the only solutions of the equations that we are clever enough to find always describe special idealized situations that will not generally arise in practice. The limited computational competence of human beings, and the inability of computers to do more than humans do (except to do it faster) means that the conclusions we draw from our theories of physics are conditioned to a considerable extent by the content of a relatively small number of exactly soluble examples. We can find solutions of Einstein's equations that describe exactly how a universe that is perfectly uniform expands in time. The real Universe is *almost* uniform over its largest dimensions—but not quite. There exist stars and planets, galaxies and clusters; these non-uniformities are, of course, essential for our own existence (there cannot be any observers in a perfectly uniform universe!). The question of how they arose, and from what initial state, is what most of modern cosmology is about. Unfortunately we know of no exact solutions of Einstein's equations that describe an expanding universe filled with a higgledy-piggledy collection of stars and galaxies. There undoubtedly do exist such realistic solutions, but they are mathematically too complicated for us to find. We rely heavily upon the real world in all its complexity being close to simple idealized situations. Often this is not the case. The horribly complicated turbulence that results at the base of a waterfall is so dissimilar to any smooth and idealized fluid flow that we have very little detailed understanding of it at all.

One way in which we can evade the need for idealization is by the use of approximations. Besides finding the exact solutions of equations for idealized situations one can also discover the approximate solution of the

equations for situations that are *almost* ideal. Although the Universe does not expand at exactly the same rate in every direction, it almost does (to within one part in ten thousand, in fact), and so we believe that it will be approximately described by the idealized solution to about the same accuracy. In this case our expectation is borne out by the good agreement between our observations of the Universe and the predictions of the idealized model. However, it is quite possible that there exist idealized solutions that do not lie close, in any sense, to non-idealized ones. Such isolated examples would be atypical and quite misleading with regard to what the full solution of the equations of the theory are like. A good example is the famous special solution to Einstein's equations found by the logician Kurt Gödel. This solution showed that there is a particular solution of Einstein's equations which allows time-travel to occur. Gödel's solution describes a weird rotating universe which looks nothing like the one we live in, but this does not mean that we can stop worrying about time-travel. We need to know whether time-travel is a property of the full and realistic solutions of Einstein's equations—that would describe our own world—or whether it is a pathology of a small number of physically irrelevant solutions with weird properties.

This highlights another important point. Einstein's equations allow innumerable different solutions each describing different expanding universes. They have different initial properties: some possess galaxies while others do not. But there is, by definition, actually only one Universe. What is the selection principle that pins down the precise solution that describes *our* observed Universe, and it alone? This principle must come from outside the theory of general relativity.* The need for this 'selection principle' shows that Einstein's theory is not the best possible description of the Universe. It permits too many things that are not realized in Nature in addition to the things that are.

Many physicists have expressed the belief that when all the disparate theories of the different elements of Nature are unified in an all-encompassing 'unified field theory' then the constraints upon the shapes of the individual pieces in order that they can be fitted together will be so stringent that there will be only one possible unified picture. There may be one and only one possible theory of all fundamental phenomena, in which everything that is not forbidden will be compulsory. Steven Weinberg leans

* In fact, *any* space–time geometry is a solution of Einstein's field equations for *some* distribution of matter, just as any gravitational potential will solve Poisson's field equation for some density distribution. These equations only place restrictions upon Nature because most geometries and potentials are ruled out by the requirement that their associated matter distribution be physically realistic in various ways—for example by having the density of matter positive everywhere.

towards this hope, and once speculated that 'when you put quantum mechanics together with relativity, you find that it is nearly impossible to conceive of any possible physical system at all. Nature somehow manages to be both relativistic and quantum-mechanical; but these two requirements restrict it so much that it has only a limited choice of how to be—hopefully a very limited choice.'

If we are idealists then perhaps the future 'discovery' of such a theory is inevitable, but if we are realists then surely we can hold no such hope. For even if such a description of the ultimate workings of Nature does exist, who is to say it is within the grasp of human minds to find it? Indeed, we saw in the last chapter that it is possible to entertain the view that there is no such ultimate theory of everything at all. There is no reason, save our grandest presumption, that Nature should be fashioned with our computational incompetence in mind. And the final irony is that, even if such a super-theory does exist and we are able to write it down, we can never know that it is correct. The scientific method does not enable us to demonstrate that our theories are true: only that they are false. While we can come up with candidates for the ultimate theory that would be falsifiable, *the* ultimate theory would not be falsifiable at all. In the meantime, any theory that does not provide a unified description of everything must eventually be proven to conflict with some aspect of experience.

We must also be aware that while there may exist a unique *Theory* of Everything it may have an infinite number of cosmological *solutions*, and the actual Universe is only described by one of them when particular starting conditions are chosen. More problematic still is the fact that broken symmetry is ubiquitous in Nature. The key features of the Universe may owe their origin to random breakings of the underlying symmetry of the Theory of Everything.

The search for a completely determined set of natural laws does not proceed by logic and observation alone. Metaphysical and aesthetic criteria are called upon to guide theoretical speculation. But what are the right aesthetic criteria: beauty, harmony, symmetry, lack of symmetry, simplicity, computability, finiteness, brevity, minimal assumptions: who can say?

Beauty

> No doubt aardvarks think that their offspring are beautiful too.
>
> John Ellis

What criteria do scientists use to steer them towards successful and powerful theories of Nature? We have already seen that in practice such theories

are necessarily mathematical in form. There is good mathematics and bad mathematics, ugly mathematics and beautiful mathematics. Difficult as it is to draw the precise line between these opposites, the professional mathematician finds it no harder to draw than the difference between night and day. The theoretical physicist Paul Dirac was one of the most outspoken supporters of the idea that the deployment of 'beautiful' mathematics—that which possesses symmetry, economy of form, a depth of interconnection with other parts of mathematics, and the maximum of structure from the barest of inputs—should be the priority of the theoretical physicist seeking a description of some physical phenomenon. This one should seek in conjunction with a desire for 'simplicity' that avoids superfluous ideas and hypotheses. Of the search for mathematical laws of Nature Dirac had this to say:

> The dominating idea in this application of mathematics to physics is that the equations representing the laws of motion *should be of a simple form*. The whole success of the scheme is due to the fact that equations of simple form do seem to work. . . . The method is much restricted, however, since the *principle of simplicity* applies only to fundamental laws of motion, not to natural phenomena in general . . . What makes the theory of relativity so acceptable to physicists in spite of its going against the principle of simplicity is its great *mathematical beauty*. This is a quality which cannot be defined, any more than beauty in art can be defined, but which people who study mathematics have no difficulty in appreciating. The theory of relativity introduced mathematical beauty to an unprecedented extent into the study of Nature.
>
> We can now see that we have to change the principle of simplicity into a *principle of mathematical beauty* . . . It often happens that the requirements of simplicity and of beauty are the same, but where they clash the latter must take precedence.

Dirac's position is quite extreme, since in some circumstances he did not wish even negative experimental evidence to deflect him from a particular line of inquiry that he had pursued because of its mathematical beauty and elegance. He continues:

> If there is not complete agreement between the results of one's work and experiment, one should not allow oneself to be too discouraged, because the discrepancy may well be due to minor features that are not properly taken into account and that will get cleared up with further developments of the theory.

These two statements are very striking. They are good examples of the influence of selection effects of a particular sort. We certainly have an aesthetic sense that we cannot easily explain. Some of its features seem to

15

be universal. Many would seek to argue that it is entirely a consequence of natural selection, but it seems to possess elements that are unnecessarily sophisticated for this purpose. In *The Sense of Beauty* George Santayana suggested that we perceive 'beauty' to reside in patterns and appearances that offer sufficient novelty to arouse our curiosity, but not so much that their complexity is beyond our understanding. He points to the starry night sky as a manifestation of this tantalizing property. The eighteenth-century Dutch writer Hemsterhuis defined beauty as that which provokes the greatest number of ideas in the shortest time. Scientists have found Nature as a whole to be beautiful in these peculiar senses: it presents challenging problems that offer the possibility of solution. It invites and satisfies curiosity. When all is said and done, scientists want equations that are analysable rather than merely simple or beautiful. And we should remember that it is the purest speculation that Nature and her laws are 'beautiful' in our (or indeed any) sense. Scientific observation alone cannot eliminate the possibility that one of David Hume's 'superannuated deities' produced the Universe we see as a faulty precursor to the real thing implemented in another time and another place!

Another definition of mathematical beauty is that suggested by the Indian Nobel laureate, Suhrahmanyan Chandrasekhar. He has proposed that Einstein's theory of general relativity has an aesthetic basis primarily because of the miraculous way in which it proved compatible with other laws of Nature which played no role in its conception and formation. However, it would be more accurate to say merely that within the compass of the general relativity theory there exists a core with this unforeseen harmony. There are other parts of general relativity, like the presence of solutions allowing time-travel and naked singularities, which would create conflict with all the other laws of physics if they were realized in Nature.

Relatively simple mathematics is useful in describing Nature, not just the most abstract and difficult—although some of that is useful as well. Perhaps this feature lured gifted individuals like Paul Dirac into the subject of mathematical physics. But it is more likely that, having found mathematics and physics to be the activity which most nearly fitted the perspectives he possessed, Dirac then sought to stress those aspects of physics that were closest to his ideals. And if they are often successful in producing great discoveries then they come to dominate one's view.

In this regard it is instructive to consider the interest that mathematicians and physicists have displayed towards the spectacular pictures of fractal curves that computers have recently generated in profusion. Collections of these have been displayed in many famous art galleries. The pictures have an undeniable beauty, which is enhanced by the skilful choice of false colour-coding introduced by the computer scientists. But beyond that there

is an aspect that connects our aesthetic appreciation to that of Nature itself. The intricate structures of fractal curves, like the Mandelbrot set (p. 220), are closely related to the self-similar structures we see around us in the natural world—the branching of a leafless tree, the pattern of frosted snowflakes, the crenellated patterns of a mountain landscape—all exhibit a non-linear invariance which we find deeply appealing. It has taken us a long time to find the type of mathematical algorithms that can generate such structures systematically, but their discovery is surely important in locating the centroid of our aesthetic appreciation of visual symmetry and mathematical harmony.

We know of no reason why the laws of Nature should be either 'simple', 'beautiful', or anything else that appeals to us. Indeed, the theory that Dirac thought the ugliest and most unsatisfactory part of physics—quantum electrodynamics—is the most accurate of all the fundamental scientific theories that we possess. It is the quantum theory that describes the interaction between light and electric and magnetic forces. Its theoretical predictions are confirmed by experiment good to ten decimal places. Some physicists find this theory 'beautiful' too!

All we can conclude from this equally subjective discussion is that scientists do possess conscious and unconscious biases towards developing certain types of descriptions and laws of Nature. The more mathematical the science the more powerful will be these influences. The search for symmetries and invariances is a goal of the mathematical physicist. There is no real evidence that Nature is in any well-defined sense 'simple' or 'beautiful'. Indeed, the hallmark of most natural phenomena is a deep complexity masquerading as simplicity. One is reminded of the story of the astronomer who began a public lecture on the nature of stars with the statement 'Stars are very simple objects', only to be met with a cry from the back of the hall that 'you'd look pretty simple too from a distance of two hundred light-years!' What is more remarkable is the extent to which theories and descriptions that we know now to be seriously incomplete or just plain wrong proved to be such reliable temporary guides to a substantial fraction of the truth for so long in the past.

