Space-time and Cosmology

ROGER PENROSE

THE TANNER LECTURES ON HUMAN VALUES

Delivered at

Cambridge University
February 13–15, 1995

SIR ROGER PENROSE is Rouse Ball Professor of Mathematics at Oxford University, as well as Francis and Helen Pentz Distinguished Professor of Physics and Mathematics at Pennsylvania State University. He was educated at University College, London, and received his Ph.D. from Cambridge University. He is a fellow of the Royal Society and recipient of numerous prizes and awards, including the 1988 Wolf Prize that he shared with Steven Hawking for their understanding of the universe. His numerous publications include *Shadows of the Mind* (1994), *The Emperor's New Mind* (1994), *Spinors and Space Time* (1984 and 1986, coauthor), and *Techniques of Differential Topology* (1972). He was knighted in 1994 for services to science.

The title of these lectures is "The Large, the Small, and the Human Mind" and the subject of this first lecture is the Large.[1] The first and second lectures are concerned with our physical Universe, which I represent very schematically as the "sphere" in my first picture.

However, these will not be "botanical" lectures, telling you in detail what is here and what is there in our Universe, but rather I want to concentrate upon understanding of the actual laws that govern the way the world behaves. One of the reasons that I have chosen to divide my descriptions of the physical laws between two lectures, namely, the Large and the Small, is that the laws that govern the large-scale behaviour of the world and those that govern its small-scale behaviour seem to be very different. The fact that they seem to be so different, and what we might have to do about his seeming discrepancy, is central to the subject of the third lecture —which is where the human mind comes in.

Since I shall be talking about the physical world in terms of the physical theories that underlie its behaviour, I shall also have to say something about another world, the Platonic world of absolutes, in its particular role as the world of mathematical truth.

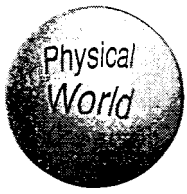


FIGURE 1.

[1] This is the first of three lectures with the overall title "The Large, the Small, and the Human Mind." All three lectures will be published by Cambridge University Press.

One can well take the view that the "Platonic world" contains other absolutes, such as the Good and the Beautiful, but I shall be concerned here only with the Platonic concepts of mathematics. Some people find it hard to conceive of this world as existing on its own. They may prefer to think of mathematical concepts merely as idealisations of our physical world —and, on this view, the mathematical world would be thought of as emerging from the world of physical objects (figure 2).

Now, this is not how I think of mathematics, nor, I believe, is it how most mathematicians or mathematical physicists think about the world. They think about it in a rather different way, as a structure precisely governed according to timeless mathematical laws. Thus, they prefer to think of the physical world, more appropriately, as emerging out of the ("timeless") world of mathematics, as illustrated in figure 3. This picture will have importance for what I shall say in the third lecture, and it also underlies most of what I shall say in the first two.

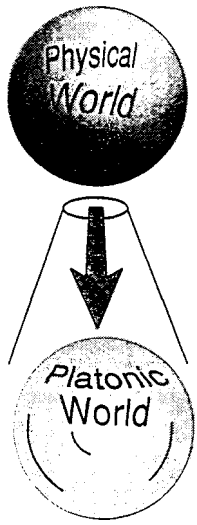


FIGURE 2.

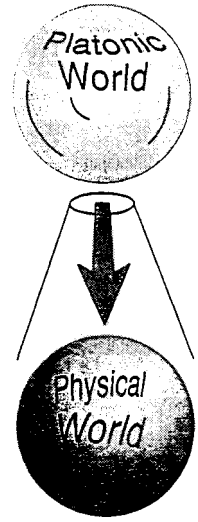


FIGURE 3.

One of the remarkable things about the behaviour of the world is how it seems to be grounded in mathematics to a quite extraordinary degree of accuracy. The more we understand about the physical world, and the deeper we probe into the laws of nature, the more it seems as though the physical world almost evaporates and we are left only with mathematics. The more deeply we understand the laws of physics, the more we are driven into this world of mathematics and of mathematical concepts.

