CHAPTER 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 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!

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." This assumes that the universe is described by a Friedmann model, right back to the big bang. In such models one 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.) 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 different particle/antiparticle pairs would be produced – and although some of these particles would annihilate on hitting antiparticles, they would be produced more rap-idly 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 directly. 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 (–273ºC). 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 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. 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 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 agreement with all the observational evidence that we have today. Nevertheless, it leaves a number of important questions unanswered:

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 back-ground 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 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 w*ould* 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 configuration 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. 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

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 picture in which the earth is a medium-sized planet orbiting around an average star in the outer suburbs of an ordinary spiral galaxy, which is itself only one of about a million million galaxies in the observable universe. Yet the strong anthropic principle would claim that this whole vast construction exists simply for our sake. This is very hard to believe. Our Solar System is certainly a prerequisite for our existence, hand one might extend this to the whole of our galaxy to allow for an earlier generation of stars that created the heavier elements. But there does not seem to be any need for all those other galaxies, nor for the universe to be so uniform and similar in every direction on the large scale.

One would feel happier about the anthropic principle, at least in its weak version, if one could show that quite a number of different initial configurations for the universe would have evolved to produce a universe like the one we observe. If this is the case, a universe that developed from some sort of random initial conditions should contain a number of regions that are smooth and uniform and are suitable for the evolution of intelligent life. On the other hand, if the initial state of the universe had to be chosen extremely carefully to lead to something like what we see around us, the universe would be unlikely to contain any region in which life would appear. In the hot big bang model described above, there was not enough time in the early universe for heat to have flowed from one region to another. This means that the initial state of the universe would have to have had exactly the same temperature everywhere in order to account for the fact that the microwave back-ground has the same temperature in every direction we look. The initial rate of expansion also would have had to be chosen very precisely for the rate of expansion still to be so close to the critical rate needed to avoid recollapse. This means that the initial state of the universe must have been very carefully chosen indeed if the hot big bang model was correct right back to the beginning of time. It would be very difficult to explain why the universe should have begun in just this way, except as the act of a God who intended to create beings like us.

In an attempt to find a model of the universe in which many different initial configurations could have evolved to something like the present universe, a scientist at the Massachusetts Institute of Technology, Alan Guth, suggested that the early universe might have gone through a period of very rapid expansion. This expansion is said to be “inflationary,” meaning that the universe at one time expanded at an increasing rate rather than the decreasing rate that it does today. According to Guth, the radius of the universe increased by a million million million million million (1 with thirty zeros after it) times in only a tiny fraction of a second.

Guth suggested that the universe started out from the big bang in a very hot, but rather chaotic, state. These high temperatures would have meant that the particles in the universe would be moving very fast and would have high energies. As we discussed earlier, one would expect that at such high temperatures the strong and weak nuclear forces and the electromagnetic force would all be unified into a single force. As the universe expanded, it would cool, and particle energies would go down. Eventually there would be what is called a phase transition and the symmetry between the forces would be broken: the strong force would become different from the weak and electromagnetic forces. One common example of a phase transition is the freezing of water when you cool it down. Liquid water is symmetrical, the same at every point and in every direction. However, when ice crystals form, they will have definite positions and will be lined up in some direction. This breaks water’s symmetry.

In the case of water, if one is careful, one can “supercool” it: that is, one can reduce the temperature below the freezing point (O°C) without ice forming. Guth suggested that the universe might behave in a similar way: the temperature might drop below the critical value without the symmetry between the forces being broken. If this happened, the universe would be in an unstable state, with more energy than if the symmetry had been broken. This special extra energy can be shown to have an antigravitational effect: it would have acted just like the cosmological constant that Einstein introduced into general relativity when he was trying to construct a static model of the universe. Since the universe would already be expanding just as in the hot big bang model, the repulsive effect of

this cosmological constant would therefore have made the universe expand at an ever-increasing rate. Even in regions where there were more matter particles than average, the gravitational attraction of the matter would have been outweighed by the repulsion of the effective cosmological constant. Thus these regions would also expand in an accelerating inflationary manner. As they expanded and the matter particles got farther apart, one would be left with an expanding universe that contained hardly any particles and was still in the supercooled state. Any irregularities in the universe would simply have been smoothed out by the expansion, as the wrinkles in a balloon are smoothed away when you blow it up. Thus the present smooth and uniform state of the universe could have evolved from many different non-uniform initial states.