Dirac's remarks about the seductive character of elegant mathematics are the natural views of a realist who believes that Nature *is* intrinsically mathematical. Thus, Nature and her mathematical representation are regarded as equivalent, and the unquestioned beauty of the former can be sought in our gropings towards the latter. Yet, scientists of a different metaphysical persuasion have argued in quite a different direction to Dirac. Some operationalists, like Bridgman, actually regard the search for aesthetic mathematical structures in physical descriptions as a dangerous metaphysical diversion. With regard to general relativity, which both Dirac,

16

Chandrasekhar, and most other physicists regard as the most 'beautiful' of theories, Bridgman writes,

> The metaphysical element I feel to be active in the attitude of many cosmologists to mathematics. By metaphysical I mean the assumption of the 'existence' of validities for which there can be no operational control . . . At any rate, I should call metaphysical the conviction that the universe is run on exact mathematical principles, and its corollary that it is possible for human beings by a fortunate *tour de force* to formulate these principles. I believe that this attitude is back of the sentiment of many cosmologists towards Einstein's differential equations of generalized relativity theory—when, for example, I ask an eminent cosmologist in conversation why he does not give up the Einstein equations if they make him so much trouble, and he replies that such a thing is unthinkable, that these are the only things that we are really sure of.

Of course, Bridgman's philosophy of science regarded mathematics as simply a tool for the construction of operational definitions of physical quantities (although he had to admit that the operationalist doctrine could not cope with a subject like cosmology, where one needs to talk about quantities like 'the mass of the Universe' which cannot be defined operationally). He adopted a constructivist interpretation of mathematics which bordered upon formalism, and regarded this as the natural complement to his operationalist philosophy in physics. Proceeding from his recognition of our 'metaphysical' tendency to assume for Nature a mathematical structure, he argues that this unquestioned assumption is a dangerous human bias:

> I believe that there are dangers in any subject in which there is such an unavoidable mixture of purely 'scientific' and 'human' elements. It seems to me that there is a particular danger of introducing actual inconsistencies into the structure if the metaphysical attitude with regard to mathematics is so far adopted as to obscure the perfectly legitimate use of mathematics in attaining simplicity of formulation.

Thus, while Dirac has so much faith in the mathematical beauty and economy of Nature that he is willing to follow it as a guiding principle that might at times rule over experiment, Bridgman argues that it is our willingness to follow such a seductive Pied Piper that undermines his faith in the intrinsically mathematical character of Nature.

Einstein had a view that intertwined the two considerations of Bridgman and Dirac. Unlike Bridgman, he believed that 'the supreme task of the physicist is to arrive at those universal elementary laws from which the cosmos can be built up by pure deduction'. He realized that, even if one

adopted an operationalist stance to the creation of physical laws, they would still require a mathematical representation that could not be gleaned uniquely from experience. At this point we see that the physicist has some artistic licence in the way that he pursues the presentation and development of his theory.

> If, then, it is true that the axiomatic basis of theoretical physics cannot be extracted from experience but must be freely invented, can we ever hope to find the right way? Nay, more, has the right way any existence outside our illusions? Can we hope to be guided safely by experience at all when there exist theories (such as classical mechanics) which to a large extent do justice to experience without getting to the root of the matter? I answer without hesitation that there is, in my opinion, a right way, and that we are capable of finding it. Our experience hitherto justifies us in believing that nature is the realization of the simplest conceivable mathematical ideas. I am convinced that we can discover by means of purely mathematical constructions the concepts and the laws connecting them with each other, which furnish the key to the understanding of natural phenomena. Experience may suggest the appropriate mathematical concepts, but they most certainly cannot be deduced from it. Experience remains, of course, the sole criterion of the physical utility of a mathematical construction. But the creative principle resides in mathematics. In a certain sense, therefore, I hold it true that pure thought can grasp reality, as the Ancients dreamed.

Dirac's remarks about the secondary role of experiment relative to theory on some occasions are also worth expanding upon. What he means by them is not quite what Eddington implied by his famous warning against embracing sensational experimental results too readily, 'do not believe any experimental result until it is predicted by theory'; but rather that one may conceive a wonderful new mathematical model for the working of, say, the atomic nucleus, but find it to possess one obstinate defect which puts it in disagreement with some experiment. In this circumstance you should not necessarily lose faith in your good idea. It may be that you have omitted some comparatively minor (or even major) ingredient in your new theory, but this omission can be straightforwardly repaired in the future. This approach is much in evidence in modern theoretical research into elementary particle physics. In this area of research, speculative theory races way ahead of experiment because of the vast cost and sophistication of the necessary experiments. They require huge energies to be attained for particle bombardments, and batteries of sophisticated computers to record and interpret the results. Unaided by experimental data, a theorist may pick upon a gauge theory (of the sort we discussed in Chapter 4) defined by

17

some group of mathematical operations, and work out some of its experimental consequences. Occasionally it will be found that almost all the pieces fall into place. Neat, unified explanations appear for facts that were previously independent and *ad hoc*. But there now exists some horrible new consequence that is at variance with experiment. Faced with this state of affairs the theorists may well ignore that problem and pursue the positive aspects of the theory in the hope that the awkward discrepancy can be sorted out later by a generalization of the theory that he has written down. This is often a good strategy, because the gauge theories under study do not claim to be theories of everything in the microscopic world. They are incomplete descriptions of Nature, and the theorist always likes to believe that it is this contemporary incompleteness alone that is the source of its bad predictions. When the theory is made greater in scope its superficial pathologies may well disappear. And sometimes they do. Part of the art of being an innovative theorist is to discern what are temporary difficulties born of the simple prototype model one is creating and what are fatal diseases endemic to any theory of the sort being proposed.

Dirac cites an example drawn from Schrödinger's experience to support his argument that formal beauty should sometimes override experimental data in the evaluation of a theory's merits:

> I heard from Schrödinger of how, when he first got the idea for his equation, he immediately applied it to the behaviour of the electron in the hydrogen atom and then he got results that did not agree with the experiment. The disagreement arose because at that time it was not known that the electron has a spin. That, of course, was a great disappointment to Schrödinger, and it caused him to abandon the work for some months. Then he noticed that if he applied the theory in a more approximate way, not taking into account the refinements required by relativity, to this rough approximation his work was in agreement with observation.

It is this story that provoked Dirac to make his much-quoted declaration; he continues:

> I think there is a moral to this story, namely that it is more important to have beauty in one's equations than to have them fit experiment. If Schrödinger had been more confident in his work, he could have published it some months earlier, and he could have published a more accurate equation.

There is another reason for sometimes taking Dirac's secondary view of experiment: experiments are sometimes wrong! It is interesting to recall a couple of cases where the existing experimental evidence was contrary to

the predictions of theorists, but so strong was the belief in the elegance of the theoretical models that the adverse experimental evidence was correctly disregarded as unreliable.

When told of experimental evidence contradicting his theory of relativity Einstein's immediate reaction was that the experiments must be wrong—and he was right. Later, when the estimates of the age of the Universe obtained from the evolution of stars were found to be in disagreement with those predicted by relativistic cosmology, Einstein again stood by the predictions of his theory on the grounds that it was based upon a firmer theoretical foundation than the theory of stellar evolution required to interpret the observational data—and he was right again.

As a last example of the victory of theory over experiment (they are only notable because there are innumerable defeats) one can cite the 1958 paper of Richard Feynman and Murray Gell-Mann on the structure of the weak interaction. The theory was developed using what the two authors described as the 'predilection' of one of them for a particular type of equation. However, the results of this theory were in disagreement with the observed distribution of electron neutrinos in the decay of helium-6. But the authors were not dismayed. On the contrary, they wrote of the disagreement in their paper, that

> These theoretical arguments seem to the authors to be strong enough to suggest that the disagreement with the He-6 recoil experiments and with some other less accurate experiments indicates that these experiments are wrong.

And, indeed, they were correct in this surmise, as later experiments revealed.

There is a further interesting consequence of erroneous experimental results that has become apparent in recent years on the interface between particle physics and cosmology. The 'Big Bang' theory of the origin and evolution of the Universe indicates that the Universe must have been a hotter, denser, and more crowded place in the distant past. As we extrapolate backwards in time we encounter universal conditions of ever-increasing temperature in which particles collide with each other at higher and higher energies. In short, the entire Universe resembled a gigantic experiment in ultra high-energy particle physics during its early stages. Cosmologists and particle physicists have therefore joined forces in the study of the early history of the Universe. The particle physicist sees a new environment about which predictions can be made using the very latest theoretical ideas and speculations. If these ideas lead to predictions that the present-day Universe should possess certain bizarre properties that it evidently does not—like containing no galaxies for instance—then we can

18

discount these new ideas as false. Likewise, the cosmologist can examine the latest brain-children of the particle physicists to see if they result in the Universe containing new species of elementary particle which might resolve some of the gaps in our picture of the Universe's make-up. This close collaboration between particle physicists and cosmologists began in earnest in 1978.

Between about 1979 and 1983 a number of dramatic experimental results were announced. It was claimed that the neutrino possessed a small mass, and that different types of neutrino could transmute back and forth into each other; it was claimed that an isolated magnetic charge (the so-called 'magnetic monopole' whose existence had been proposed by Paul Dirac long before) had been detected; it was claimed that the effects of an electric charge asymmetry within the neutron had been detected, and there were claims to have seen protons decay. Each of these experiments generated an enormous amount of theoretical interest in the areas of cosmology and astrophysics they touched upon. Innumerable popular articles and books were written about them for the general public, and the number of conferences bringing together particle physicists and astrophysicists to discuss them spiralled to such an extent that it became possible to spend one's entire life either at or in transit between such gatherings! Looking back on this explosion of interest it must now be said that all of these stimulating experimental results have now either been withdrawn, or such doubt has been cast upon them that no theoreticians lean upon them at all. This disappointing situation is rather unusual, and merits some explanation. Fundamental experiments are generally performed carefully and correctly. All of the above-mentioned experiments were peculiar either in being chance observations or detections of events right at the limit of an experiment's sensitivity to resolve real measurements from the noise produced by the measuring instruments themselves. What is most notable about this collection of unfounded experimental claims is that they led to far more theoretical progress in exploring new ideas in both particle physics and cosmology than have correct experiments!

The Anthropic Principle

> There is no such person as a philosopher; no one is detached; the observer, like the observed, is in chains.
>
> E. M. Forster

The Big Bang picture of the history of the Universe is the central paradigm within which cosmologists work to understand what we do and don't know about the Universe in which we find ourselves. But what has all this got to

do with you and me? The problem of fitting human life into the impersonal tapestry of cosmic space and time has been pondered by mystics, philosophers, theologians, and scientists of all ages. The views they have come up with straddle the entire range of options. At one extreme is painted the depressing materialist picture of human life as a local accident, totally disconnected from and irrelevant to the inexorable march of the Universe into a future 'Big Crunch' of devastating heat or the eternal oblivion of the 'heat death', while at the other extreme is preached the anthropocentric teleological view that the Universe was tailor-made for human life by some form of providential design. The latter view was strongly held in many cultures, reaching its zenith amongst English scientists in the eighteenth century. It remained the view of many biologists until, in the mid-nineteenth century, Charles Darwin and Alfred Russel Wallace recognized the evolutionary adaptation of organisms to their environment by natural selection. Since that time biologists have rejected any notion of evolution as being goal-directed. There is no grand goal (Mankind?) to which the entire evolutionary process is directed. If the environment were to change in some unusual way so as to render intelligence a liability then we would cease to be well adapted to survive, and might well face the same sort of demise as did the dinosaurs.

