THE ILLUSTRATED

A BRIEF HISTORY OF TIME

UPDATED AND EXPANDED EDITION

THE UNIVERSE IN A NUTSHELL

STEPHEN HAWKING

BANTAM BOOKS

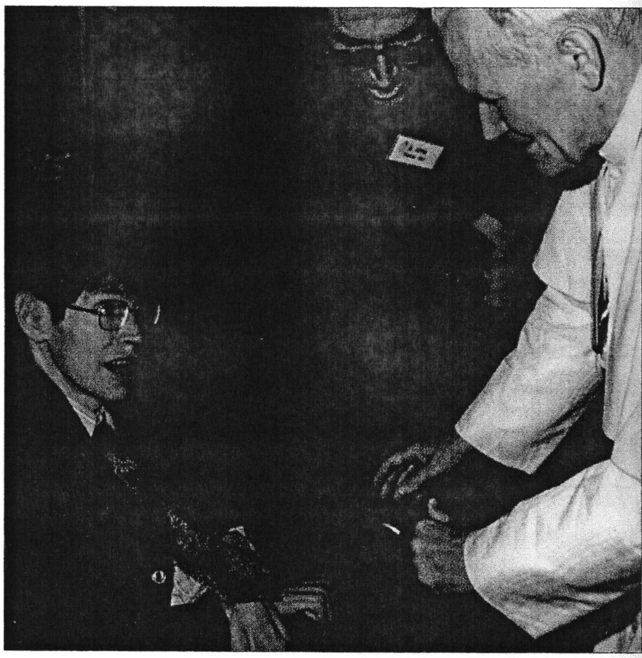
NEW YORK TORONTO LONDON SYDNEY AUCKLAND

8

The Origin and Fate of the Universe

EINSTEIN'S GENERAL THEORY of relativity, on its own, predicted that space-time began at the big bang singularity and would come to an end either at the big crunch singularity (if the whole universe recollapsed), or at a singularity inside a black hole (if a local region, such as a star, were to collapse). Any matter that fell into the hole would be destroyed at the singularity, and only the gravitational effect of its mass would continue to be felt outside. On the other hand, when quantum effects were taken into account, it seemed that the mass or energy of the matter would eventually be returned to the rest of the universe, and that the black hole, along with any singularity inside it, would evaporate away and finally disappear. Could quantum mechanics have an equally dramatic effect on the big bang and big crunch singularities? What really happens during the very early or late stages of the universe, when gravitational fields are so strong that quantum effects cannot be ignored? Does the universe in fact have a beginning or an end? And if so, what are they like?

Throughout the 1970s I had been mainly studying black holes, but in 1981 my interest in



The author meeting with Pope Paul in 1981.

questions about the origin and fate of the universe was reawakened when I attended a conference on cosmology organized by the Jesuits in the Vatican. The Catholic Church had made a bad mistake with Galileo when it tried to lay down the law on a question of science, declaring that the sun went round the earth. Now, centuries later, it had decided to invite a number of experts to advise it on cosmology. At the end of the conference the participants were granted an audience with the Pope. He told us that it was all right to study the evolution of the universe after the big bang, but we should not inquire into the big bang itself because that was the moment of Creation and therefore the work of God. I was glad then that he did not know the subject of the talk I had just given at the conference — the possibility that space-time was finite but had no boundary, which means that it had no beginning, no moment of Creation. I had no desire to share the fate of Galileo, with whom I feel a strong sense of identity, partly because of the coincidence of having been born exactly 300 years after his death!

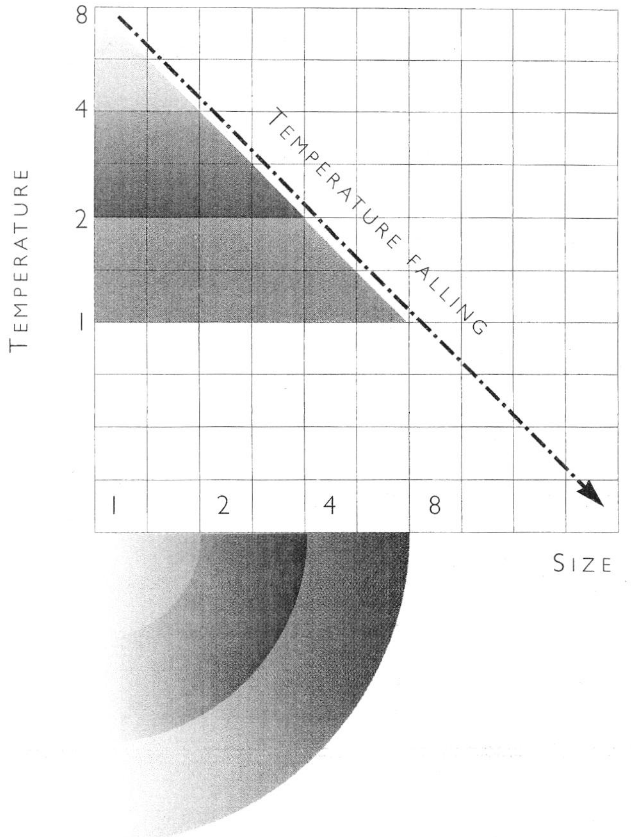
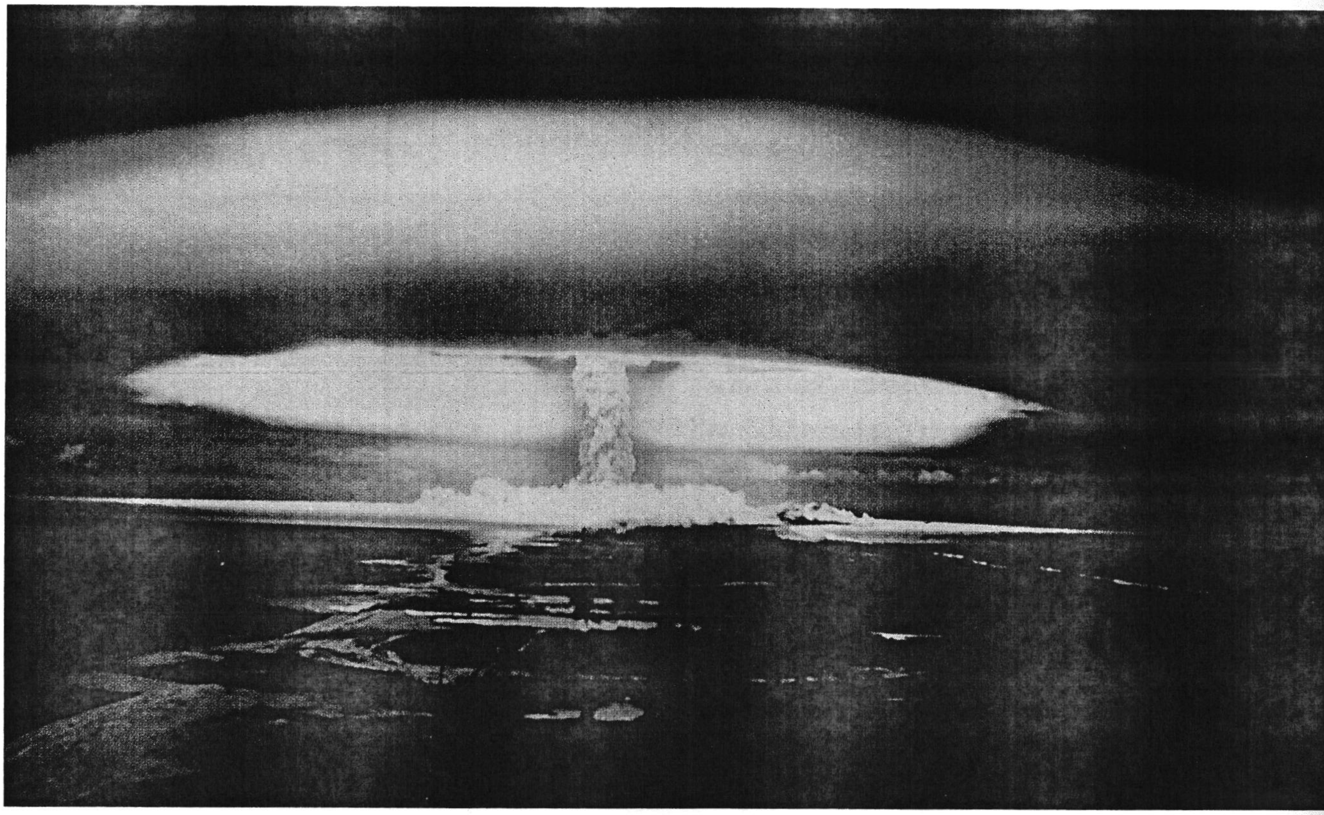