In such a universe, in which the expansion was accelerated by a cosmological constant rather than slowed down by the gravitational attraction of matter, there would be enough time for light to travel from one region to another in the early universe. This could provide a solution to the problem, raised earlier, of why different regions in the early universe have the same properties. Moreover, the rate of expansion of the universe would automatically become very close to the critical rate determined by the energy density of the universe. This could then explain why the rate of expansion is still so close to the critical rate, without having to assume that the initial rate of expansion of the universe was very carefully chosen.

The idea of inflation could also explain why there is so much matter in the universe. There are something like ten million million million million million million million million million million million million million million million (1 with eighty zeros after it) particles in the region of the universe that we can observe. Where did they all come from? The answer is that, in quantum theory, particles can be created out of energy in the form of particle/antiparticle pairs. But that just raises the question of where the energy came from. The answer is that the total energy of the universe is exactly zero. The matter in the universe is made out of positive energy. However, the matter is all attracting itself by gravity. Two pieces of matter that are close to each other have less energy than the same two pieces a long way apart, because you have to expend energy to separate them against the gravitational force that is pulling them together. Thus, in a sense, the gravitational field has negative energy. In the case of a universe that is approximately uniform in space, one can show that this negative gravitational energy exactly cancels the positive energy represented by the matter. So the total energy of the universe is zero.

Now twice zero is also zero. Thus the universe can double the amount of positive matter energy and also double the negative gravitational energy without violation of the conservation of energy. This does not happen in the normal expansion of the universe in which the matter energy density goes down as the universe gets bigger. It does happen, however, in the inflationary expansion because the energy density of the supercooled state remains constant while the universe expands: when the universe doubles in size, the positive matter energy and the negative gravitational energy both double, so the total energy remains zero. During the inflationary phase, the universe increases its size by a very large amount. Thus the total amount of energy available to make particles becomes very large. As Guth has remarked, "It is said that there's no such thing as a free lunch. But the universe is the ultimate free lunch."

The universe is not expanding in an inflationary way today. Thus there has to be some mechanism that would eliminate the very large effective cosmological constant and so change the rate of expansion from an accelerated one to one that is slowed down by gravity, as we have today. In the inflationary expansion one might expect that eventually the symmetry between the forces would be broken, just as super-cooled water always freezes in the end. The extra energy of the unbroken symmetry state would then be released and would reheat the universe to a temperature just below the critical temperature for symmetry between the forces. The universe would then go on to expand and cool just like the hot big bang model, but there would now be an explanation of why the universe was expanding at exactly the critical rate and why different regions had the same temperature.

In Guth's original proposal the phase transition was supposed to occur suddenly, rather like the appearance of ice crystals in very cold water. The idea was that "bubbles" of the new phase of broken symmetry would have formed in the old phase, like bubbles of steam surrounded by boiling water. The bubbles were supposed to expand and meet up with each other until the whole universe was in the new phase. The trouble was, as I and several other people pointed out, that the universe was expanding so fast that even if the bubbles grew at the speed of light, they would be moving away from each other and so could not join up. The universe would be left in a very non-uniform state, with some regions still having symmetry between the different forces. Such a model of the universe would not correspond to what we see.

In October 1981, I went to Moscow for a conference on quantum gravity. After the conference I gave a seminar on the inflationary model and its problems at the Sternberg Astronomical Institute. Before this, I had got someone else to give my lectures for me, because most people could not understand my voice. But there was not time to prepare this seminar, so I gave it myself, with one of my graduate students repeating my words. It worked well, and gave me