Let us look at the scales we have to deal with in the Universe and also the role of our place in the Universe. I can summarise all these scales in a single diagram (figure 4). On the left-hand side

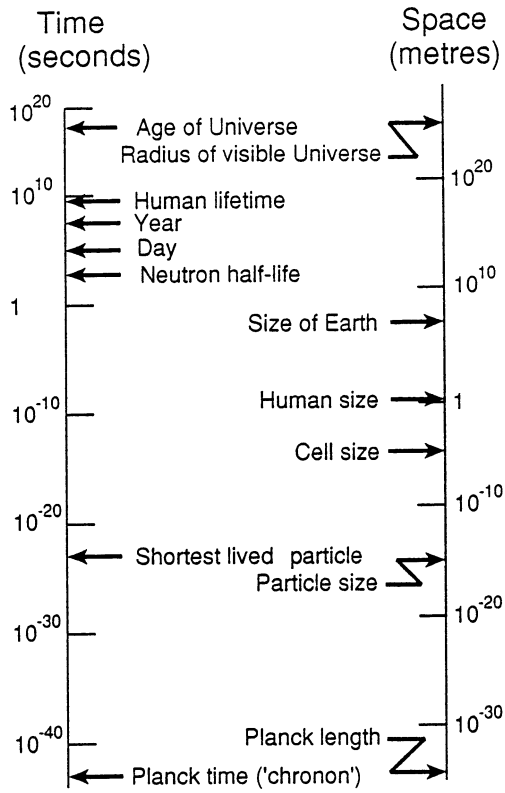


FIGURE 4. Sizes and time-scales in the Universe.

of the diagram, time-scales are shown. At the bottom of the diagram, on the left-hand side, is the very shortest time scale that is physically meaningful. This time-scale is about 10^{-43} of a second and is often referred to as the *Planck time-scale* or a "chronon." This time-scale is much shorter than anything experienced in particle physics. For example, the shortest-lived particles, called resonances, last for about 10^{-23} of a second. Further up the diagram, on the left, the day and the year are shown, and, at the top of the diagram, the present age of the Universe is shown.

On the right-hand side of the diagram, distances corresponding to these time-scales are depicted. The length corresponding to the Planck time (or chronon) is the fundamental unit of length, called the *Planck length*. These concepts of the Planck time and the Planck length fall out naturally when one tries to combine the physical theories that describe the large and the small, that is, combining Einstein's General Relativity, which describes the physics of the very large, with quantum mechanics, which describes the physics of the very small. When these theories are brought together, these Planck lengths and times turn out to be fundamental. The translation from the left-hand to the right-hand axis of the diagram is via the speed of light so that times can be translated into distances by asking how far a light signal could travel in that time.

The sizes of the physical objects represented on the diagram range from about 10^{-15} of a metre for the characteristic sizes of particles to about 10^{27} meters for the radius of the observable Universe at the present time, which is roughly the age of the Universe multiplied by the speed of light. It is intriguing to note where *we* are in the diagram, namely, the human scale. With regard to spatial dimensions, it can be seen that we are more or less in the middle of the diagram. We are enormous compared with the Planck length; even compared with the size of particles, we are very large. Yet, compared with the distance scale of the observable Universe, we are very tiny. Indeed, we are as small compared with

it as we are large compared with particles. In contrast, with regard to temporal dimensions, the human lifetime is almost as long as the Universe! People talk about the ephemeral nature of existence but, when you look at the human lifetime as shown in the diagram, it can be seen that we are not ephemeral at all — we live more or less as long as the Universe itself! Of course, this is looking on a "logarithmic scale," but this is the natural thing to do when we are concerned with such enormous ranges. To put it another way, the number of human lifetimes that make up the age of the Universe is very, very much less than the number of Planck times, or even lifetimes of the shortest-lived particles, that make up a human lifetime. Thus, we are really very stable structures in the Universe. As far as spatial sizes are concerned, we are very much in the middle — we directly experience neither the physics of the very large nor the physics of the very small. We are very much in-between. In fact, looked at logarithmically, all living objects from single cells to human beings are roughly the same in-between size.

What kinds of physics apply on these different scales? Let me introduce the diagram that summarizes the whole of physics. I have had to leave out a few details, of course, such as all the equations! But the essential basic theories that physicists use are indicated.