The lesson we have drawn from the problem of 'selection effects' is that we must be aware of any in-built biases in our measuring instruments toward preferentially gathering evidence of a particular sort. If we were to observe the Universe only with the human eye then we would conclude that all the radiation in the Universe lies in that range of wavelengths which we call the 'visible waveband', spanning the spectrum from red to violet. But there exists radiation of other wavelengths that the human eye cannot detect. Light of longer wavelength has too little energy to record its reception on the rhodopsin molecules at the back of the retina, and light of much shorter wavelength than the visible is so energetic that its reception would destroy the eye. Our human physiology circumscribes the range of astronomical observations we can make unaided, and hence what the ancients could learn about astronomy.

Today, we compensate for the eye's limited observational range by building artificial 'eyes' of far greater power and scope. These take the form of traditional optical telescopes like those at Mt. Palomar, or radio telescopes at Jodrell Bank and Arecibo. These are now complemented by infra-red telescopes high in the mountains of Hawaii, and X-ray, infra-red, and ultraviolet detectors orbiting the Earth in satellites. But there is a more dramatic aspect of our human physiology that simply building better telescopes cannot overcome.

Human beings are complex biochemical computers. This is not all they

are but it is their irreducible minimum specification. They are composed of self-reproducing helical molecules of DNA (= deoxyribonucleic acid) composed of atoms of carbon, nitrogen, phosphorus, and oxygen. How such intricate molecular structures arose on the Earth is not known for certain. It is possible they were generated initially by random interactions and mutations during the Earth's early history. If large numbers of different types of complex molecules were produced at different times in a primordial soup, it is quite plausible that those able to effect copies of themselves as a result of their interactions with other molecules would rapidly come to dominate the population at the expense of the non-replicators. But what is the origin of the carbon, nitrogen, oxygen, and phosphorus that compose the DNA molecules of life? Of this we are much more certain: it lies in the stars.

The atomic elements heavier than hydrogen and helium could not have been produced during the inferno of the Big Bang. The Universe expands and cools too rapidly for the heavier nuclei to be synthesized by nuclear reactions. The natural nuclear reactor we call the Big Bang shut itself down after the Universe had been expanding for about three minutes, but in that short time it transformed 25% of the mass of the Universe into helium, leaving almost 75% as hydrogen. I say 'almost' because the nuclear fusion of hydrogen into helium leaves tiny traces of deuterium, lithium, and the isotope helium—3 in relative abundances of 1/1000%, 1/100,000,000%, and 1/1000% by mass respectively. These predicted abundances correspond exactly to the fractions measured in the Universe today. This remarkable finding is one of the corner-stones of the Big Bang cosmological theory.

We have learned that the complex phenomenon we call 'life' is built upon elements that are heavier and more complex than the hydrogen and helium which emerges from the Big Bang. Most biochemists believe that the element carbon, on which our own organic chemistry is based, is the only viable foundation from which chemical life can arise *spontaneously*. Living systems on Earth are based upon the subtle chemical properties of carbon, and its relationships with hydrogen, nitrogen, oxygen, and phosphorus. Other elements play important roles but these five are the leading actors in the game of life. In order to create these five building blocks of life (and also silicon, if we foresee a future for the non-spontaneous evolution of 'life' evolved from current silicon technology), the simple nuclei made in the Big Bang must be cooked at high temperatures for billions of years. The furnaces that Nature has provided are the interiors of the stars. There, the hydrogen and helium surviving the Big Bang is slowly burnt into the heavier elements necessary for you and me. When stars have exhausted their nuclear fuel resources they implode at the centre, and expel their outer layers into

space. These dramatic death throes, which we witness as supernovae, serve to disperse the biological elements through space where they become incorporated into planets, asteroids, and other forms of interstellar debris. Ultimately, they find their way into our bodies. We are the ashes of the stars.

The most important fact about this stellar alchemy upon which life hinges is the length of time it all takes. At least ten billion years of stellar burning are required to produce essential elements like carbon. It is this simple fact that renders our study of the Universe and its properties the victim of an all-embracing selection effect: our own existence. As an example, let us take the question of the *size* of the visible Universe.

The present speed of the Universe's expansion and its deceleration rate indicate that the expansion has been occurring for a time somewhere between 13 and 18 billion years. This state of expansion means that the size of the Universe is inextricably entwined with its age. The reason that the Visible Universe is more than 13 billion light-years in size today is that it is more than 13 billion years old. A Universe that contained just one galaxy like our own Milky Way, with its 100 billion stars, each perhaps surrounded by planetary systems, might seem a reasonable economy if one were in the universal construction business. But such a universe, with more than a 100 billion fewer galaxies than our own, could have expanded for little more than a few months. It could have produced neither stars nor biological elements. It could contain no astronomers. We should not be surprised to discover that the Universe is so vast in scale, because we could not exist in one that was significantly smaller. This realization, that some of the key structural features of the Universe may be necessary prerequisites for the existence of observers, must influence our view of many issues. Many a philosopher has argued against the ultimate significance of human life on the grounds that it occupies such an minuscule fraction of the known universe. Some modern astronomers see the vastness of the Universe as persuasive testimony to the overwhelming probability that the Galaxy is teeming with other intelligent life-forms with whom we might communicate. But the Universe needs to be as big as it is to support just one solitary outpost of life. It is a sobering thought that the global and possibly infinite structure of the Universe is so linked to the conditions necessary for the evolution of life on a planet like Earth.

This recognition, that there are types of universe which we could not expect to observe, is often called the *Weak Anthropic Principle*. It is at root an extension of our caution in requiring a full understanding of the in-built biases present in our measuring apparati when doing experimental science. It tells us that our astonishment at many properties of the Universe which appear unusual a priori must be tempered by the recognition that many of

20

them simply must be present if a universe is to be studied by intelligent observers.

Cosmologists view the Weak Anthropic Principle as a qualification of the famous stricture of Copernicus, who by announcing that the Sun, and not the Earth, was at the centre of the solar system (which constituted the entire Universe as far as pre-nineteenth-century astronomy was concerned), removed the prejudice of centuries that humanity lay at the centre of the physical Universe. We should be careful not to confuse Copernicus's important lesson that we must not regard our position in the Universe as special in *every* way with the spurious belief that our position in the Universe cannot therefore be special in *any* way. We could not exist within a star; we could not exist when the Universe was less than a million years old. If the Universe did happen to possess a centre (there is no evidence that it does), and conditions were only conducive to the evolution and continued existence of life near that centre, then we should not be surprised to find ourselves living there.

It seems that this symbiotic relationship between the Universe and observers of it has other more mysterious features that are impressive but hard to evaluate objectively. The great success of Einstein's general relativity theory in describing the past and present of our Universe, with its highly regular expansion, low density, and patchwork of stars and galaxies, has provoked cosmologists to study the other types of universe that Einstein's equations can describe. The more we examine the other types of universe that the laws of physics appear to allow, so the more special and unusual do the properties of the actual Universe appear to be. Its uniqueness is impressed upon us most forcefully by the fact that we can seemingly conceive of so many alternatives. Whether the preponderance of theoretical possibilities in the face of the fact of the uniqueness of the Universe is telling us something about the initial conditions allowed for universes or the laws we think govern their evolution was something we discussed in Chapter 4.

For illustration we can pick on just a few unexplained large-scale properties of the Universe. We have seen why we must find it to be so large and so old. And we can see why it is unlikely that we would be around when it is more than ten times older and larger: by this epoch the stars will all have died; the resources of our planet will be unable to support us. It may well be that we will have become resourceful enough to continue living elsewhere in another way, but our continued existence will be more improbable than it is today when the Sun is shining in the prime of its nuclear life. But what of the galaxies? We don't know for certain how galaxies form, but we know a basic physical process that will lead to the development of structures like galaxies. If a collection of particles exert attractive forces upon one another

then, unless they are distributed in a *perfectly* uniform manner, they will tend to become clumped and non-uniformly distributed as time passes. In the context of the Universe the particles are the atoms and grains of matter emerging from the Big Bang, and the attractive force is gravity. The result is a process of gravitational aggregation whereby the denser regions of space get denser at the expense of the sparser ones. By this inevitable process a Universe that is not perfectly smooth—and none could ever be so in view of the inevitable quantum ambiguity in the positions of elementary particles within it—will, over a period of billions of years, pass from being almost smooth and unstructured into a state in which matter is aggregated into dense islands scattered throughout a sea of lower density gas and dust. Only in these dense islands, which we identify with galaxies, can material attain the densities necessary to form stars. This much we understand, although not the later stages of how and why some dense islands of cosmic material wind up to resemble spiral galaxies whilst others end up as the giant ellipsoidal balls of orbiting stars we call elliptical galaxies. Yet although we know how early stages of the process of aggregation develop with the passage of time there is one important piece of the picture missing. How large was the irregularity at the beginning? Only when we know the answer to this question can we tell whether our explanation for the existence of galaxies is the correct one. One thing we can say about this starting value is that it must be very specially tuned if galaxies are to form in time. A small increase above the optimum value and the dense islands form too early, and collapse catastrophically under the pull of gravity to form giant black holes before stars can ever shine. A small decrease below the optimum value results in islands of denser material that are too feeble ever to condense into galaxies: stars never form. In either case there appears to be little chance of life evolving spontaneously. Likewise, the relative concentrations of matter and radiation emerging from the Big Bang are characterized by a particular value that inhabits a small niche that allows life as we know it to evolve. One of the goals of the inflationary universe theory is to provide an explanation of these small irregularities. The theory has the potential to predict the level of non-uniformity expected in the Universe. The simplest version of the inflationary universe makes particular predictions about what we should see in the Universe, if the irregularities that inflation seeded were of a scale and level that is sufficient to create the galaxies and clusters of galaxies that we see in the Universe today. These irregularities will slowly amplify as the Universe ages and leave tell-tale patterns in the background radiation left over from the early stages of the Universe. Satellites have been able to detect these fluctuations in the temperature of the radiation as we receive it from separate directions. The differences are just two parts in one hundred thousand when we look in directions on the sky separated by more

21

than ten degrees. Moreover, these differences do not change from this value as we compare temperatures over larger and larger separations, up to 180 degrees. This agrees with the predictions of the simplest inflationary theory. However, if we could examine the differences between the temperature in directions that are separated by angles 10 to 100 times smaller then we predict that very distinctive peaks in the temperature variations should be seen. Already experiments based at remote sites on the Earth's surface have found tentative evidence for variations like this. But the Earth's atmosphere spoils the signal, and earth-based observatories can only see a small portion of the sky. New satellites, MAP from NASA and Planck Surveyor from the European Space Agency, scheduled for launch in 2000 and 2007, will map these smaller-scale variations in exquisite detail. If all goes to plan we will know whether or not our visible universe underwent inflation far in the past and whether that process gave rise to the irregularities that subsequently matured into the spectacle of stars and galaxies we see today.