much more contact with my audience. In the audience was a young Russian, Andrei Linde, from the Lebedev Institute in Moscow. He said that the difficulty with the bubbles not joining up could be avoided if the bubbles were so big that our region of the universe is all contained inside a single bubble. In order for this to work, the change from symmetry to broken symmetry must have taken place very slowly inside the bubble, but this is quite possible according to grand unified theories. Linde's idea of a slow breaking of symmetry was very good, but I later realized that his bubbles would have to have been bigger than the size of the universe at the time! I showed that instead the symmetry would have broken everywhere at the same time, rather than just inside bubbles. This would lead to a uniform universe, as we observe. I was very excited by this idea and discussed it with one of my students, Ian Moss. As a friend of Linde's, I was rather embarrassed, however, when I was later sent his paper by a scientific journal and asked whether it was suitable for publication. I replied that there was this flaw about the bubbles being bigger than the universe, but that the basic idea of a slow breaking of symmetry was very good. I recommended that the paper ¿ published as it was because it would take Linde several months to correct it, since anything he sent to the West would have to be passed by Soviet censorship, which was neither very skillful nor very quick with scientific papers. Instead, I wrote a short paper with Ian Moss in the same journal in which we pointed out this problem with the bubble and showed how it could be resolved.

The day after I got back from Moscow I set out for Philadelphia, where I was due to receive a medal from the Franklin Institute. My secretary, Judy Fella, had used her not inconsiderable charm to persuade British Airways to give herself and me free seats on a Concorde as a publicity venture. However, I .was held up on my way to the airport by heavy rain and I missed the plane. Nevertheless, I got to Philadelphia in the end and received my medal. I was then asked to give a seminar on the inflationary universe at Drexel University in Philadelphia. I gave the same seminar about the problems of the inflationary universe, just as in Moscow.

A very similar idea to Linde's was put forth independently a few months later by Paul Steinhardt and Andreas Albrecht of the University of Pennsylvania. They are now given joint credit with Linde for what is called "the new inflationary model," based on the idea of a slow breaking of symmetry. (The old inflationary model was Guth's original suggestion of fast symmetry breaking with the formation of bubbles.)

The new inflationary model was a good attempt to explain why the universe is the way it is. However, I and several other people showed that, at least in its original form, it predicted much greater variations in the temperature of the microwave background radiation than are observed. Later work has also cast doubt on whether there could be a phase transition in the very early universe of the kind required. In my personal opinion, the new inflationary model is now dead as a scientific theory, although a lot of people do not seem to have heard of its demise and are still writing papers as if it were viable. A better model, called the chaotic inflationary model, was put forward by Linde in 1983. In this there is no phase transition or supercooling. Instead, there is a spin 0 field, which, because of quantum fluctuations, would have large values in some regions of the early universe. The energy of the field in those regions would behave like a cosmological constant. It would have a repulsive gravitational effect, and thus make those regions expand in an inflationary manner. As they expanded, the energy of the field in them would slowly decrease until the inflationary expansion changed to an expansion like that in the hot big bang model. One of these regions would become what we now see as the observable universe. This model has all the advantages of the earlier inflationary models, but it does not depend on a dubious phase transition, and it can moreover give a reasonable size for the fluctuations in the temperature of the microwave background that agrees with observation.

This work on inflationary models showed that the present state of the universe could have arisen from quite a large number of different initial configurations. This is important, because it shows that the initial state of the part of the universe that we inhabit did not have to be chosen with great care. So we may, if we wish, use the weak anthropic principle to explain why the universe looks the way it does now. It cannot be the case, however, that every initial configuration would have led to a universe like the one we observe. One can show this by considering a very different state for the universe at the present time, say, a very lumpy and irregular one. One could use the laws of science to evolve the universe back in time to determine its configuration at earlier times. According to the singularity theorems of classical general relativity, there would still have been a big bang singularity. If you evolve such a universe forward in time according to the laws of science, you will end up with the lumpy and irregular state you started with. Thus there must have been initial configurations that would not have given rise to a universe like the one we see today. So even the inflationary model does not tell us why the initial configuration was not such as to produce something very different from what we observe. Must we turn to the anthropic principle for an explanation? Was it all just a lucky chance? That would seem a counsel of despair, a negation of all our hopes of understanding the underlying order of the universe.