Fig. 8.1

Each time the size of the universe doubles, its temperature falls by half

In order to explain the ideas that I and other people have had about how quantum mechanics may affect the origin and fate of the universe, it is necessary first to understand the generally accepted history of the universe, according to what is known as the "hot big bang model." (See Fig. 8.2 on page 148.) This assumes that the universe is described by a Friedmann model, right back to the big bang. In such models one



Nuclear bomb test at Bikini Atoll, 1954. At the heart of an atom bomb explosion, man can create temperatures of some ten thousand million degrees. This is comparable to the temperature of the universe one second after the big bang.

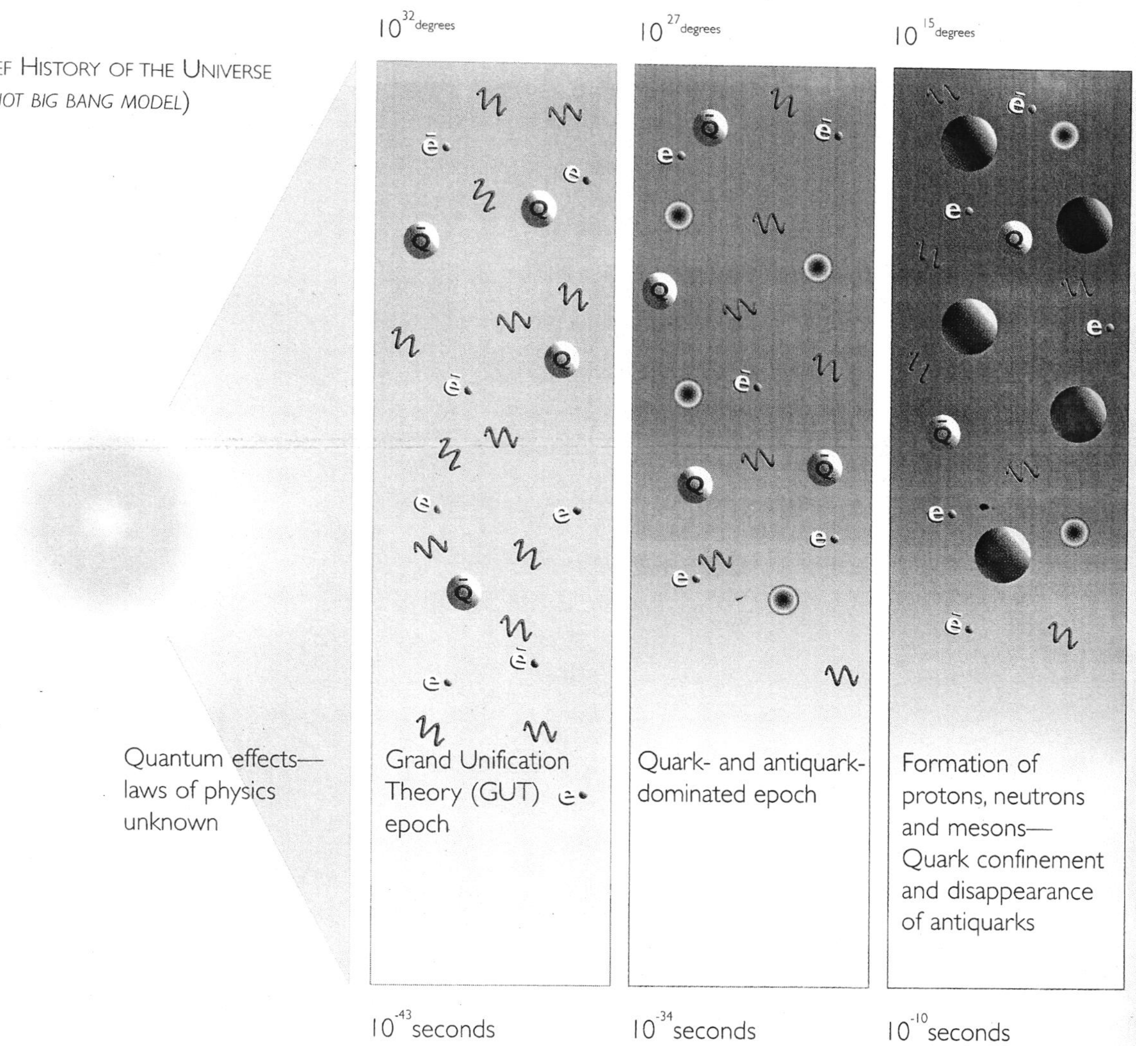
finds that as the universe expands, any matter or radiation in it gets cooler. (When the universe doubles in size, its temperature falls by half. See Fig. 8.1.) Since temperature is simply a measure of the average energy — or speed — of the particles, this cooling of the universe would have a major effect on the matter in it. At very high temperatures, particles would be moving around so fast that they could escape any attraction toward each other due to nuclear or electromagnetic forces, but as they cooled off one would expect particles that attract each other to start to clump together. Moreover, even the types of particles that exist in the universe would depend on the temperature. At high enough temperatures, particles have so much energy that whenever they collide many differ-

George Gamow emerges in this collage as the genie of a bottle of "Ylem," the hypothetical primordial material of the big bang. It was Gamow and Ralph Alpher, who also appears in this picture, who first proposed that the universe had a very hot early stage.

ent particle/antiparticle pairs would be produced — and although some of these particles would annihilate on hitting antiparticles, they would be produced more rapidly than they could annihilate. At lower temperatures, however, when colliding particles have less energy, particle/antiparticle pairs would be produced less quickly — and annihilation would become faster than production.