Even if inflation can provide an answer to the problem of the inhomogeneity level of the Universe, it still requires an application of the Anthropic Principle. We recall that the idea of inflation is that each microscopic region of the Universe can, in its earliest stages, undergo an accelerated phase of expansion. Thus, the entire visible universe today may reflect the conditions that existed within a single prehistoric quantum fluctuation. However, the early universe possesses many (an infinite number if the universe is infinite) microscopic regions at the time when inflation can occur. Each will inflate by an amount determined by the local conditions within it. The result will be a universe resembling a foam of bubbles of all sizes. Some regions will inflate a lot, some only a little. There is no way in which the usual scientific method of prediction or falsification applies here. If we live in one of these bubbles it is a historical question, and history is the science of things that are not repeated. If inflation happened then we must find ourselves inhabiting one of the inflated bubbles which underwent at least ten or fifteen billion light-years of inflation. Beyond the horizon of our visible Universe there will exist other inflated bubbles with unpredictable sizes.

As if this complexification of cosmic geography is not enough, Alex Vilenkin and Andrei Linde have shown that an analogous complexification is inevitable for cosmic history. In general, if a bubble inflates then it will create within itself and its progeny the conditions for further inflation of those bubbles to take place. This process appears to be self-perpetuating and never ending. We do not know whether it needs to have a beginning or not. This 'eternal' inflationary universe gives us a new perspective on our place in the cosmic scheme of things. We may be the inflated image of a single patch of an infinite universe that happened to get big enough for

stars to form and carbon-based life to evolve on solid planets. Such events may occur from time to time in the infinite history of the Universe. All we can hope to do is to evaluate the probability of it having happened where we live. As cosmologists have explored these spectacular cosmic scenarios in further detail, exploring the theories that predict them for all their other consequences, they have found more remarkable possibilities. The different regions that can inflate by different amounts in different parts of the Universe can differ in more profound ways than by size or level of regularity. The constants of physics can fall out differently in different regions; even the number of dimensions of space that grow large can vary from region to region, and the number of separate forces of Nature can change as we move from one region to another. Only in some of these regions, will things coincide so that life can arise. Only some of these inflated domains are inhabitable by conscious observers. Inevitably, we find ourselves observing one of those domains where the dice fell out right.

Coincidences

> Although we talk so much about coincidence we do not really believe in it. In our heart of hearts we think better of the universe, we are secretly convinced that it is not such a slipshod, haphazard affair, that everything in it has meaning.
>
> J. B. Priestley

The Weak Anthropic Principle should not be viewed as a falsifiable theory or theorem. It is a methodological principle which one ignores at one's peril. There have been examples of how a lack of recognition of it allows the cosmologist to speculate unnecessarily, and, in fact, to develop quite unnecessary new theories of gravity. Let us recall the most striking of these, because it was the stimulus for the first explicit cosmological statement of the Weak Anthropic Principle by Robert Dicke in 1957.

In the 1930s (while he was on his honeymoon, in fact) Paul Dirac drew attention to a peculiar coincidence of Nature: that the number of particles in the observable universe is roughly equal to 10^{78}, whereas the ratio of the strengths of electromagnetic to gravitational forces between two protons is close to 10^{39}. That these numbers are so huge is strange enough, but the fact that one is the square of the other suggests that they are not totally unrelated, perhaps by an equation like:

$$(\text{Ratio of strengths of electric and gravitational forces})^2 = \text{Number of atoms in the observable Universe} \qquad (7.1)$$

22

However, the suggestion that the square of the first ratio is actually *equal* to the latter (as $10^{78} = 10^{39} \times 10^{39}$) creates a dilemma: the ratio of the intrinsic strengths of electric to gravitational forces is fixed by *constants of Nature* (the charge of the electron, the mass of the proton, and the Newtonian gravitational constant), and is believed to be the same everywhere and everywhen. By contrast, the number of particles in the *observable* Universe is continually increasing. At every moment, light rays that began their journey to us from huge distances away are reaching our telescopes for the first time. Objects that are further away from us than the speed of light multiplied by the time for which the Universe has been expanding have not yet been seen. We are surrounded by a spherical horizon about 15 billion light-years away which separates the observable part of the Universe (in its interior) from the, as yet unobserved, part beyond the horizon. But as time passes we can see farther and farther, and the number of atoms or particles in the visible part of the Universe will steadily increase—in direct proportion to the age of the Universe. The only way in which Dirac's equality (7.1) can hold is if either the strength of the electromagnetic force or of the gravitational force were to change with time.

Either suggestion is extremely radical. Dirac suggested that it was the intrinsic strength of gravity that weakened in inverse proportion to the age of the Universe. This idea subsequently generated a vast amount of theoretical and experimental physics as mathematicians showed how Einstein's theory of gravitation could be changed to include this feature, and various experiments were constructed to check whether the strength of gravity was changing as the Universe aged. To this day there is no evidence that gravity is weakening with time. As a result of the Viking space missions to Mars, we know that if gravity is weakening with time then it can have changed by no more than one per cent in the entire 15 billion year history of the Universe. This is a hundred times smaller than the change predicted by Dirac.

The interesting point about this story is that, in 1964, Robert Dicke, an American physicist working at Princeton, pointed out that Dirac's coincidence between the square of the relative strength of gravity and electromagnetism and the number of particles in the visible universe today was, in fact, one upon which our own existence depends. It tells us that we live close to the time when stars have started to burn their hydrogen into helium. Observers can only arise when the Universe has aged sufficiently for Dirac's coincidence to hold. Universes which do not display Dirac's coincidence are unlikely to contain observers. It is an anthropic selection effect, and no varying gravitational force strengths need be invoked to explain our observation of it.

The speculative Anthropic Principle

> There are two times in a man's life when he should not speculate: when he can't afford it, and when he can.
>
> Mark Twain

The style of argument we have just been discussing is rather striking, and it provoked a number of cosmologists to indulge in more speculative extensions of it. Just suppose, they suggested, that there is an infinity of all possible universes having all possible sizes, ages, temperatures, shapes, and contents. Even imagine that the strengths of gravity and electromagnetism take on different combinations of values in each one of them. Then how large is the collection of possible universes within which observers could arise? In some sense it seems to be very small. If the fundamental forces of Nature are imagined to possess slightly different strengths than they do, then chemistry becomes impossible: there are no stars, no carbon compounds, and apparently no observers. Thus, it appears that of all the universes *we* can conceive, very few are able to support life. Most are stillborn, unable to produce the basic building-blocks of life or provide an environment in which evolution by natural selection can produce non-trivial results.

It is hard to know what to make of this type of argument at present. It may be true that the Universe could have been different. It may not. If it is true that universes with all possible structures can exist, and the collection of possible universes in which life can arise is very small, then there need be no further explanation for many of the Universe's observed properties. Indeed, if the Universe is infinitely large in spatial extent (a view that current observations favour) and its initial conditions are random, then somewhere within that randomly infinite set of starting conditions there must exist an infinite collection of sub-regions that will expand into the uniform and isotropic expanding visible region we call the presently observable universe. In this situation there is no deeper explanation for its *observed* large-scale properties. This is a rather unsavoury state of affairs for the scientist whose goal is to explain the complexity of the Universe we witness by recourse to a small number of all-embracing laws of Nature. There may exist such a 'simple' explanation, but then again there may not. We may well be living in a habitable portion of an infinite and random universe whose initial state obeyed no laws of Nature at all.

23

Life and observership

> I am always surprised when a young man tells me he wants to work at cosmology; I think of cosmology as something that happens to one, not something one can choose.
>
> W. H. McCrea

When we list the medley of conditions that must be satisfied in order that any type of chemical life evolve in the Universe, we find that a large number of very finely balanced 'coincidences' must exist in order that the Universe give rise to observers. If we were to imagine a whole collection of hypothetical 'other universes' in which all the quantities that define the structure of our Universe take on all possible permutations of values, then we find that almost all of these other possible universes we have created on paper are stillborn, unable to give rise to that type of chemical complexity that we call 'life'. This discovery led Brandon Carter to suggest that there might exist some more speculative metaphysical aspect to the Universe which he termed the *Strong* Anthropic Principle, to distinguish it from the uncontroversial Weak Anthropic Principle discussed above. The Strong Anthropic Principle suggests that, because there appear to exist such a large number of remarkable and apparently disconnected 'coincidences' which conspire to allow life to be possible in the Universe, the Universe *must* give rise to observers at some stage in its history.

Now this sounds rather strange. Cosmologists are talking about 'other universes'. Where are they? How can one say that our Universe is better suited to the evolution of life than another? We also speak of 'life' and 'observers' as though they have some role to play in physics. How can this be? In the search for answers modern physics points us in some surprising directions.

With regard to the way the Universe began we have two options which are considered seriously by theoretical cosmologists. Over the last twenty years favour has continually ebbed and flowed between the two. First, it could be that there is only one type of Universe that is logically possible. All the presently unexplained values of the fundamental constants of Nature would, in such a unique scheme, be found to possess no possible arbitrariness. There will be found to exist a single 'Theory of Everything', and this branch of scientific inquiry will then be complete. The current great excitement amongst theoretical physicists for 'superstrings' has arisen because this idea provides the first good candidate for a 'Theory of Everything'. The other possibility is that there are elements of randomness in the make-up of both the structure of the Universe and the fundamental constants of Nature. This randomness can emerge in various ways. The Universe as a whole may vary greatly in composition from place to place. If the constants of Nature arise from the breaking of some symmetry then this could have happened in different ways in different places, and may even be different elsewhere in the Universe today. The phenomenon of inflation could ensure that the laws and constants of Nature are similar only within the co-ordinated primordial region that inflated to form our visible Universe.

The conclusion of the second scenario is that the Universe could have been different. It is in some sense a particular asymmetric manifestation of deeper, but now partially hidden, symmetry. The symmetries of the laws of Nature are hidden by the need for particular outworkings of them to occur.

If the Universe is uniquely prescribed by some higher internal logic then we must judge ourselves extremely fortunate that this unique self-consistent arrangement happened to allow the evolution of observers to witness it, and we are unable to conclude anything further about the connection between life and the Universe without appealing to metaphysical or religious beliefs. If the Universe possesses some random aspect in its make-up then the verdict is rather different. We must accept that, contrary to the prejudice of many scientists, there are aspects of the large-scale structure of the Universe which do not have any explanation in the conventional sense. They arise as random events in the first moments of the Universe's history. They could have been otherwise (and may even *be* otherwise elsewhere in the Universe). We could not exist in the majority of possible universes, where the outcomes of such accidents lead to universes that cannot support life.

This still does not really introduce observers in any way that makes them necessary for the existence of the Universe, rather than just for the observing of it. We could imagine a Universe empty of life. It seems a lonely place—meaningless perhaps—but it doesn't seem logically impossible or physically inconsistent in any way. Or does it?

The greatest achievement of physical science in modern times has been the development and use of quantum theory. It is this branch of physics which underpins our everyday existence. Our understanding of its workings is so good that we are able to use it to develop lasers, transistors, microchips, and computers; the whole of our technological society is built upon it in a thousand different ways. But this totally pragmatic science which allows us to understand the microscopic structure of matter in fantastic detail, and which governs the behaviour of every atom of Nature and every DNA helix within our bodies, contains a deep mystery of physics at its heart. In Chapter 3 we saw how the standard form in which it was developed by Niels Bohr during the pre-war era maintains that no phenomenon exists until it is observed. And when it is observed, the state in which it is seen is determined unpredictably by the act of observership. All that

24

definitely can be predicted about it is the probability that a particular measurement will be recorded when the state is observed.