In order to predict how the universe should have started off, one needs laws that hold at the beginning of time. If the classical theory of general relativity was correct, the singularity theorems that Roger Penrose and I proved show that

the beginning of time would have been a point of infinite density and infinite curvature of space-time. All the known laws of science would break down at such a point. One might suppose that there were new laws that held at singularities, but it would be very difficult even to formulate such laws at such badly behaved points, and we would have no guide from observations as to what those laws might be. However, what the singularity theorems really indicate is that the gravitational field becomes so strong that quantum gravitational effects become important: classical theory is no longer a good description of the universe. So one has to use a quantum theory of gravity to discuss the very early stages of the universe. As we shall see, it is possible in the quantum theory for the ordinary laws of science to hold everywhere, including at the beginning of time: it is not necessary to postulate new laws for singularities, because there need not be any singularities in the quantum theory.

We don't yet have a complete and consistent theory that combines quantum mechanics and gravity. However, we are fairly certain of some features that such a unified theory should have. One is that it should incorporate Feynman's proposal to formulate quantum theory in terms of a sum over histories. In this approach, a particle does not have just a single history, as it would in a classical theory. Instead, it is supposed to follow every possible path in space-time, and with each of these histories there are associated a couple of numbers, one represent-ing the size of a wave and the other representing its position in the cycle (its phase). The probability that the particle, say, passes through some particular point is found by adding up the waves associated with every possible history that passes through that point. When one actually tries to perform these sums, however, one runs into severe technical problems. The only way around these is the following peculiar prescription: one must add up the waves for particle histories that are not in the "real" time that you and I experience but take place in what is called imaginary time. Imaginary time may sound like science fiction but it is in fact a well-defined mathematical concept. If we take any ordinary (or "real") number and multiply it by itself, the result is a positive number. (For example, 2 times 2 is 4, but so is – 2 times – 2.) There are, however, special numbers (called imaginary numbers) that give negative numbers when multiplied by themselves. (The one called i, when multiplied by itself, gives – 1, 2i multiplied by itself gives – 4, and so on.)

One can picture real and imaginary numbers in the following way: The real numbers can be represented by a line going from left to right, with zero in the middle, negative numbers like – 1, – 2, etc. on the left, and positive numbers, 1, 2, etc. on the right. Then imaginary numbers are represented by a line going up and down the page, with i, 2i, etc. above the middle, and – i, – 2i, etc. below. Thus imaginary numbers are in a sense numbers at right angles to ordinary real numbers.

To avoid the technical difficulties with Feynman's sum over histories, one must use imaginary time. That is to say, for the purposes of the calculation one must measure time using imaginary numbers, rather than real ones. This has an interesting effect on space-time: the distinction between time and space disappears completely. A space-time in which events have imaginary values of the time coordinate is said to be Euclidean, after the ancient Greek Euclid, who founded the study of the geometry of two-dimensional surfaces. What we now call Euclidean space-time is very similar except that it has four dimensions instead of two. In Euclidean space-time there is no difference between the time direction and directions in space. On the other hand, in real space-time, in which events are labeled by ordinary, real values of the time coordinate, it is easy to tell the difference – the time direction at all points lies within the light cone, and space directions lie outside. In any case, as far as everyday quantum mechanics is concerned, we may regard our use of imaginary time and Euclidean space-time as merely a mathematical device (or trick) to calculate answers about real space-time.

A second feature that we believe must be part of any ultimate theory is Einstein's idea that the gravitational field is represented by curved space-time: particles try to follow the nearest thing to a straight path in a curved space, but because space-time is not flat their paths appear to be bent, as if by a gravitational field. When we apply Feynman's sum over histories to Einstein's view of gravity, the analogue of the history of a particle is now a complete curved space-time that represents the history of the whole universe. To avoid the technical difficulties in actually performing the sum over histories, these curved space-times must be taken to be Euclidean. That is, time is imaginary and is indistinguishable from directions in space. To calculate the probability of finding a real space-time with some certain property, such as looking the same at every point and in every direction, one adds up the waves associated with all the histories that have that property.

In the classical theory of general relativity, there are many different possible curved space-times, each corresponding to a different initial state of the universe. If we knew the initial state of our universe, we would know its entire history. Similarly, in the quantum theory of gravity, there are many different possible quantum states for the universe. Again, if we knew how the Euclidean curved space-times in the sum over histories behaved at early times, we would know the quantum state of the universe.