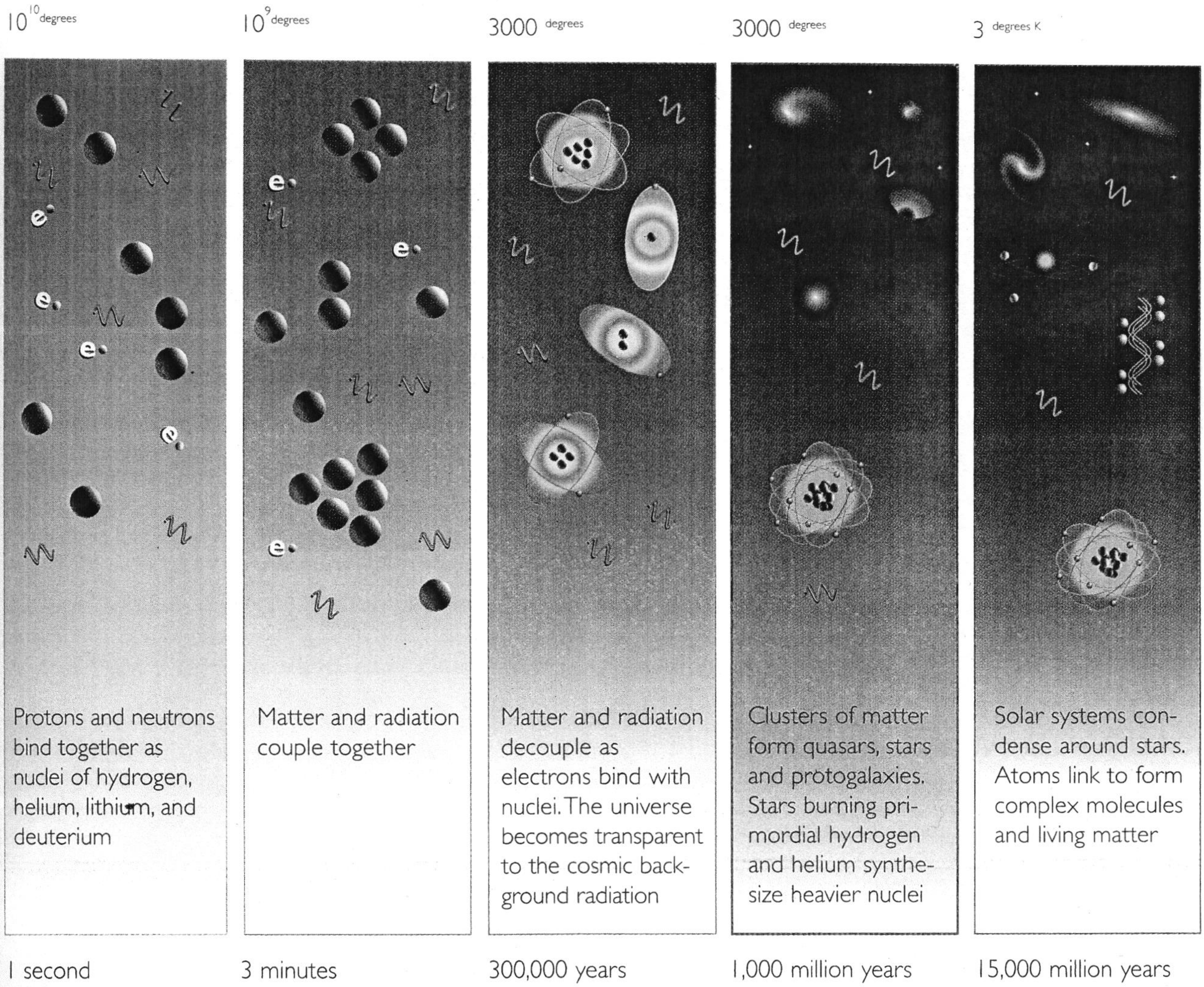
At the big bang itself the universe is thought to have had zero size, and so to have been infinitely hot. But as the universe expanded, the temperature of the radiation decreased. One second after the big bang, it would have fallen to about ten thousand million degrees. This is about a thousand times the temperature at the center of the sun, but temperatures as high as this are reached in H-bomb explosions. At this time the universe would have contained mostly photons, electrons, and neutrinos (extremely light particles that are affected only by the weak force and gravity) and their antiparticles, together with some protons and neutrons. As the universe continued to expand and the temperature to drop, the rate at which electron/antielectron pairs were being produced in collisions would have fallen below the rate at which they were being destroyed by annihilation. So most of the electrons and antielectrons would have annihilated with each other to produce more photons, leaving only a few electrons left over. The neutrinos and antineutrinos, however, would not have annihilated with each other, because these particles interact with themselves and with other particles only very weakly. So they should still be around today. If we could observe them, it would provide a good test of this picture of a very hot early stage of the universe. Unfortunately, their energies nowadays would be too low for us to observe them direct-

Fig. 8.2 A BRIEF HISTORY OF THE UNIVERSE (THE HOT BIG BANG MODEL)



ly. However, if neutrinos are not massless, but have a small mass of their own, as suggested by some recent experiments, we might be able to detect them indirectly: they could be a form of "dark matter," like that mentioned earlier, with sufficient gravitational attraction to stop the expansion of the universe and cause it to collapse again.

About one hundred seconds after the big bang, the temperature would have fallen to one thousand million degrees, the temperature inside the hottest stars. At this temperature protons



and neutrons would no longer have sufficient energy to escape the attraction of the strong nuclear force, and would have started to combine together to produce the nuclei of atoms of deuterium (heavy hydrogen), which contain one proton and one neutron. The deuterium nuclei would then have combined with more protons and neutrons to make helium nuclei, which contain two protons and two neutrons, and also small amounts of a couple of heavier elements, lithium and beryllium. One can calculate that in the hot big bang model about a quarter of the protons and neutrons would have been converted into helium nuclei, along with a small amount of heavy hydrogen and other elements. The remaining neutrons would have decayed into protons, which are the nuclei of ordinary hydrogen atoms.