According to Bohr, the only real properties of natural phenomena are observed phenomena. We can no longer maintain the old Cartesian view that we can observe Nature like a bird-watcher with a perfect hide. There is an unbreakable connection between the observer and the observed. The eminent American physicist and long-time co-worker of Bohr's, John A. Wheeler, has proposed that taken at face value this interpretation of quantum mechanics requires 'observers' in order to bring the quantum world into being. Thus, according to Bohr's extreme interpretation of quantum theory, the quantum reality of the distant stars and galaxies cannot be granted until they are 'observed'. In Wheeler's words 'observers may be necessary to bring the Universe into being'.

We still do not fully understand what properties are necessary to constitute an 'observer' in quantum physics. Some argue that any device for storing information will suffice, but others, most notably the Nobel laureate Eugene Wigner, have argued that the self-reflective property of human consciousness is necessary.

Bohr's interpretation of quantum theory is strange, but it is the one that working physicists adopt pragmatically without worrying about how it works. Such subtleties do not impinge upon its practical use in the laboratory. However, in recent years cosmologists have begun to consider the implications of applying quantum theory to the Universe as a whole—quantum cosmology. Such a programme immediately faces an impasse that can only be overcome by coming to grips with the meaning of quantum observership. If we only ascribe reality to what is observed, who observes the Universe? If we can only make statements about the probability of the Universe being observed in a particular state what does this mean when there is only one Universe? The 'Many Worlds' interpretation of quantum reality maintains that every time an observation is made there is a splitting of the observer or the world into two states—one for each possible outcome of the observation. Thus, the Universe evolves by successively splitting into an ever-increasing collection of different worlds in which everything that can logically occur eventually will. The randomness of quantum measurement, which Bohr regarded as intrinsic to the inseparability of the observer and the observed, is an illusion created by the fact that we experience just one path through the network of world-splittings. This interpretation of quantum reality is adopted by quantum cosmologists because it does not require the Universe to be observed. It ascribes an equal reality to universes other than the one we experience and observe. It ascribes no special significance to life at the quantum level, but there must exist branches where life evolves, because all possible outworkings of the laws of Nature are explored in the different branches of the Everett worlds. This array of universes is equivalent to the picture in which there are random elements in the make-up of our single Universe, although here the probability distribution is realized rather than potential.

As yet we do not know whether Bohr or Everett was right about the meaning of quantum mechanics. Whatever the answer to the riddle of quantum reality, the correct assessment of the role and meaning of observers in the Universe must await the outcome of the confrontation of the Cosmos with the quantum. The marriage of these unlikely partners will bring cosmologists face to face with the question of the origin of the Universe: a conundrum in which we are found to play a mysterious and unexpected part.

Is the Anthropic Principle an argument for the existence of God?

> Give us the power, O Lord, for if thou doest not give us the power we shall not give thee the glory, and who will be the gainer by that, O Lord?
>
> Old prayer

History reveals that most past cultures, be they Eastern or Western, harboured a deep intuitive belief that Nature was providentially designed for them by a benevolent God or gods. The beauty of Nature, the availability of natural resources, the recurrence of night and day, summer and winter, seed-time and harvest, seemed to bear eloquent witness to such a state of affairs. The Old Testament view that fashions so much of our own cultural and scientific heritage is no exception. It structured and reflected the early Jewish view of Nature. It would, of course, never have occurred to a pious Jew that Nature should be used to prove the existence of God. There was no doubt of His existence. Nature was something to be celebrated and to be part of. It was also something secular. There were no nature gods. Although the underlying assumption and belief was grounded in teleology it was not always naïvely anthropocentric, as the book of Job bears eloquent witness. Later, the Greek ideas of Aristotle attained a pre-eminence in philosophical thinking in Europe. Aristotle laid great stress upon the purpose of things as revealing their true meaning and significance. He was interested in 'why' things happened as well as 'how'. The Aristotelian view became merged with the Judaeo-Christian tradition, and was used to frame many arguments for the existence of God from the existence of apparent 'design' in Nature. Later, the revolutionary change of method and emphasis in science brought about by figures like Copernicus, Galileo, and Newton tailored the objectives of science to answering the 'how' questions and not the 'why'.

25

One might have thought that this would lead to a demise of arguments for the existence of God being formed from the apparent life-supporting purpose of the world. Nothing could be farther from the truth. The dramatic unfolding of the laws of Nature, culminating in Newton's great works, led only to a change of emphasis. The evidence for the existence of a Deity was taken to be the meticulous mathematical precision and regularity of the laws of Nature themselves rather than individual events. Alongside this remained a more naïve argument which cited the match between human and animal physiology, and their needs for survival in the environments in which they were found, as evidence for the providential design of Nature. It was the latter view that Darwin's theory of natural selection completely undermined. However, as was widely appreciated at the time, the Darwinian revolution had nothing to say about the other type of Newtonian Design Argument based upon the mathematical harmony of the unchanging laws of Nature.

For some there still exists an irresistible temptation to draw a strong metaphysical conclusion from the fact that we have found a whole collection of coincidences in the make-up of the physical Universe, which contrive in concert to make our existence possible. Is this not evidence for the existence of a God who has created the Universe with mortal man in mind, they ask? This type of 'natural theology', as it became known, originated as a systematic study in medieval times, but reached its peak amongst English scientists in the seventeenth century. Flushed with the success of the Newtonian revolution, they sought to understand the order they had found in terms of their religious beliefs. The meticulous clockwork precision and regularity of the underlying laws of Nature was cited as primary evidence for the existence of a Grand Designer behind the Universe, who was identified with the God of the theologians. Such ideas were absorbed within the mainstream of Protestant theology of the day and eloquently expressed by the hymn-writer, whose famous lines, 'Laws which never shall be broken/ For their guidance hath He made' sprang from the contemplation of the newly revealed laws of Nature. And indeed, Newton was pleased with this use of his ideas. In the introduction of the *Principia* he remarked that in its writing he had an 'eye upon arguments' for belief in a Deity. And most intriguingly, he wrote to Richard Bentley that 'There is yet another argument for a Deity wch I take to be a very strong one, but till ye principles on wch tis grounded be better received I think it more advisable to let it sleep.' Newton never revealed this new argument. It is possible, in view of the context of the remark, that Newton had deduced an age of a hundred million years or so for the past lifetime of the solar system, based upon his law of gravitation.

Today, this type of theological deduction is still attractive to many

individuals, and the merest suspicion of it seems to spur its opponents to man the trenches just as the theological exploitation of Newton's work led to the critical reaction of David Hume and Immanuel Kant against any argument for God from the so-called design of Nature. Heinz Pagels believes that some scientists regard the Anthropic Principle as a form of substitute religion; he claims,

> Of course, some scientists, believing science and religion mutually exclusive . . . [when] . . . faced with questions that do not fit into the framework of science . . . are loath to resort to religious explanation; yet their curiosity will not let them leave matters unaddressed.

Maybe this is not so unusual a charge as it first sounds. Others have remarked upon the curious resemblance between traditional religious ideas about 'salvation' and the motivations of some eminent searchers for extra-terrestrial intelligence, who believe that contact with advanced civilizations will reveal to us the secret of successful world government, and 'save us from ourselves'.

Motivated by these controversies, John Updike has recently felt the need to write an entire novel in which the English natural theological tradition is pitted against the Barthian stance of the utter transcendance and inaccessibility of God to mundane scientific arguments. The cyclopean enthusiasm of a young computer student out to develop a computer code that will lead from Nature up to Nature's God, aided by the anthropic coincidences amongst the fundamental constants of Nature, is foiled by the arid scepticism of a liberal theologian convinced of Barth's words, that there is 'no way from us to God', and alarmed at the prospect because, 'The god who stood at the end of some human way . . . would not be God'. It is interesting that Updike has used many (although often garbled) cosmological coincidences to decorate the dialogue of the novel, and acknowledges a selection of popular scientific articles as their source.

There are two simple things one can say about well-meaning logical and scientific quests for the existence of God (or gods). The logical arguments are all of a piece. They begin with some assumptions ('axioms' as the logicians like to call them), and then proceed to deduce the existence of God by a series of inexorable logical steps. But in the last analysis we are left not with a conclusion, but a choice. Only if we believe the assumptions at the outset must we believe the conclusions. There cannot be an ineluctable logical proof of God's existence or non-existence. There will always be a choice about the credibility of assumptions. Furthermore, one suspects that even the great propounders of logical arguments for the existence of God, like Thomas Aquinas, had a personal faith that would not have been perturbed one iota by the undermining of their logical or scientific

demonstrations, because it was grounded elsewhere. By the same token, they could not honestly have expected their arguments to sway anyone else to accept their conclusion.

It is for reasons of this sort that the Strong Anthropic coincidences cannot be the basis of a cogent argument for God's existence from apparent anthropocentric design in the Universe, although they are quite consistent with such a conclusion. The wide range of remarkable coincidences between values of constants of Nature which have allowed complex living things to evolve are only conditions *necessary* for the existence of life. They are not sufficient to guarantee it. Modern biologists reject the notion that the evolution of life in the Universe is in any sense inevitable. Such a teleological view—that there is some future goal to which Nature is directed or magnetically attracted—finds no support in known facts, although it recaptured popular attention in the 1950s and 1960s following its semi-poetic espousal by the Catholic scientist and mystic Teilhard de Chardin.

The Anthropic Principle merely identifies coincidences which are *necessary* for the evolution of complex chemistry of the sort that biochemists believe to be essential for the spontaneous evolution of life by natural selection. The fact that it finds these 'coincidences' to be numerous and surprising does not allow the conclusion to be drawn that they also guarantee the presence of conscious observers in the Universe.

The time of your life

> I have not found out why we humans think of time as a line going from backwards, forwards, whilst it may be in all directions like everything else in the system of the world.
>
> Ferruccio Busoni

We are all aware of the subjectivity of time. Although we sense the arrival of the future, we have no sure sense of the rate of passage of time. Yet, despite this we hold to the belief (because we have read it) that there is a definite time behind the subjectivity of our experience, and that time distinguishes the future from the past in an absolute way. These thoughts take us back to issues that we have raised in earlier chapters. In Chapter 3 we encountered the dilemma of the Second Law of thermodynamics, which picks out an 'arrow' of time by the direction of entropy increase. In Chapter 4 we encountered the discovery of the expanding universe, which provides us with another arrow of time in the sense of expansion. The paradox of quantum measurement which arose in Chapter 3, wherein the time-reversible evolution of the quantum wave function is supplanted by the irreversible effect of quantum measurement, gives rise to another breed of

irreversibility. In principle, all these 'arrows' which distinguish the future from the past might be distinct from the subjective consciousness we have of the direction of future time. Some have speculated that they might all be linked in some deep way. In his suggestion that there exists a fundamental 'entropy' which gauges the evolution of the complexity of the gravitational field of the Universe, Roger Penrose hypothesizes that such a quantity plays some role in uniting the irreversible antics of thermodynamics and quantum measurement with the overall evolution of the Universe.

Connections between the local thermodynamic arrows of time and the expansion of the Universe have been suggested before. It was once a popular speculation to link the two with the result that, if the Universe were one day to reverse its expansion into contraction, then the local arrows of time would also reverse, and we would see entropy decrease in the future. Our desks would grow spontaneously tidy; perpetual-motion machines would abound. Such a conclusion does not really stand up to close analysis though. The reversal of the expansion dynamics of the Universe is a global phenomenon, whereas the arrow of time in some microscopic physical process giving rise to frictional resistance here and now is a local one. How can the local process 'know' that the Universe elsewhere has expanded to its maximum extent. If we attribute the local arrow of thermodynamics to the *local* expansion dynamics then we have a chaotic situation. In a realistic universe some places will reach their expansion maxima before others. The universe will be composed of regions, some expanding, some contracting, with different thermodynamic arrows of time.