In the classical theory of gravity, which is based on real space-time, there are only two possible ways the universe can behave: either it has existed for an infinite time, or else it had a beginning at a singularity at some finite time in the past. In the quantum theory of gravity, on the other hand, a third possibility arises. Because one is using Euclidean space-times, in which the time direction is on the same footing as directions in space, it is possible for space-time to be finite in extent and yet to have no singularities that formed a boundary or edge. Space-time would be like the surface of the earth, only with two more dimensions. The surface of the earth is finite in extent but it doesn't have a boundary or edge: if you sail off into the sunset, you don't fall off the edge or run into a singularity. (I know, because I have been round the world!)

If Euclidean space-time stretches back to infinite imaginary time, or else starts at a singularity in imaginary time, we have the same problem as in the classical theory of specifying the initial state of the universe: God may know how the universe began, but we cannot give any particular reason for thinking it began one way rather than another. On the other hand, the quantum theory of gravity has opened up a new possibility, in which there would be no boundary to space-time and so there would be no need to specify the behavior at the boundary. There would be no singularities at which the laws of science broke down, and no edge of space-time at which one would have to appeal to God or some new law to set the boundary conditions for space-time. One could say: "The boundary condition of the universe is that it has no boundary." The universe would be completely self-contained and not affected by anything outside itself. It would neither be created nor destroyed, It would just BE.

It was at the conference in the Vatican mentioned earlier that I first put forward the suggestion that maybe time and space together formed a surface that was finite in size but did not have any boundary or edge. My paper was rather mathematical, however, so its implications for the role of God in the creation of the universe were not generally recognized at the time (just as well for me). At the time of the Vatican conference, I did not know how to use the "no boundary" idea to make predictions about the universe. However, I spent the following sum-mer at the University of California, Santa Barbara. There a friend and colleague of mine, Jim Hartle, worked out with me what conditions the universe must satisfy if space-time had no boundary. When I returned to Cambridge, I continued this work with two of my research students, Julian Luttrel and Jonathan Halliwell.

I'd like to emphasize that this idea that time and space should be finite "without boundary" is just a *proposal:* it cannot be deduced from some other principle. Like any other scientific theory, it may initially be put forward for aesthetic or metaphysical reasons, but the real test is whether it makes predictions that agree with observation. This, how-ever, is difficult to determine in the case of quantum gravity, for two reasons. First, as will be explained in Chapter 11, we are not yet sure exactly which theory successfully combines general relativity and quantum mechanics, though we know quite a lot about the form such a theory must have. Second, any model that described the whole universe in detail would be much too complicated mathematically for us to be able to calculate exact predictions. One therefore has to make simplifying assumptions and approximations – and even then, the problem of extracting predictions remains a formidable one.

Each history in the sum over histories will describe not only the space-time but everything in it as well, including any complicated organisms like human beings who can observe the history of the universe. This may provide another justification for the anthropic principle, for if all the histories are possible, then so long as we exist in one of the histories, we may use the anthropic principle to explain why the universe is found to be the way it is. Exactly what meaning can be attached to the other histories, in which we do not exist, is not clear. This view of a quantum theory of gravity would be much more satisfactory, however, if one could show that, using the sum over histories, our universe is not just one of the possible histories but one of the most probable ones. To do this, we must perform the sum over histories for all possible Euclidean space-times that have no boundary.

Under the "no boundary" proposal one learns that the chance of the universe being found to be following most of the possible histories is negligible, but there is a particular family of histories that are much more probable than the others. These histories may be pictured as being like the surface of the earth, with the distance from the North Pole representing imaginary time and the size of a circle of constant distance from the North Pole representing the spatial size of the universe. The universe starts at the North Pole as a single point. As one moves south, the circles of latitude at constant distance from the North Pole get bigger, corresponding to the universe expanding with imaginary time Figure 8:1. The universe would reach a maximum size at the equator and would contract with increasing imaginary time to a single point at the South Pole. Ever though the universe would have zero size at the North and South Poles, these points would not be singularities, any more than the North aid South Poles on the earth are singular. The laws of science will hold at them, just as they do at the North and South Poles on the earth.