This picture of a hot early stage of the universe was first put forward by the scientist George Gamow in a famous paper written in 1948 with a student of his, Ralph Alpher. Gamow had quite a sense of humor — he persuaded the nuclear scientist Hans Bethe to add his name to the paper to make the list of authors "Alpher, Bethe, Gamow," like the first three letters of the Greek alphabet, alpha, beta, gamma: particularly appropriate for a paper on the beginning of the universe! In this paper they made the remarkable prediction that radiation (in the form of photons) from the very hot early stages of the universe should still be around today, but with its temperature reduced to only a few degrees above absolute zero (-273°C). It was this radiation that Penzias and Wilson found in 1965. At the time that Alpher, Bethe, and Gamow wrote their paper, not much was

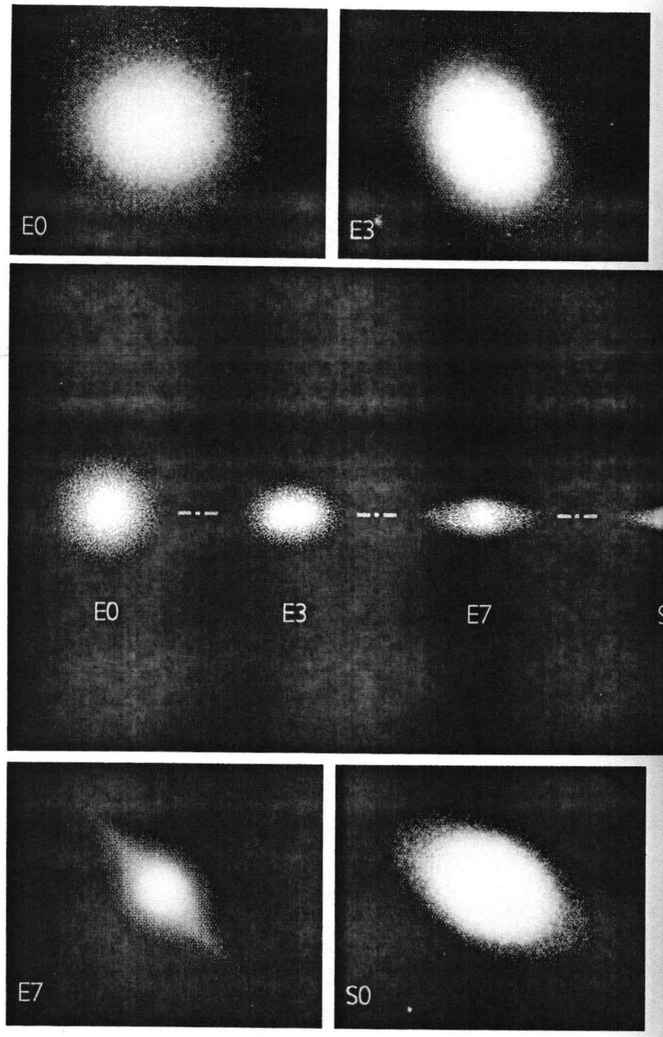
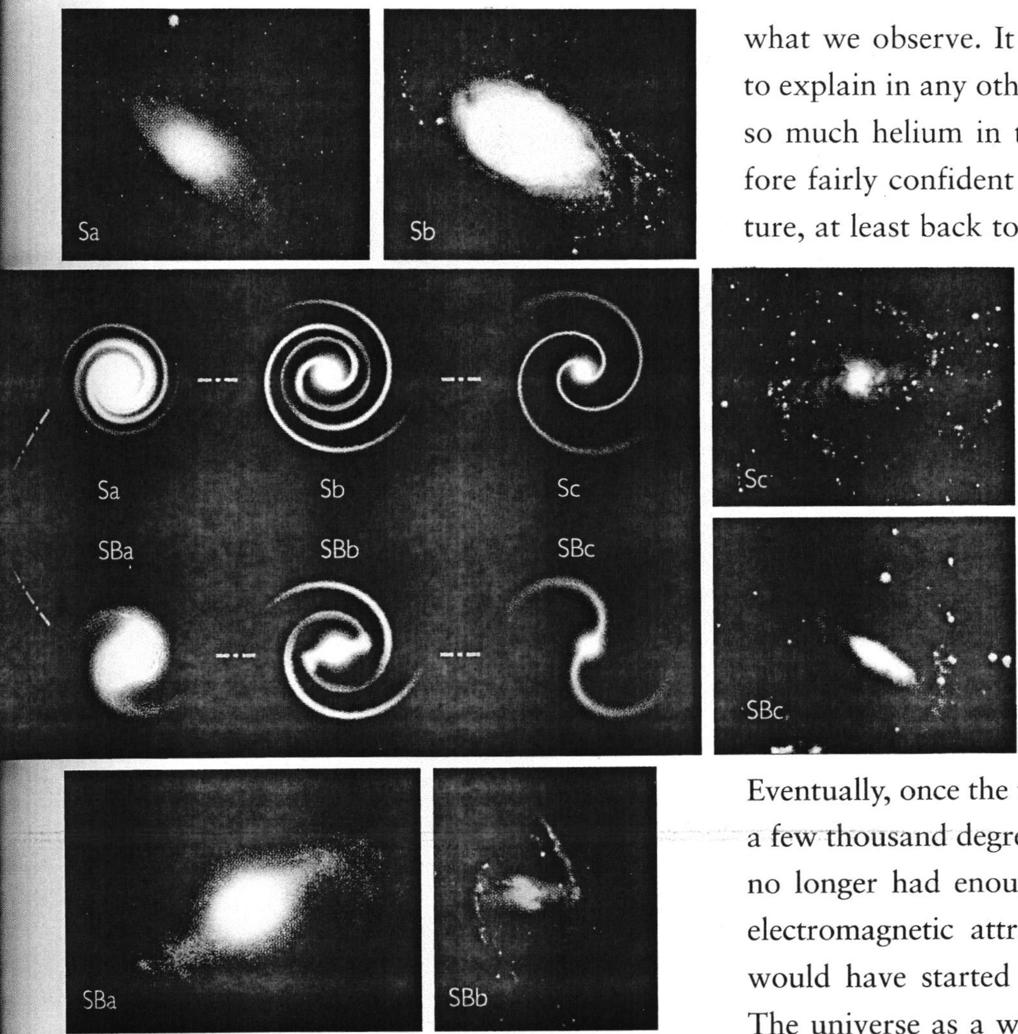
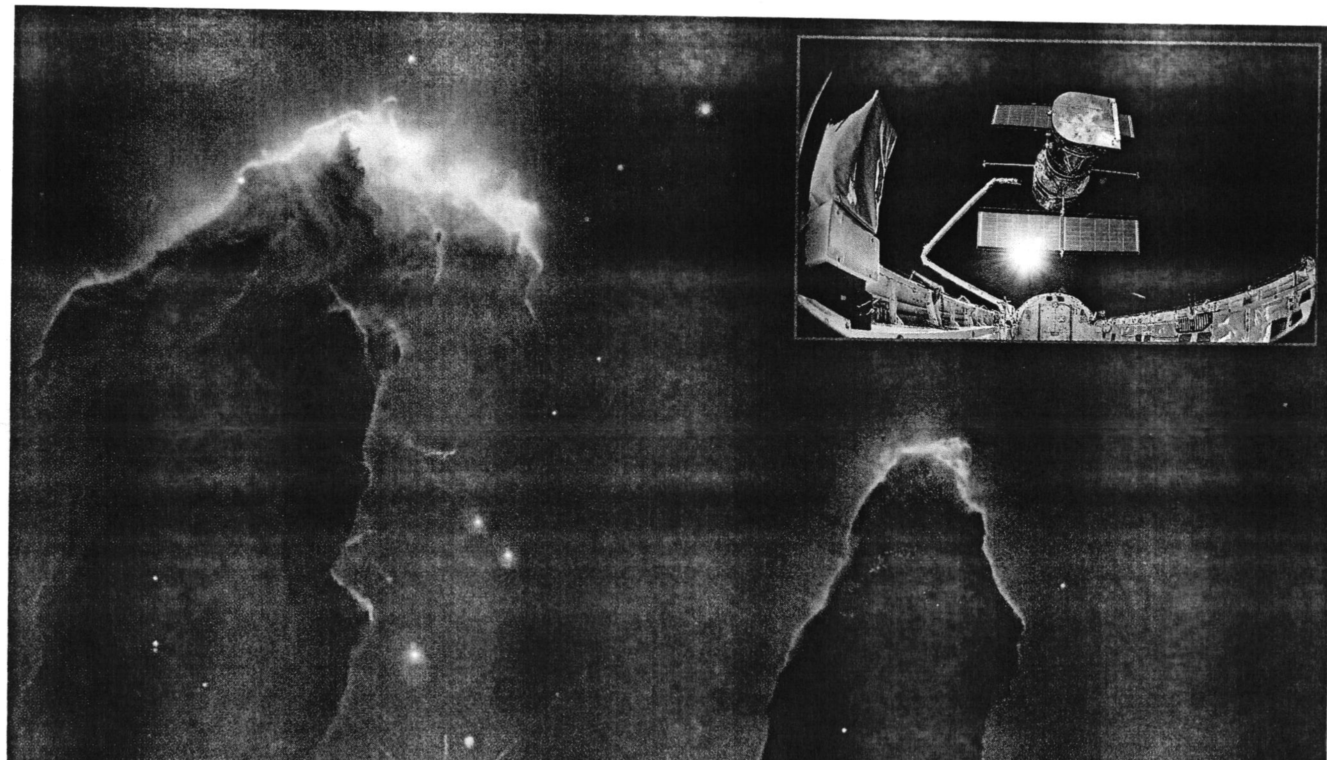