The key point is that, in physics, we use two very different types of procedure. To describe the small-scale behaviour, we use quantum mechanics — what I have described as the quantum level in figure 5. I shall say much more about this in the second lecture. One of the things that people say about quantum mechanics is that it is fuzzy and indeterministic, but this is not true. So long as you remain at this level, quantum theory is deterministic and precise. In its most familiar form, quantum mechanics involves use of the equation known as Schrödinger's equation, which governs the behaviour of the physical state of a quantum system — called its *quantum state* — and this is a deterministic equation. I have used the letter U to describe this quantum level activity. Indeterminacy

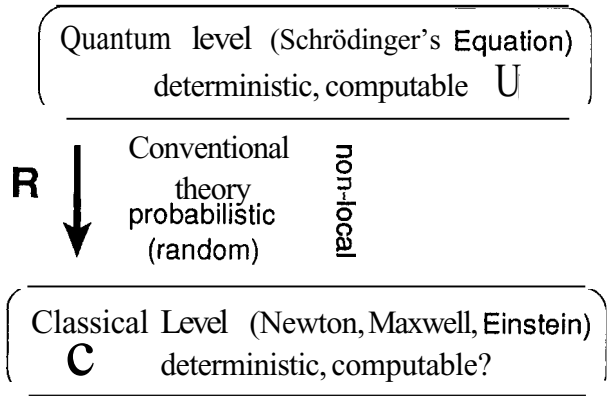


FIGURE 5.

in quantum mechanics arises only when you perform what is called "making a measurement" and that involves magnifying an event from the quantum level to the classical level. I shall say quite a lot about this in the second lecture.

On the large scale, we use classical physics, which is entirely deterministic —these classical laws include Newton's laws of motion, Maxwell's laws for the electromagnetic field, which incorporate electricity, magnetism, and light, and Einstein's theories of relativity, the Special Theory, which deals with large velocities, and the General Theory, which deals with large gravitational fields. These laws apply very, very accurately on the large scale.

Just as a footnote to figure 5, it can be seen that I have included a remark about "computability" in quantum and classical physics. This has no relevance to the present lecture or the next, but it will have importance for the third, and I shall return to the issue of computability in that lecture.

For the rest of the present lecture, I shall be primarily concerned with Einstein's theory of relativity —specifically, how theory works, its extraordinary accuracy, and something about its elegance as a physical theory. But let us first consider Newtonian theory. Newtonian physics, just as in the case of relativity, allows a space-

time description to be used. This was first precisely formulated by E.-J. Cartan for Newtonian gravity, some time after Einstein had presented his General Theory of relativity. The physics of Galileo and Newton is represented in space-time for which there is a global time coordinate, here depicted as running up the diagram (figure 6); and for each constant value of the time, there is a space section which is a Euclidean 3-space, here depicted as horizontal planes. An essential feature of the Newtonian space-time picture is that these space-slices, across the diagram, represent moments of simultaneity.

Thus, everything that occurs on Monday at noon lies on one horizontal slice through the space-time diagram; everything that happens on Tuesday at noon lies on the next slice shown in the