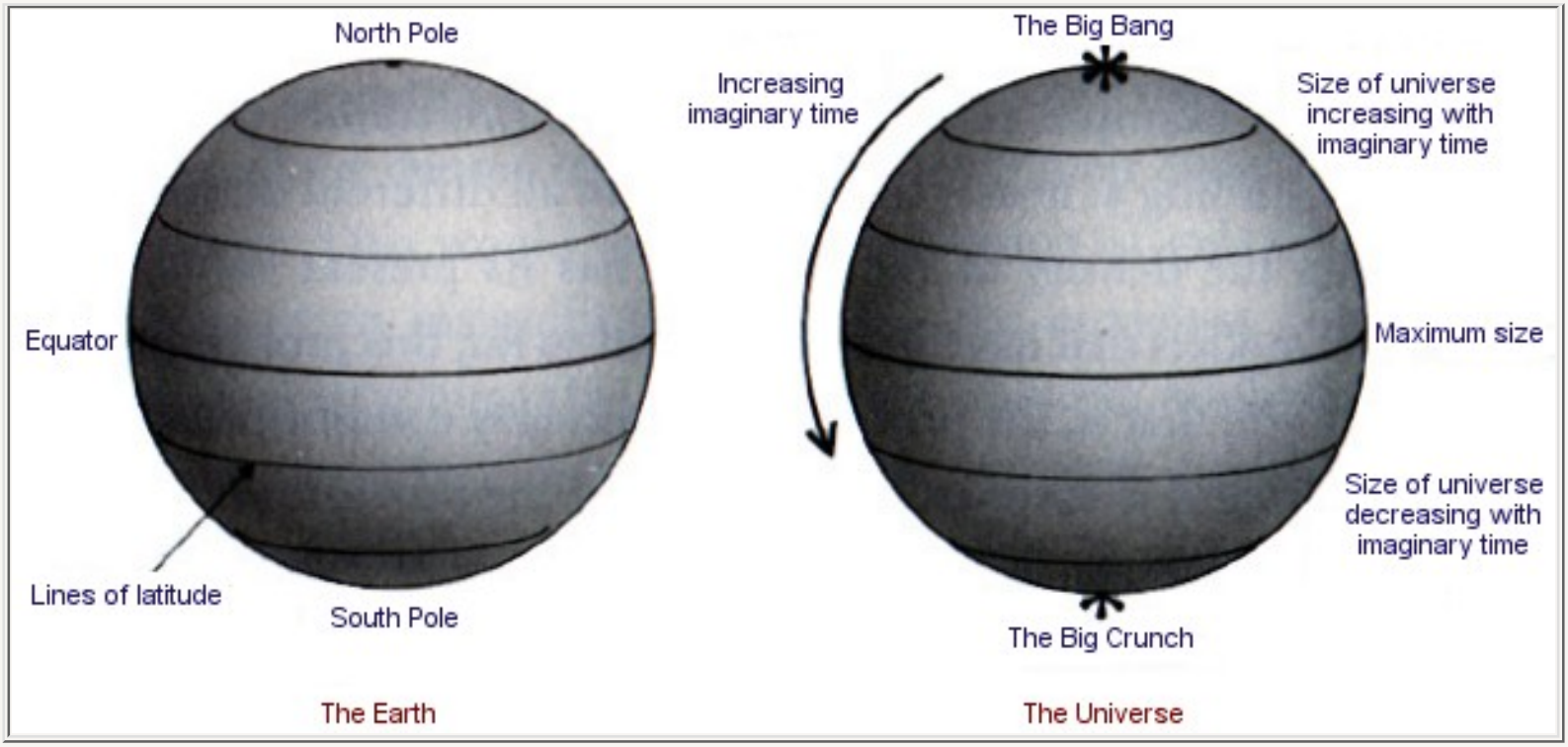

Figure 8:1

The history of the universe in real time, however, would look very different. At about ten or twenty thousand million years ago, it would have a minimum size, which was equal to the maximum radius of the history in imaginary time. At later real times, the universe would expand like the chaotic inflationary model proposed by Linde (but one would not now have to assume that the universe was created somehow in the right sort of state). The universe would expand to a very large size Figure 8:1 and eventually it would collapse again into what looks like a singularity in real time. Thus, in a sense, we are still all doomed, even if we keep away from black holes. Only if we could picture the universe in terms of imaginary time would there be no singularities.