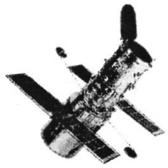
Fig. 8.3 *A revised version of the galactic classification system proposed by Edwin Hubble and Milton Humason in 1936. On the left are the four featureless elliptical, non-rotating systems, E0, E3, E7, and S0. The upper group on the right shows the spiral galaxies, Sa, Sb, and Sc, while those below are the barred spirals, SBa, SBb, SBc. The three categories a, b, and c in each group denote the central nuclear regions becoming smaller while the galactic arms become wider and more open.*



known about the nuclear reactions of protons and neutrons. Predictions made for the proportions of various elements in the early universe were therefore rather inaccurate, but these calculations have been repeated in the light of better knowledge and now agree very well with what we observe. It is, moreover, very difficult to explain in any other way why there should be so much helium in the universe. We are therefore fairly confident that we have the right picture, at least back to about one second after the big bang.

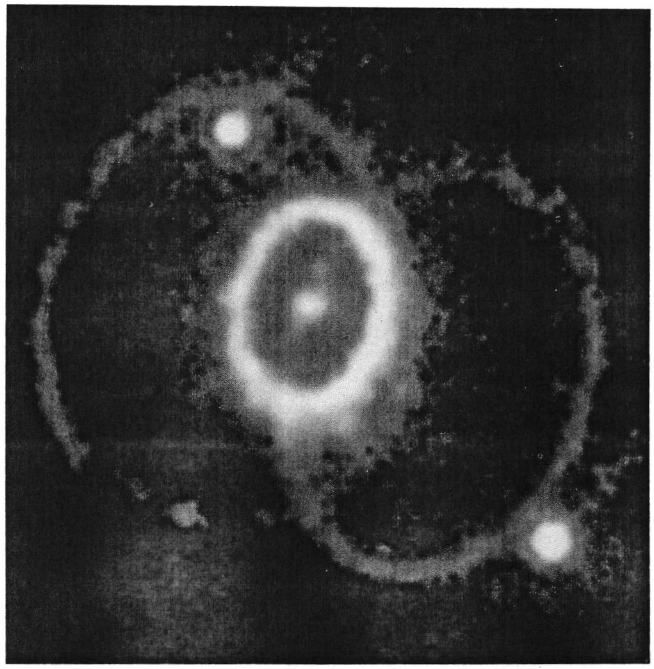
Within only a few hours of the big bang, the production of helium and other elements would have stopped. And after that, for the next million years or so, the universe would have just continued expanding, without anything much happening. Eventually, once the temperature had dropped to a few thousand degrees, and electrons and nuclei no longer had enough energy to overcome the electromagnetic attraction between them, they would have started combining to form atoms. The universe as a whole would have continued expanding and cooling, but in regions that were slightly denser than average, the expansion would have been slowed down by the extra gravitational attraction. This would eventually stop expansion in some regions and cause them to start to recollapse. As they were collapsing,





the gravitational pull of matter outside these regions might start them rotating slightly. As the collapsing region got smaller, it would spin faster — just as skaters spinning on ice spin faster as they draw in their arms. Eventually, when the region got small enough, it would be spinning fast enough to balance the attraction of gravity, and in this way disklike rotating galaxies were born (Fig. 8.3). Other regions, which did not happen to pick up a rotation, would become oval-shaped objects called elliptical galaxies. In these, the region would stop collapsing because individual parts of the galaxy would be orbiting stably round its center, but the galaxy would have no overall rotation.

As time went on, the hydrogen and helium gas in the galaxies would break up into smaller clouds that would collapse under their own gravity. As these contracted, and the atoms within them collided with one another, the temperature of the gas would increase, until eventually it became hot enough to start nuclear fusion reactions. These would convert the hydrogen into more helium, and the heat given off would raise the pressure, and so stop the clouds from contracting any further. They would remain stable in this state for a long time as stars like our sun, burning hydrogen into helium and radiating the resulting energy as heat and light. More massive stars would need to be hotter to balance