The idea that there might be different arrows of thermodynamic time is an old one which pre-dates the idea that time's arrow might be connected with the expansion of the Universe. In Chapter 3 we discussed how the thermodynamic arrow of entropy increase is a reflection of the relative probabilities of various states. Ordered states are far more improbable than disordered ones, and so it is far more likely that a system will evolve from a state of order into one of chaos. Boltzmann saw that this view precipitated a subjective view of the Second Law of thermodynamics and the direction of time which it defines. If the Universe varies from place to place in its initial state of disorder, then there will be some places which begin in an improbable state, and from which entropy and disorder tend to increase, but there will exist other regions which begin in probable disordered states from which the evolution can proceed to states of lower entropy and disorder. In this way the local thermodynamic arrows of time in the Universe would be different. Boltzmann explains:

> We have the choice of two kinds of picture. Either we assume that the whole universe is at present in a very improbable state. Or else we assume

> that the aeons during which this improbable state lasts, and the distance from here to Sirius, are *minute* if compared with the age and size of the whole universe. In such a universe, which is in thermal equilibrium as a whole and therefore dead, relatively small regions of the size of our galaxy will be found here and there; regions (which we may call 'worlds') which deviate significantly from thermal equilibrium for relatively short stretches of these 'aeons' of time. Among these worlds the probability of their state will increase as often as they decrease. In the universe as a whole the two directions of time are indistinguishable, just as in space there is no up or down. However, just as at a certain place on the earth's surface we can call 'down' the direction towards the centre of the earth, so a living organism that finds itself in such a world at a certain period of time can define the 'direction' of time as going from the less probable state to the more probable one (the former will be the 'past' and the latter the 'future'), and . . . he will find that his own small region, isolated from the rest of the universe is 'initially' always in an improbable state. It seems to me that this way of looking at things is the only one which allows us to understand the validity of the second law, and the heat death of each individual world, without invoking a unidirectional change of the entire universe from a definite initial state to the final state.

Events in the parts of Boltzmann's world where entropy decreases would be very strange. Poincaré argued that familiar concepts like 'prediction' would be hopeless. Friction would cease to be a retarding force. Objects would be spontaneously accelerated. In the future of our subjective time the oceans would not tend to equilibrate their temperatures. Inequalities would grow in an unstable fashion. This is not the type of world where life can either evolve or survive. Thus Boltzmann's world need not conflict with the world we see. The Weak Anthropic Principle persuades us that we could only exist in one of the entropy-increasing islands.

As a final speculation upon this world of many times we might consider what would happen at the interfaces between regions where the thermodynamic arrows of time differ. Norbert Wiener cites the following conundrum:

> It is a very interesting intellectual experiment to make the fantasy of an intelligent being whose time should run the other way to our own. To such a being all communication with us would be impossible. Any signal he might send would reach us with a logical stream of consequents from his point of view, antecedent from ours. These antecedents would already be in our experience, and would have served to us as the natural explanation of his signal, without presupposing an intelligent being to have sent it.

Reichenbach proposed one idea for communication which might be explored in order to ascertain that the other beings did have a counter-oriented thermodynamic arrow:

> That such a system is developing in the opposite time direction might be discovered by us from some radiation travelling from the system to us and perhaps exhibiting a shift in spectral lines upon arrival . . . the radiation travelling from the system to us would . . . not leave that system but arrive at it. Perhaps the signal could be interpreted by inhabitants of that system as a message from our system telling them that our system develops in the reverse time direction. We have here a connecting light ray which, for each system, is an arriving light ray annihilated in some absorption process.

It is not difficult to conceive of better tests than this which exploit the properties expected in sources of radio waves and noise in transmission signals, but we shall resist the urge to speculate further. It is left, in the immortal words of the textbook-writer, as an exercise for reader.

Cosmology, stars, and life

> As the mystic said to the hot-dog vendor, 'Make me one with everything'
>
> Laurence Kushner

Prior to the discovery of the expansion of the Universe there was little that cosmology could contribute to the question of extraterrestrial life aside from probabilities and prejudices. After our discovery of the expansion and evolution of the Universe the situation changed significantly. The entire cosmic environment was recognized as undergoing steady change. The history of the Universe took on the complexion of an unfolding drama in many acts, with the formations of first atoms and molecules, then galaxies and stars, and most recently, planets and life. The most important and simplest feature of the overall change in the Universe that the expansion produces is the rate at which it occurs. This is linked to the age of the expanding universe and that of its constituents.

In the 1930s, the distinguished biologist J. B. S. Haldane took an interest in Milne's proposal that there might exist two different timescales governing the rates of change of physical processes in the Universe: one, t, for 'atomic changes' and another, τ for 'gravitational changes' with $\tau = \log(t/t_0)$ and t_0 constant. Haldane explored how changing from one timescale to the other could alter your picture of when conditions in the Universe would become suitable for the evolution of biochemical life. In particular, he

28

argued that it would be possible for radioactive decays to occur with a decay rate that was constant on the t timescale but which grew in proportion to t when evaluated on the τ scale. The biochemical processes associated with energy derived from the breakdown of adenosine triphosphoric acid would yield energies which, while constant on the t scale, would grow as t^2 on the τ scale. Thus there would be an epoch of cosmic history on the tau scale before which life was impossible but after which it would become increasingly likely. Milne's theory subsequently fell into abeyance, although the interest in gravitation theories with a varying Newtonian 'constant' of gravitation led to detailed scrutiny of the paleontological and biological consequences of such hypothetical changes for the past history of the Earth. Ultimately, this led to the formulation of the collection of ideas now known as the Anthropic Principles.

Another interface between the problem of the origin of life and cosmology has been the perennial problem of dealing with finite probabilities in situations where an infinite number of potential trials seem to be available. For example, in a universe that is infinite in spatial volume (as would be expected for the case for an expanding, open universe with non-compact topology), any event that has a finite probability of occurring should occur not just once but infinitely often with probability one if the spatial structure of the Universe is exhaustively random. In particular, in an infinite universe we conclude that there should exist an infinite number of sites where life has progressed to our stage of development. In the case of the steady-state universe, it is possible to apply this type of argument to the history of the universe as well as its geography because the universe is assumed to be infinitely old. Every past-directed world line should encounter a living civilization. Accordingly, it has been argued that the steady state universe makes the awkward prediction that the universe should now be teeming with life along every line of sight.

The key ingredient that modern cosmology introduces into considerations of biology is that of *time*. The observable universe is expanding, and not in a steady state. The density and temperature are steadily falling as the expansion proceeds. This means that the average ambient conditions in the universe are linked to its age. Roughly, in all expanding universes, dimensional analysis tells us that the density of matter, ρ, is related to the age t and Newton's gravitation constant, G, by means of a relation of the form $\rho \approx 1/Gt^2$.

The expanding universe creates an interval of cosmic history during which biochemical observers, like ourselves, can expect to be examining the Universe. Chemical complexity requires basic atomic building blocks which are heavier than the elements of hydrogen and helium which emerge from the hot early stages of the universe. Heavier elements, like carbon, nitrogen, and oxygen, are made in the stars, as a result of nuclear reactions that take billions of years to complete. Then, they are dispersed through space by supernovae after which they find their way into grains, planets, and ultimately, into people. This process takes billions of years to complete and allows the expansion to produce a universe that is billions of light years in size. Thus we see why it is inevitable that the universe is seen to be so large. A universe that is billions of years old and hence billions of light years in size is a necessary prerequisite for observers based upon chemical complexity. Biochemists believe that chemical life of this sort, and the form based upon carbon in particular, is likely to be the only sort able to evolve spontaneously. Other forms of living complexity (for example that being sought by means of silicon physics) almost certainly can exist but it is by being developed with carbon-based life-forms as a catalyst rather than by spontaneous evolution.

The inevitability of universes that are big and old as habitats for life also leads us to conclude that they must be rather cold on average because significant expansion to large size reduces the average temperature inversely in proportion to the size of the universe. They must also be sparse, with a low average density of matter and large distances between different stars and galaxies. This low temperature and density also ensures that the sky is dark at night (the so called 'Olbers' Paradox' first noted by Halley) because there is too little energy available in space to provide significant apparent luminosity from all the stars. We conclude that many aspects of our Universe which, superficially, appear hostile to the evolution of life are necessary prerequisites for the existence of any form of biological complexity in the Universe.

Life needs to evolve on a timescale that is intermediate between the typical time scale that it takes for stars to reach a state of stable hydrogen burning, the so called main-sequence lifetime, $t\{\text{star}\}$, the timescale on which stars exhaust their nuclear fuel and gravitationally collapse. This timescale is determined by a combination of fundamental constants of Nature.

Evidently, in our solar system, life first evolved quite soon after the formation of a hospitable terrestrial environment. Suppose the typical time that it takes for life to evolve is denoted by some timescale $t\{\text{bio}\}$, then from the evidence presented by the solar system, which is about 4.6 billion years old, it seems that $t\{\text{bio}\} \approx t\{\text{star}\}$. A first sight we might assume that the microscopic biochemical processes and local environmental conditions that combine to determine the magnitude of $t\{\text{bio}\}$ are independent of the nuclear astrophysical and gravitational processes that determine the typical stellar main sequence lifetime $t\{\text{star}\}$. However, this assumption leads to the striking conclusion that we should expect extraterrestrial forms of life

to be exceptionally rare. The argument, in its simplest form, is as follows. If $t\{\text{bio}\}$ and $t\{\text{star}\}$ are independent then the time that life takes to arise is random with respect to the stellar timescale $t\{\text{star}\}$. So it is most likely that either $t\{\text{bio}\}$ is far greater than $t\{\text{star}\}$, or that $t\{\text{bio}\}$ is far smaller than $t\{\text{star}\}$. But if $t\{\text{bio}\}$ is far less than $t\{\text{star}\}$, we must ask why it is that the first observed inhabited solar system (that is, us) has $t\{\text{bio}\} \approx t\{\text{star}\}$. This would seem to be extraordinarily unlikely. On the other hand, if $t\{\text{bio}\}$ is much bigger than $t\{\text{star}\}$, then the first observed inhabited solar system (us) is most likely to have $t\{\text{bio}\} \approx t\{\text{star}\}$, since systems in which $t\{\text{bio}\}$ is much bigger than us have yet to evolve. Thus we are a rarity, one of the first living systems to arrive on the scene. Generally, we are led to a conclusion, and a somewhat pessimistic one for searchers after extraterrestrial life, that $t\{\text{bio}\}$ greatly exceeds $t\{\text{star}\}$ and we are likely to be among the first cosmic outposts of life.