If the universe really is in such a quantum state, there would be no singularities in the history of the universe in imaginary time. It might seem therefore that my more recent work had completely undone the results of my earlier work on singularities. But, as indicated above, the real importance of the singularity theorems was that they showed that the gravitational field must become so strong that quantum gravitational effects could not be ignored. This in turn led to the idea that the universe could be finite in imaginary time but without boundaries or singularities. When one goes back to the real time in which we live, however, there will still appear to be singularities. The poor astronaut who falls into a black hole will still come to a sticky end; only if he lived in imaginary time would he encounter no singularities.

This might suggest that the so-called imaginary time is really the real time, and that what we call real time is just a figment of our imaginations. In real time, the universe has a beginning and an end at singularities that form a boundary to space-time and at which the laws of science break down. But in imaginary time, there are no singularities or boundaries. So maybe what we call imaginary time is really more basic, and what we call real is just an idea that we invent to help us describe what we think the universe is like. But according to the approach I described in Chapter 1, a scientific theory is just a mathematical model we make to describe our observations: it exists only in our minds. So it is meaningless to ask: which is real, “real” or “imaginary” time? It is simply a matter of which is the more useful description.

One can also use the sum over histories, along with the no boundary proposal, to find which properties of the universe are likely to occur together. For example, one can calculate the probability that the universe is expanding at nearly the same rate in all different directions at a time when the density of the universe has its present value. In the simplified models that have been examined so far, this probability turns out to be high; that is, the proposed no boundary condition leads to the prediction that it is extremely probable that the present rate of expansion of the universe is almost the same in each direction. This is consistent with the observations of the microwave background radiation, which show that it has almost exactly the same intensity in any direction. If the universe were expanding faster in some directions than in others, the intensity of the radiation in those directions would be reduced by an

additional red shift.

Further predictions of the no boundary condition are currently being worked out. A particularly interesting problem is the size of the small departures from uniform density in the early universe that caused the formation first of the galaxies, then of stars, and finally of us. The uncertainty principle implies that the early universe cannot have been completely uniform because there must have been some uncertainties or fluctuations in the positions and velocities of the particles. Using the no boundary condition, we find that the universe must in fact have started off with just the minimum possible non-uniformity allowed by the uncertainty principle. The universe would have then undergone a period of rapid expansion, as in the inflationary models. During this period, the initial non-uniformities would have been amplified until they were big enough to explain the origin of the structures we observe around us. In 1992 the Cosmic Background Explorer satellite (COBE) first detected very slight variations in the intensity of the microwave background with direction. The way these non-uniformities depend on direction seems to agree with the predictions of the inflationary model and the no boundary proposal. Thus the no boundary proposal is a good scientific theory in the sense of Karl Popper: it could have been falsified by observations but instead its predictions have been confirmed. In an expanding universe in which the density of matter varied slightly from place to place, gravity would have caused the denser regions to slow down their expansion and start contracting. This would lead to the formation of galaxies, stars, and eventually even insignificant creatures like ourselves. Thus all the complicated structures that we see in the universe might be explained by the no boundary condition for the universe together with the uncertainty principle of quantum mechanics.

The idea that space and time may form a closed surface without boundary also has profound implications for the role of God in the affairs of the universe. With the success of scientific theories in describing events, most people have come to believe that God allows the universe to evolve according to a set of laws and does not intervene in the universe to break these laws. However, the laws do not tell us what the universe should have looked like when it started – it would still be up to God to wind up the clockwork and choose how to start it off. So long as the universe had a beginning, we could suppose it had a creator. But if the universe is really completely self-contained, having no boundary or edge, it would have neither beginning nor end: it would simply be. What place, then, for a creator?

PREVIOUS NEXT