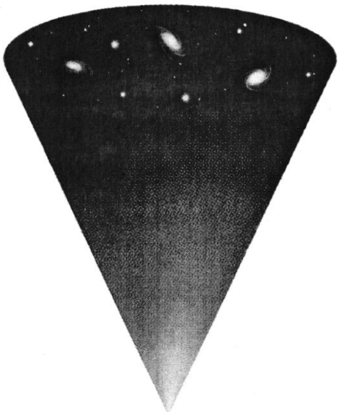
Above: *The aftermath of supernova 1987a. The central ring is an expanding doughnut of material blown off by the explosion. The central spot is a new neutron star.* Opposite: *New stars being born inside clouds of dust and gas in the Eagle Cluster. Both of these photos were taken by the Hubble Space Telescope, seen being repaired in orbit (inset).*

their stronger gravitational attraction, making the nuclear fusion reactions proceed so much more rapidly that they would use up their hydrogen in as little as a hundred million years. They would then contract slightly, and as they heated up further, would start to convert helium into heavier elements like carbon or oxygen. This, however, would not release much more energy, so a crisis would occur, as was described in the chapter on black holes. What happens next is not completely clear, but it seems likely that the central regions of the star would collapse to a very dense state, such as a neutron star or black hole. The outer regions of the star may sometimes get blown off in a tremendous explosion called a supernova, which would outshine all the other stars in its galaxy. Some of the heavier elements produced near the end of the star's life would be flung back into the gas in the galaxy, and would provide some of the raw material for the next generation of stars. Our own sun contains about 2 percent of these heavier elements because it is a second- or third-generation star, formed some five thousand million years ago out of a cloud of rotating gas containing the debris of earlier supernovas. Most of the gas in that cloud went to form the sun or got blown away, but a small amount of the heavier elements collected together to form the bodies that now orbit the sun as planets like the earth.

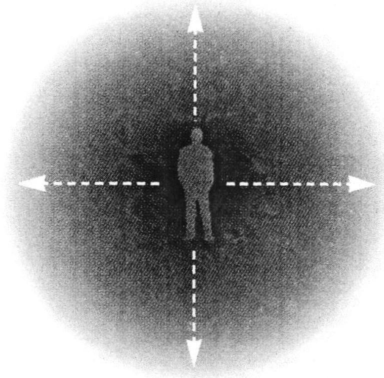
The earth was initially very hot and without an atmosphere. In the course of time it cooled and acquired an atmosphere from the emission of gases from the rocks. This early atmosphere was not one in which we could have survived. It contained no oxygen, but a lot of other gases that are poisonous to us, such as hydrogen sulfide (the gas that gives rotten eggs their smell). There are, however, other primitive forms of life that can flourish under such conditions. It is thought that they developed in the oceans, possibly as a result of chance combinations of atoms into large structures, called macromolecules, which were capable of assembling other atoms in the ocean into similar structures. They would thus have reproduced themselves and multiplied. In some cases there would be errors in the reproduction. Mostly these errors would have been such that the new macromolecule could not reproduce itself and eventually would have been destroyed. However, a few of the errors would have produced new macromolecules that were even better at reproducing themselves. They would have therefore had an advantage and would have tended to replace the original macromolecules. In this way a process of evolution was started that led to the development of more and more complicated, self-reproducing organisms. The first primitive forms of life consumed various materials, including hydrogen sulfide, and released oxygen. This gradually changed the atmosphere to the composition that it has today, and allowed the development of higher forms of life such as fish, reptiles, mammals, and ultimately the human race.

This picture of a universe that started off very hot and cooled as it expanded is in agree-

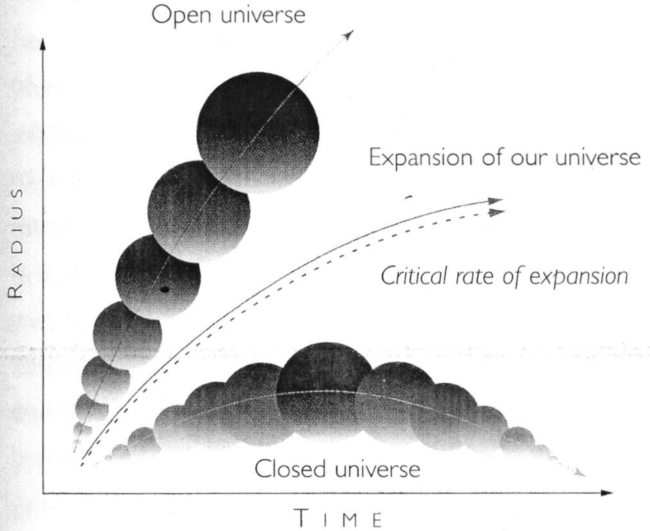
Fig. 8.4



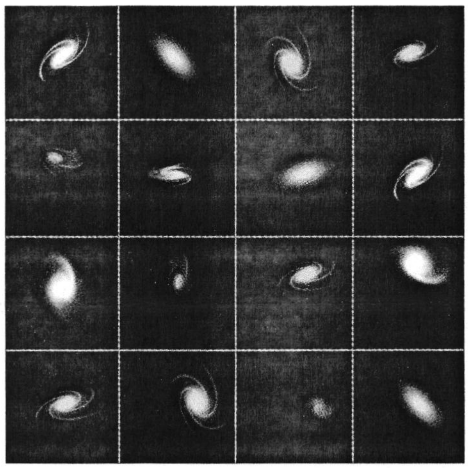
The universe started very hot



The temperature of the microwave background is almost exactly the same in all directions



The universe is on a knife edge between continuous expansion and recollapse

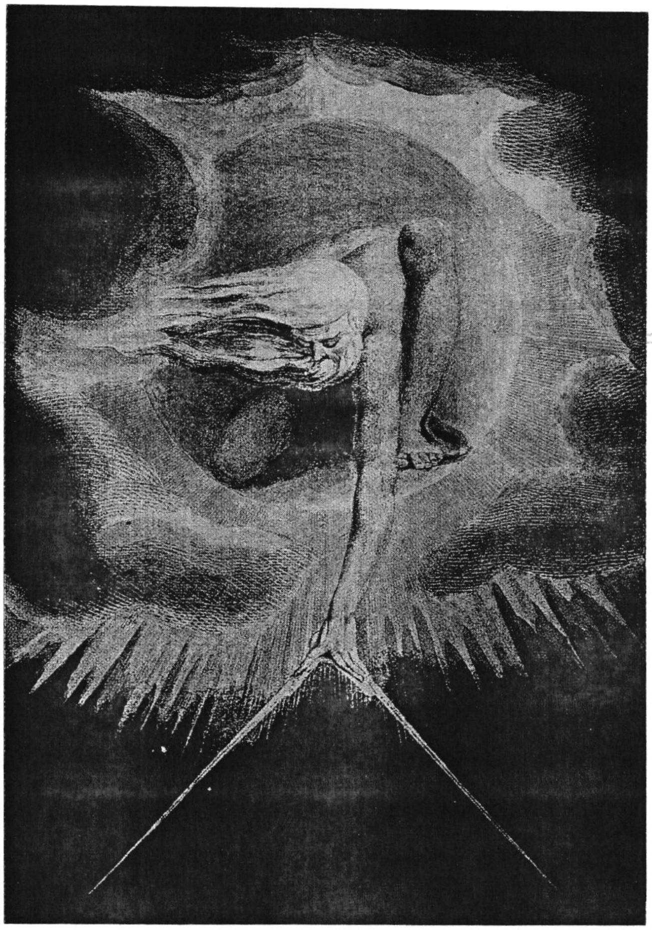


Small fluctuations in the density of the universe give rise to galaxies and stars

ment with all the observational evidence that we have today. Nevertheless, it leaves a number of important questions unanswered (Fig. 8.4):

(1) Why was the early universe so hot?