In order to escape from this conclusion we have to undermine one of the assumptions underlying the argument that leads to it. For example, if we suppose that $t\{\text{bio}\}$ is not independent of $t\{\text{star}\}$, then things look different. If $t\{\text{bio}\}$ increases as $t\{\text{star}\}$ increases, then it is becomes likely that we will find $t\{\text{bio}\} \approx t\{\text{star}\}$. The astrophysicist Mario Livio has given a simple model of how it could be that $t\{\text{bio}\}$ and $t\{\text{star}\}$ are related in this way. He takes a very simple model of the evolution of a life-supporting planetary atmosphere, like the Earth's, to have two key phases determine the fraction of oxygen in the atmosphere. During phase 1, oxygen is released by the photodissociation of water vapour. On Earth this took 2.4 billion years and led to a build-up of oxygen to about one-thousandth of its present value. During phase 2, the oxygen and ozone levels grow to about one tenth of their present levels. This is sufficient to shield the Earth's surface from lethal levels of ultra-violet radiation in the 2000–3000 Ångstrom band (nucleic acid and protein absorption of ultra-violet radiation peaks in the 2600–2700 and 2700–2900 Ångstrom bands, respectively). On Earth this phase took about 1.6 billion years.

Now the length of Phase 1 might be expected to be inversely proportional to the intensity of radiation in the wavelength interval 1000–2000 Ångstroms, where the key molecular levels for absorption by water molecules lie. Studies of stellar evolution allow us to determine this time interval and provide a rough numerical estimate of the resulting link between the biological evolution time (assuming it to be determined closely by the photodissociation time) and the main-sequence stellar lifetime.

This simple model indicates a possible route towards establishing a link between the biochemical time-scales for the evolution of life and the astrophysical time-scales that determine the time required to create an environment supported by a stable hydrogen-burning star. There are obvious weak

links in the argument. It provides only a necessary condition for life to evolve, not a sufficient one. We know that there are many other events that need to occur before life can evolve in a planetary system. We could imagine being able to derive an expression for the probability of planet formation around a star. This should involve many other factors which determine the amount of material available for the formation of solid planets with atmospheres at distances which permit the presence of liquid water and stable surface conditions. Unfortunately, we know that there were many 'accidents' of the planetary formation process in the solar system which have subsequently played a major role in the existence of long-lived stable conditions on Earth. For example, the presence of resonances between the precession rates of rotating planets and the gravitational perturbations they feel from all other bodies in their solar system can easily produce chaotic evolution of the tilt of a planet's rotation axis with respect to the orbital plane of the planets over times much shorter than the age of the system. The planet's surface-temperature variations, insolation levels, and sea levels are sensitive to this angle of tilt. It determines the climatic differences between what we call 'the seasons'. In the case of the Earth, the modest angle of tilt (approximately 23 degrees) would have experienced this erratic evolution had it not been for the presence of the Moon. The French astronomer Jacques Laskar and his colleagues have shown that the Moon is large enough for its gravitational effects to dominate the resonances which occur between the Earth's precessional rotation and the frequency of external gravitational perturbations from the other planets. As a result, the Earth's tilt wobbles only by one half a degree around 23 degrees over hundreds of thousands of years. Enough perhaps to cause some climatic change—ice ages even—but not catastrophic for the evolution of life.

This shows how the causal link between stellar lifetimes and biological evolution times may be rather a minor factor in the chain of fortuitous circumstances that must occur if habitable planets are to form and sustain viable conditions for the evolution of life over long periods of time. The problem remains to determine whether other decisive astronomical factors in planet formation are functionally linked to the surface conditions needed for biochemical processes.

We know that several of the distinctive features of the large scale structure of the visible Universe play a rôle in meeting the conditions needed for the evolution of biochemical complexity within it. The first example is the proximity of the expansion dynamics to the 'critical' state which separates an ever-expanding future from one of eventual contraction, to better than ten per cent. Universes that expanded far faster than this would be unable to form galaxies and stars, and hence the building blocks of biochemistry would be absent. The rapid expansion would prevent islands of material

separating out from the global expansion and becoming bound by their own self-gravitation. By contrast, if the expansion rate were far below that characterizing the critical rate then the material in the Universe would have condensed into dense structures and black holes long before stars could form.

The second example is that of the uniformity of the Universe. The non-uniformity level on the largest scales is very small, approximately one part in 100,000. If it were significantly larger then galaxies would have rapidly degenerated into dense structures within which planetary orbits would be disrupted by tidal forces, and black holes would form rapidly before life-supporting environments could be established. If it were significantly smaller then the non-uniformities in the density would be gravitationally too feeble to collapse into galaxies and no stars would form. Again, the Universe would be bereft of the biochemical building blocks of life.

In recent years, the most popular theory of the very early evolution of the Universe has provided a possible explanation as to why the Universe expands so close to the critical life-supporting divide and why the fluctuation level has the value observed. This is the picture of 'inflation' that we have already encountered. Recall that it proposes that during a short interval of time when the temperature was very high, the expansion of the Universe accelerated. This required the material content of the Universe to be temporarily dominated by forms of matter which effectively anti-gravitated for that period of time. Remarkably, modern theories of particle physics predict that many such forms of matter should exist in Nature. As yet, we have not got particle colliders energetic enough to produce them, and so we cannot be sure that the type of anti-gravitation they display is strong enough and persistent enough to produce inflationary expansion of the Universe in its early stages. In fact, astronomers hope that by working out the very detailed cosmological consequences of these particles they may be able to discover whether or not they exist more easily than the particle physicists.

The inflation is envisaged to end because the matter fields responsible decay into other forms of matter, like radiation, which do not display anti-gravitation. After this occurs, the expansion resumes the state of decelerating expansion that it possessed before its inflationary episode began.

If inflation occurs it offers the possibility that the whole of the visible part of the Universe (roughly fifteen billion light-years in extent today) has expanded from a region that was small enough to be causally linked by light signals at the very high temperatures and early times when inflation occurred. If inflation does not occur then the visible Universe would have expanded from a region that is far larger than the distance that light can circumnavigate at these early times, and so its smoothness today is a mystery. If inflation occurs it will transform the irreducible quantum statistical fluctuations in space into distinctive patterns of fluctuations in the microwave background radiation, which future satellite observations will be able to detect if they were of an intensity sufficient to have produced the observed galaxies and clusters by the process of gravitational instability.

As the inflationary universe scenario has been explored in greater depth it has been found to possess a number of unexpected properties which, if they are realized, would considerably increase the complexity of the global cosmological problem and create new perspectives on the existence of life in the Universe.

It is possible for inflation to occur in different ways in different places in the early Universe. The effect is rather like the random expansion of a foam of bubbles. Some inflate considerably while others hardly inflate at all. This is 'chaotic inflation'. Of course, we have to find ourselves in one of the regions that underwent sufficient inflation so that the expansion lasted for longer than $t\{star\}$, and stars could produce biological elements. In such a scenario the global structure of the Universe is predicted to be highly inhomogeneous. Our observations of the microwave background temperature structure will only be able to tell us whether the region which expanded to encompass out visible part of the Universe underwent inflation in its past. An important aspect of this theory is that for the first time it has provided us with a positive reason to expect that the observable Universe is not typical of the structure of the Universe beyond our visible horizon, fifteen billion light years away.

We have already disucssed (p. 400) how, under fairly general conditions, inflation can be self-reproducing. That is, quantum fluctuations within each inflating bubble will necessarily create conditions for further inflation of microscopic regions to occur. This process or 'eternal inflation' appears to have no end and may not have had a beginning. Thus life will be possible only in bubbles with properties which allow self-organized complexity to evolve and persist.

It has been found that there is further scope for random variations in these chaotic and eternal inflationary scenarios. In the standard picture we have just sketched, properties like the expansion rate and temperature of each inflated bubble can vary randomly from region to region. However, it is also possible for the strengths and number of low-energy forces of Nature to vary. It is even possible for the number of dimensions of space which have expanded to large size to be different from region to region. We know that we cannot produce the known varieties of organized biochemical complexity if the strengths of forces change by relatively small amounts, or in dimensions other than three because of the impossibility of creating chemical or gravitational bound states.

31

It is possible to have random variations like this because inflation is ended by the decay of some matter field which displays anti-gravitation due to its negative pressure. This corresponds to the field evolving to a minimum in its self-interaction potential. If that potential has a single minimum then the characteristic physics that results from that ground state will be the same everywhere. But if the potential has many minima (for example, like a piece of corrugated roofing) then each minimum will have different low-energy physics, and different parts of the Universe can emerge from inflation in different minima and with different effective laws of interaction for elementary particles. In general, we expect the symmetry breaking which chooses the minima in different regions to be independent and random.

Considerations like these, together with the light that superstring theories have shed upon the origins of the constants of Nature, mean that we should assess how narrowly defined the existing constants of Nature need to be in order to permit biochemical complexity to exist in the Universe. For example, if we were to allow the ratio of the electron and proton masses (β) and the fine structure constant (α) to change their values (assuming no other aspect of physics is changed by this assumption—which is clearly likely to be false!) then the allowed variations are very constraining. Increase β too much and there can be no ordered molecular structures because the small value of β ensures that electrons occupy well-defined positions in the Coulomb field created by the protons in the nucleus; if β exceeds about $0.005\alpha^{1/2}$, then there would be no stars; if modern grand unified gauge theories are correct then α must lie in a narrow range between about 1/180 and 1/85 in order that protons do not decay too rapidly and a fundamental unification of the forces of Nature can occur. If, instead, we consider the allowed variations in the strength of the strong nuclear force, α_s, and in α, then α_s must be less than about $0.3\alpha^{1/2}$ if biologically useful elements like carbon are to be stable. If we increase α_s by just 4 per cent disaster follows, because the helium-2 isotope can exist (it just fails to exist in practice) and allows very fast direct nuclear burning of hydrogen to helium. Stars would rapidly exhaust their fuel and collapse to very dense states or black holes. In contrast, if α_s were decreased by about 10 per cent then the deuterium nucleus would cease to be bound and the nuclear astrophysical pathways to the build-up of biological elements would be blocked. Again, the conclusion is that there is a rather small region of parameter space in which the basic building blocks of chemical complexity can exist.

We should stress that conclusions regarding the fragility of living systems with respect to variations in the values of the constants of Nature are not really rigorous in all cases. The values of the constants are simply assumed to take different constant values to those that they are observed to take and

the consequences of changing them one at a time are examined. However, if the different constants are fully linked together, as we might expect for many of them if a unified Theory of Everything exists, then many of these independent variations may not be possible. The consequences of a small change in one constant would have further necessary ramifications for the allowed values of other constants. One would expect the overall effect to be more constraining on the allowed variations that are life-supporting.

These considerations are likely to have a bearing on interpreting any future quantum cosmological theory. Such a theory, by its quantum nature, will make probabilistic predictions. It will predict that it is 'most probable' that we find the Universe (or its forces and constants) to take particular values. This presents an interpretational problem because it is not clear that we should expect the most probable values to be the ones that we observe. Since only a narrow range of the allowed values for, say, the fine structure constant will permit observers to exist in the Universe, we must find ourselves in the narrow range of possibilities which permit them, no matter how improbable they may be. This means that in order to test fully the predictions of future Theories of Everything, we must have a thorough understanding of all the ways in which the possible existence of observers is constrained by variations in the structure of the Universe, in the values of the constants that define its properties, and in the number of dimensions it possesses.