(2) Why is the universe so uniform on a large scale? Why does it look the same at all points of space and in all directions? In particular, why is the temperature of the microwave background

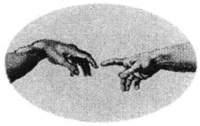


"The Ancient of Days" by William Blake (1757–1827).

radiation so nearly the same when we look in different directions? It is a bit like asking a number of students an exam question. If they all give exactly the same answer, you can be pretty sure they have communicated with each other. Yet, in the model described above, there would not have been time since the big bang for light to get from one distant region to another, even though the regions were close together in the early universe. According to the theory of relativity, if light cannot get from one region to another, no other information can. So there would be no way in which different regions in the early universe could have come to have had the same temperature as each other, unless for some unexplained reason they happened to start out with the same temperature.

(3) Why did the universe start out with so nearly the critical rate of expansion that separates models that recollapse from those that go on expanding forever, that even now, ten thousand million years later, it is still expanding at nearly the critical rate? If the rate of expansion one second after the big bang had been smaller by even one part in a hundred thousand million million, the universe would have recollapsed before it ever reached its present size.

(4) Despite the fact that the universe is so uniform and homogeneous on a large scale, it contains local irregularities, such as stars and galaxies. These are thought to have developed from small differences in the density of the early universe from one region to another. What was the origin of these density fluctuations?

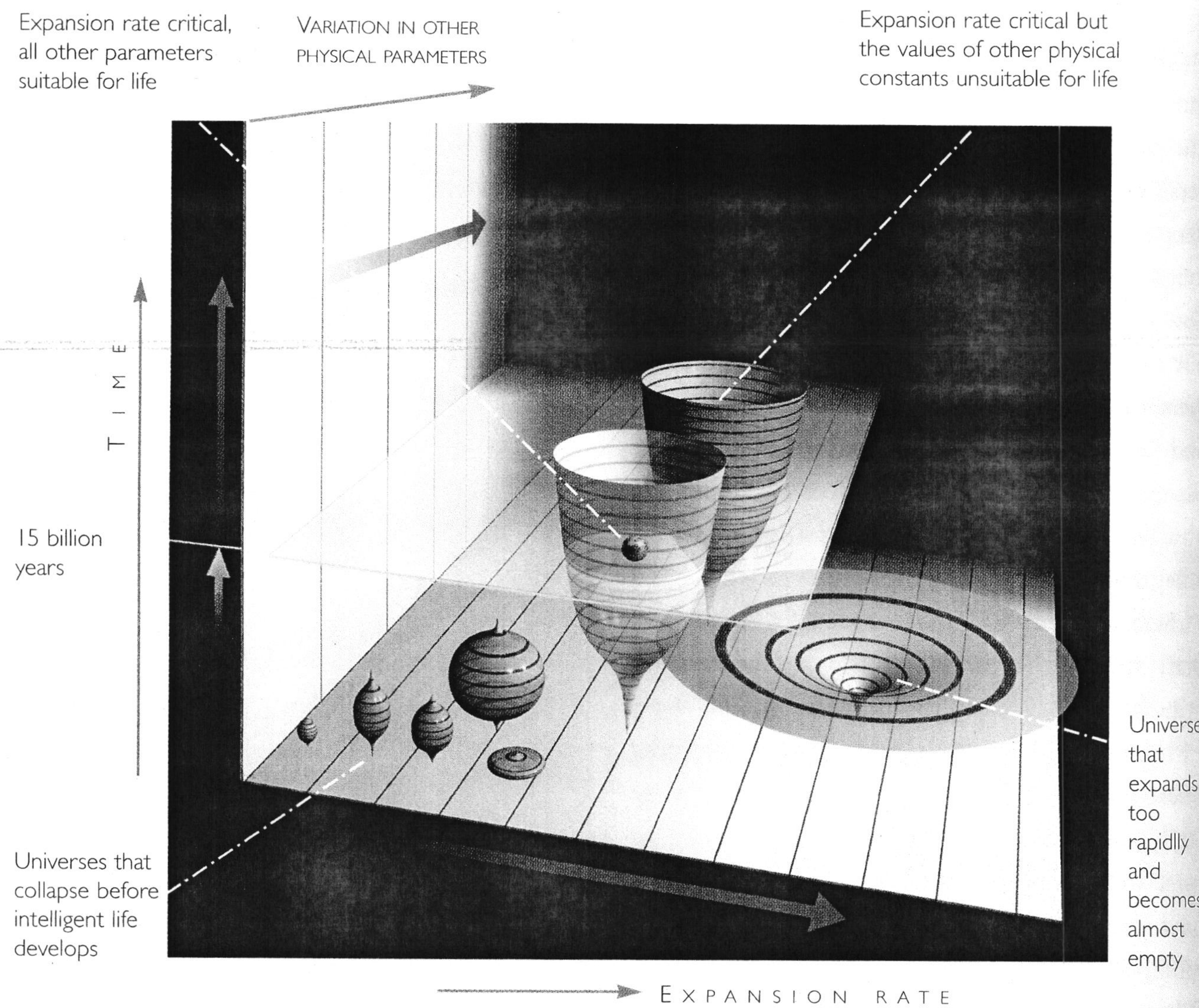


The general theory of relativity, on its own, cannot explain these features or answer these questions because of its prediction that the universe started off with infinite density at the big bang singularity. At the singularity, general relativity and all other physical laws would break down: one couldn't predict what would come out of the singularity. As explained before, this means that one might as well cut the big bang, and any events before it, out of the theory, because they can have no effect on what we observe. Space-time *would* have a boundary — a beginning at the big bang.

Science seems to have uncovered a set of laws that, within the limits set by the uncertainty principle, tell us how the universe will develop with time, if we know its state at any one time. These laws may have originally been decreed by God, but it appears that he has since left the universe to evolve according to them and does not now intervene in it. But how did he choose the initial state or configuration of the universe? What were the "boundary conditions" at the beginning of time?

One possible answer is to say that God chose the initial configuration of the universe for reasons that we cannot hope to understand. This would certainly have been within the power of an omnipotent being, but if he had started it off in such an incomprehensible way, why did he choose to let it evolve according to laws that we could understand? The whole history of science has been the gradual realization that events do not happen in an arbitrary manner, but that they reflect a certain underlying order, which may or may not be divinely inspired. It would be only natural to suppose that this order should apply not only to the laws, but also to the conditions at the boundary of space-time that specify the initial state of the universe. There may be a large number of models of the universe with different initial conditions that all obey the laws. There ought to be some principle that picks out one initial state, and hence one model, to represent our universe.

One such possibility is what are called chaotic boundary conditions. These implicitly assume either that the universe is spatially infinite or that there are infinitely many universes. Under chaotic boundary conditions, the probability of finding any particular region of space in any given configuration just after the big bang is the same, in some sense, as the probability of finding it in any other configuration: the initial state of the universe is chosen purely randomly. This would mean that the early universe would have



probably been very chaotic and irregular because there are many more chaotic and disordered configurations for the universe than there are smooth and ordered ones. (If each configu-

Fig. 8.5 *The strong anthropic principle supposes that there are many different universes with different initial expansion rates and other fundamental physical attributes. Only a few are suitable for life.*

ration is equally probable, it is likely that the universe started out in a chaotic and disordered state, simply because there are so many more of them.) It is difficult to see how such chaotic initial conditions could have given rise to a universe that is so smooth and regular on a large scale as ours is today. One would also have expected the density fluctuations in such a model to have led to the formation of many more primordial black holes than the upper limit that has been set by observations of the gamma ray background.

If the universe is indeed spatially infinite, or if there are infinitely many universes, there would probably be some large regions somewhere that started out in a smooth and uniform manner. It is a bit like the well-known horde of monkeys hammering away on typewriters — most of what they write will be garbage, but very occasionally by pure chance they will type out one of Shakespeare's sonnets. Similarly, in the case of the universe, could it be that we are living in a region that just happens by chance to be smooth and uniform? At first sight this might seem very improbable, because such smooth regions would be heavily outnumbered by chaotic and irregular regions. However, suppose that only in the smooth regions were galaxies and stars formed and were conditions right for the development of complicated self-replicating organisms like ourselves who were capable of asking the question: why is the universe so smooth? This is an example of the application of what is known as the anthropic principle, which can be paraphrased as "We see the universe the way it is because we exist."

There are two versions of the anthropic principle, the weak and the strong. The weak anthropic principle states that in a universe that is large or infinite in space and/or time, the conditions necessary for the development of intelligent life will be met only in certain regions that are limited in space and time. The intelligent beings in these regions should therefore not be surprised if they observe that their locality in the universe satisfies the conditions that are necessary for their existence. It is a bit like a rich person living in a wealthy neighborhood not seeing any poverty.

One example of the use of the weak anthropic principle is to "explain" why the big bang occurred about ten thousand million years ago — it takes about that long for intelligent beings to evolve. As explained above, an early generation of stars first had to form. These stars converted some of the original hydrogen and helium into elements like carbon and oxygen, out of which we are made. The stars then exploded as supernovas, and their debris went to form other stars and planets, among them those of our Solar System, which is about five thousand million years old. The first one or two thousand million years of the earth's existence were too hot for the development of anything complicated. The remaining three thousand million years or so have been taken up by the slow process of biological evolution, which has led from the simplest organisms to beings who are capable of measuring time back to the big bang.

Few people would quarrel with the validity or utility of the weak anthropic principle. Some, however, go much further and propose a strong version of the principle (Fig. 8.5). According to this theory, there are either many different universes or many different regions of a single universe, each with its own initial configuration and, perhaps, with its own set of laws of science. In most of these universes the conditions would not be right for the development of complicated organisms; only in the few universes that are like ours would intelligent beings develop and ask the question, "Why is the universe the way we see it?" The answer is then simple: if it had been different, we would not be here!

The laws of science, as we know them at present, contain many fundamental numbers, like the size of the electric charge of the electron and the ratio of the masses of the proton and the electron. We cannot, at the moment at least, predict the values of these numbers from theory — we have to find them by observation. It may be that one day we shall discover a complete unified theory that predicts them all, but it is also possible that some or all of them vary from universe to universe or within a single universe. The remarkable fact is that the values of these numbers seem to have been very finely adjusted to make possible the development of life. For example, if the electric charge of the electron had been only slightly different, stars either would have been unable to burn hydrogen and helium, or else they would not have exploded. Of course, there might be other forms of intelligent life, not dreamed of even by writers of science fiction, that did not require the light of a star like the sun or the heavier chemical elements that are made in stars and are flung back into space when the stars explode. Nevertheless, it seems clear that there are relatively few ranges of values for the numbers that would allow the development of any form of intelligent life. Most sets of values would give rise to universes that, although they might be very beautiful, would contain no one able to wonder at that beauty. One can take this either as evidence of a divine purpose in Creation and the choice of the

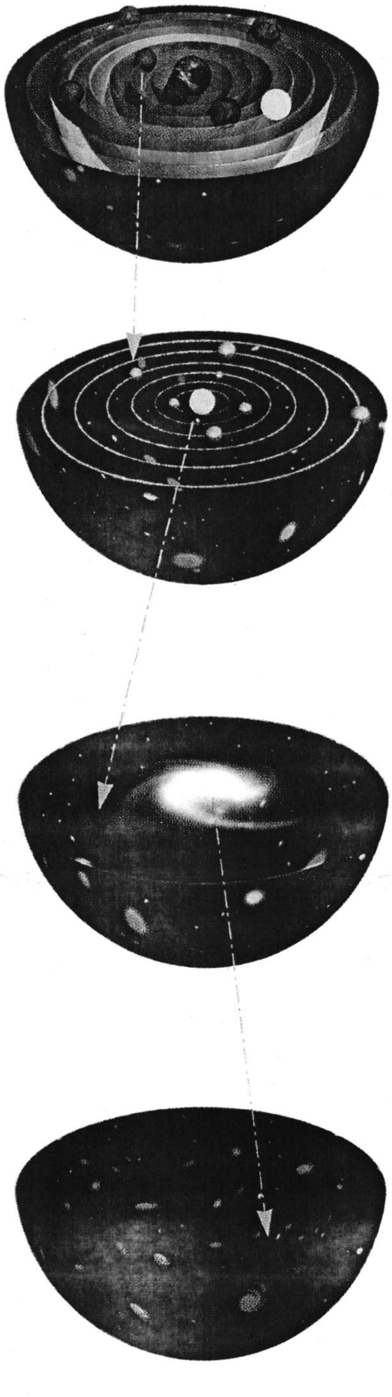


Fig. 8.6

Geocentric cosmology of Ptolemy, with Earth at the center of the universe

Heliocentric cosmology of Copernicus showing the Earth within a solar system and the stars orbiting on the outer sphere

Galactic cosmology in which Earth orbits an average star in the outer reaches of a spiral of the Milky Way

Our present picture is of the Milky Way as just one amongst one million million observable galaxies in our particular region of the universe

laws of science or as support for the strong anthropic principle.

There are a number of objections that one can raise to the strong anthropic principle as an explanation of the observed state of the universe. First, in what sense can all these different universes be said to exist? If they are really separate from each other, what happens in another universe can have no observable consequences in our own universe. We should therefore use the principle of economy and cut them out of the theory. If, on the other hand, they are just different regions of a single universe, the laws of science would have to be the same in each region, because otherwise one could not move continuously from one region to another. In this case the only difference between the regions would be their initial configurations and so the strong anthropic principle would reduce to the weak one.

A second objection to the strong anthropic principle is that it runs against the tide of the whole history of science. We have developed from the geocentric cosmologies of Ptolemy and his forebears, through the heliocentric cosmology of Copernicus and Galileo, to the modern