The misanthropists

> There is no possibility of reducing all laws to one law . . . no a priori means of excluding from the world the unique.
>
> Josiah Royce

The Weak Anthropic Principle recognizes the constraints that are placed upon what we can expect to observe in Nature by the selection effect of our own existence as observers made of carbon living billions of years after the Big Bang. This is an uncontroversial statement of truth, but is it a useful addition to our knowledge? Not everybody seems to think so. Heinz Pagels claims that

> As I thought more about the anthropic principle, however, it seemed less like a grand Darwinian selective principle and more like a far-fetched explanation for those features of the universe which physicists cannot yet explain. Physicists and cosmologists who appeal to anthropic reasoning seemed to me to be gratuitously abandoning the successful program of conventional physical science of understanding the quantitative

> properties of our universe on the basis of universal physical laws . . . We could debate its merits and demerits a long time. But such interminable debate is a symptom of what is wrong with the anthropic principle: unlike the principles of physics, it affords no way to determine whether it is right or wrong; there is no way to test it. Unlike conventional physical principles, the anthropic principle is not subject to experimental falsification . . . the influence of the cosmological principle on the development of contemporary cosmological models has been sterile: it has explained nothing . . . no knowledge has been gained by the adoption of anthropic reasoning. I would opt for rejecting the anthropic principle as needless clutter in the conceptual repertoire of science . . . My own view is that although we do not yet know the fundamental laws, when and if we find them the possibility of life in a universe governed by those laws will be written into them. The existence of life in the universe is not a selective principle acting upon the laws of nature; rather it is a consequence of time.

This enthusiastic condemnation includes most of the standard objections to the use of the Anthropic Principle. We can abstract them explicity as follows:

1. Scientists spent centuries separating philosophy from science. The Anthropic Principle is undoing this by mixing them up again.
2. The Anthropic Principle is a form of teleological reasoning that Darwin overthrew.
3. The Anthropic Principle is not testable, therefore it is not scientific. It is a quasi-religious principle.
4. The Anthropic Principle is like the 'God of the Gaps'. With every new discovery that explains a previously unexplained large-scale property of the Universe, the need for the Anthropic Principle shrinks. Inflation explains most of the cosmological properties in a more attractive way.
5. The Anthropic Principle is an inappropriate methodology. Particle physics offers the prospect of a theory of everything in which the structure of the Universe, including all the values of its physical constants, will be determined uniquely and completely. The possibility of life evolving in the Universe will be built into these laws from the beginning. Life is only a consequence of the laws of Nature.
6. The Anthropic Principle makes statements of comparative reference to other hypothetical universes. We know of, and can know of, only one Universe.
7. The Anthropic Principle appeals to the Many Worlds interpretation of quantum mechanics, but we can never test that the other quantum worlds exist.
8. The Anthropic Principle takes a parochial view of life, and assumes that all life-forms in the Universe resemble ourselves.

The most common misconception regarding the Anthropic Principle, which features in the quotation given above is that it is in some sense a rival cosmological or particle physics theory which one is being offered as an alternative to the standard picture. This is rather misleading. All that is being claimed is that the Anthropic Principle must be used as a *complement* to the standard deductive theories, otherwise there is a real danger of drawing erroneous conclusions or, more commonly, providing elaborate 'explanations' for non-existent problems. A classic example is Dicke's demonstration that Dirac's Large Number Coincidences do not require any extreme hypothesis, like the time-variation of Newton's gravitation constant, to explain them. The Weak Anthropic Principle does not explain them, but it shows that a posteriori they are not surprising. A discovery that the gravitation constant was decreasing with time in the way predicted by Dirac would falsify the anthropic explanation for these coincidences. An anthropic explanation can be ruled out by observation.

The first three objections are all of a piece. We have become so indoctrinated by the philosophers of science, with their paradigms and exemplars, their emphases upon falsification and verification, that we can easily lose sight of the fact that they are methodological principles for the expedient *practice* of science. They need have nothing whatsoever to do with whether particular theories are actually true or false. If someone writes down the correct 'theory of everything', then that will not be falsifiable either. To believe that we will be able to test and falsify all theories is just the sort of anthropocentric view of the Universe that critics of the Anthropic Principle so roundly decry elsewhere. Why should Nature be constructed upon a scale that is spanned by human intelligence? Why should what is true also be humanly falsifiable or verifiable?

The fundamental questions of cosmology and particle physics are of a very special type. Any explanation for the origin and structure of the Universe is likely to be of a very unusual sort. We would be foolish to discard certain approaches to these problems simply because they do not have analogues in more mundane scientific investigation. It is surely right to study Nature with the confident assumption that it can be fully understood, but it is not correct to reject ideas because they do not fit in with the religious view that Pagels puts forward. Needless to say, the Weak Anthropic Principle does not claim that the Universe was constructed especially for life, human or otherwise, only that the existence of life may need to be included in a right evaluation of its global properties.

Objections 4 and 5 are the most interesting. Let us consider first the question of a 'theory of everything'. This is topical at present because of the impetus provided by superstring theories. Of course, the existence of a theory of everything is just an assumption. There is no evidence for it: it is a

philosophical view that is mixed into science (compare Objection 1). Nevertheless, it is a reasonable one to entertain. But what is not reasonable to entertain is that a knowledge of laws of Nature in their unified entirety will suffice to provide a complete explanation for the structure of the Universe and our own evolution. Even if the laws of Nature are found to be uniquely determined, the solutions of those laws may not be. We know from our experience with particle physics that solutions of equations need not possess the same symmetries as the equations themselves. Even in the presence of a theory of everything it is quite reasonable to entertain the view that there will exist quasi-random elements in the Universe's structure, and even to some extent in its laws and 'almost' symmetries. This means that, even when in possession of the complete set of equations governing the evolution of the visible Universe and the logically determined values of its fundamental constants, we will not be able to predict the structure of the actual Universe uniquely, any more than we can predict the direction in which the Earth is spinning from the law of angular momentum conservation.

One must also regard as speculation the assumption that a theory of everything will provide a set of unique initial conditions for the evolution of the Universe. Such a provision seems more unlikely still if the Universe had no beginning (so 'initial' conditions are set at past temporal infinity), or has tunnelled from some earlier quantum state, in which case we could have no more than a *probability* that any particular final state arises. In all these pictures where there is an element of chance in the gross structure of the Universe, it is quite possible a priori for the Universe to expand into a state that cannot evolve and support carbon-based life-forms. A correct explanation for its structure and evolution could not neglect the a posteriori fact of our own evolution.

As we look out into the Universe there are some things that we do not look to a fundamental law of Nature to explain—why it is raining today, why the Earth has a moon, the number of planets in the solar system, the number of galaxies in the Local Group of galaxies. These are chance events, in that they could have been different without doing violence to the laws of Nature. In the local cosmic environment from which these examples are drawn it is relatively easy to pick out such events, but when we consider the large-scale structure of the Universe it is not clear which aspects require a fundamental explanation in terms of laws of Nature and which do not. Pagels assumes everything about the large-scale Universe requires, and has, a fundamental and unique explanation, whereas the Anthropic Principle recognizes that there may be elements of the observed structure of the Universe that are chance outcomes of particular symmetry breakings. We are able to observe the outcomes that we do only because they fell out in a

fashion that allows observers to arise subsequently. Both positions are assumptions, both employ unverified philosophical ideas, both could be wrong, or one might be right; it is too early to pass judgement with certainty.

With regard to the inflationary universe picture (see Objection 4) one can say more about the relationship with anthropic explanations. Far from being an alternative to the Anthropic Principle, in its purest form the inflationary universe actually has to employ the Anthropic Principle. Inflation assumes chaotic initial conditions in the Universe, and each local microscopic region then inflates by an amount determined by the degree of microscopic smoothness within it. Some regions inflate a lot, some only a little. The result is that the Universe ends up divided into domains which have very different conditions. We have to live in a domain that has inflated to at least thirteen billion light-years in extent in order that life can have formed. The inflationary hypothesis could be made to fail—simply assume that no domain inflates enough to explain the large-scale uniformity of the observable universe—but of course nobody countenances such an assumption. They use the Anthropic Principle *implicitly* to deduce that at least one region must expand to large size, and that we inhabit one of those that do.

The issue of 'other worlds' also arises in inflationary explanations. Andrei Linde, one of the inventors of the current inflationary universe model, suggests that in an infinite Universe we should regard the inflationary development of a large universe like our own as inevitable, because in a randomly infinite universe there will initially exist microscopic regions in all possible states of smoothness, which will therefore result in all possible inflated states. We will inhabit one of the large ones for no other reason than the fact that a large universe is necessary for life to evolve. There is no reason at present to believe that a theory of everything will change this argument significantly.

We see from this idea that the infinite Universe can be recast into an infinite number of causally disjoint regions where different things happen. In this case the 'other worlds' are neither speculative nor mysterious. At present, in the absence of any definitive law of initial conditions, we regard the possibility of changing the possible initial conditions of the Universe as equivalent to changing the starting conditions in solutions of Einstein's equations. Each of the universes that results is regarded as a possible 'other universe'.

The Many Worlds interpretation of quantum mechanics has grown in popularity with the study of quantum cosmology. Again, it is an example of a type of theory that opponents reject basically because they do not like it, or because we might be unable to test the existence of other worlds. Clearly, if the Universe is perpetually branching every time a quantum

34

interaction occurs we are again in the situation where a fundamental theory of everything is insufficient to explain the observed structure of the Universe. All possible degrees of inflation, all possible symmetry-breakings and the values of the fundamental constants they create actually occur in reality. The branch of it that we inhabit is chosen from the entire ensemble by the fact that the necessary conditions for the evolution of life are met within it. Whereas the standard picture described above, in which there are quasi-random elements in the evolution of the Universe, has only one of the set of possibilities extant, the Many Worlds scenario has them all occurring. In defence of the Many Worlds interpretation, it is the simplest of the interpretations on offer because it uses the minimum of additional assumptions in order to explain the things that are seen.

A common reaction to the problem of the interpretation of quantum mechanics is that of the physicist who says that quantum mechanics works, and that is all that matters. The question of the *meaning* of quantum mechanics is not one that physicists should worry about. However, this is not an attitude that we are happy to adopt elsewhere. If a student comes and asks how to solve a quadratic equation, and says he just wants to know the formula that extracts the solution but he does not want to know why it works or where it comes from, we would take a very dim view of that student. The whole scientific enterprise is based upon rejection of the view that if it 'works' then that is good enough.

The question of whether all life-forms in the Universe have to resemble ourselves in respect of being carbon-based is an interesting one. The view of the biochemists is that only life that makes use of carbon can come into existence *spontaneously*. Thus, while in the future we may create forms of artificial silicon-based intelligence meriting the title 'life', this life is secondary: it could not evolve spontaneously. In fact this issue is a red herring as far as applications of the Weak Anthropic Principle are concerned. The arguments regarding the length of time required for the stellar synthesis of carbon apply equally to the origin of silicon, nitrogen, phosphorus, oxygen, and all the heavier elements as well. One can be very sure that there are no forms of atomic life that avoid the use of elements heavier than lithium, and large, Big Bang universes are necessary for the production of all these heavier elements.

In conclusion, it is important to stress once again that the fundamental problems at the frontiers of modern cosmology and particle physics are of a unique type. They are not like the problems of laboratory physics. They are not problems which always respect the traditional dogmas about the philosophy and practice of science. They are extraordinary problems, and they possess extraordinary solutions which it will require extraordinary methods to coax from the Universe. If our methods ultimately fail, then

any boundary between fundamental science and metaphysical theology will become increasingly difficult to draw. Sight must give way to faith. Confronted with an emotionally satisfying mathematical scheme which is 'simple' enough to command universal assent, but esoteric enough to admit no means of experimental test and grandiose enough to provoke no new questions then, closeted within our world within the world, we might simply have to believe it. Whereof we cannot speak thereof we must be silent: this is the final sentence of the laws of Nature.

35