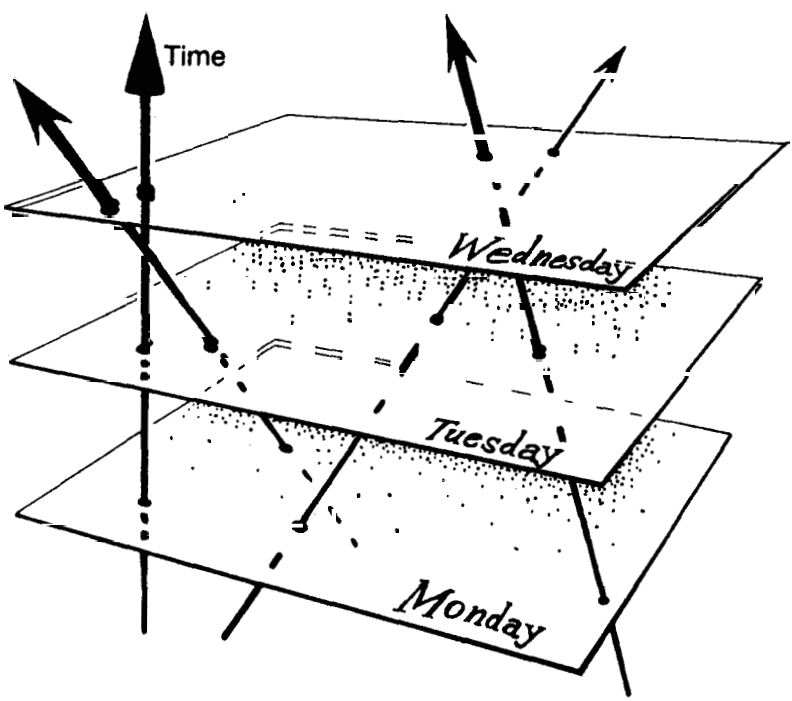


FIGURE 6. Galilean space-time: particles in uniform motion are depicted as straight lines.

diagram; and so on. Time cuts across the space-time diagram and the Euclidean sections follow one after the other as time progresses. All the observers can agree about the time when events take place, no matter how they move through the space-time, because everyone uses the same time slices to measure how time passes.

In Einstein's Special Theory of relativity, one has to adopt a different picture. In it, the space-time picture is absolutely essential —the key difference is that time is not the universal thing it is in Newtonian theory. To appreciate how the theories differ, it is necessary to understand an essential part of relativity theory, namely, those structures known as *light-cones.*

What is a light-cone? A light-cone is drawn in figure 7. We imagine a flash of light taking place at some point at some instant — that is, at an *event* in space-time — and the light waves travel outwards from this event, the source of the flash, at the speed of light. In a purely spatial picture (right-hand picture of figure 7), we can represent the paths of the light waves through space as a sphere expanding at the speed of light. We can now translate this motion of the light waves into a space-time diagram (left-hand picture of figure 7) in which time runs up the diagram and the space coordinates refer to horizontal displacements, just as

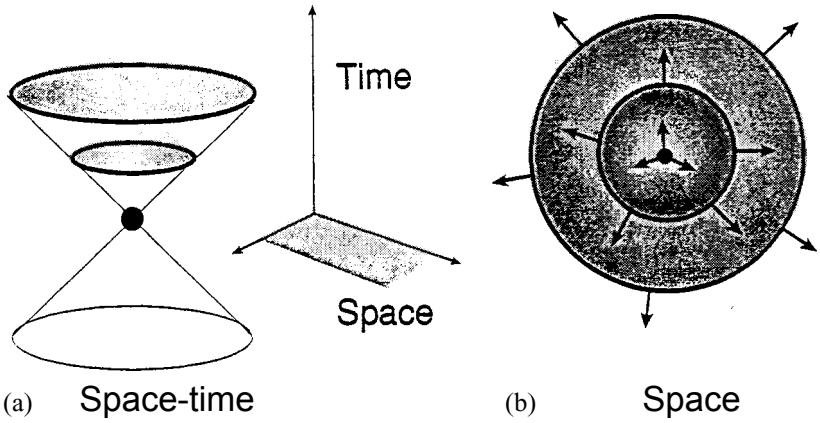


FIGURE 7. The representation of the history of a light flash in terms of its propagation in (a) space-time and (b) space.

in the Newtonian situation of figure 6. Unfortunately, in the full space-time picture, on the left of figure 7, we can represent only two spatial dimensions horizontally on the diagram, because the space of our picture is only three-dimensional, Now, we see that the flash is represented by a point (event) at the origin and that the subsequent paths of the light rays (waves) cut the horizontal "space" planes in circles, the radii of which increase at the speed of light up the diagram. It can be seen that the paths of the light rays form cones in the space-time diagram. The light-cone thus represents the history of this flash of light —light propagates away from the origin along the light-cone, which means at the speed of light, into the future. Light rays can also arrive at the origin along the light-cone from the past —that part of the light-cone is known as the past light-cone and all information carried to the observer by light waves arrives at the origin along this cone.

Light-cones represent the most important structures in space-time. In particular, they represent the limits of causal influence. The history of a particle in space-time is represented by a line travelling up the space-time diagram, and this line has to lie within the light-cone (figure 8). This is just another way of saying that a

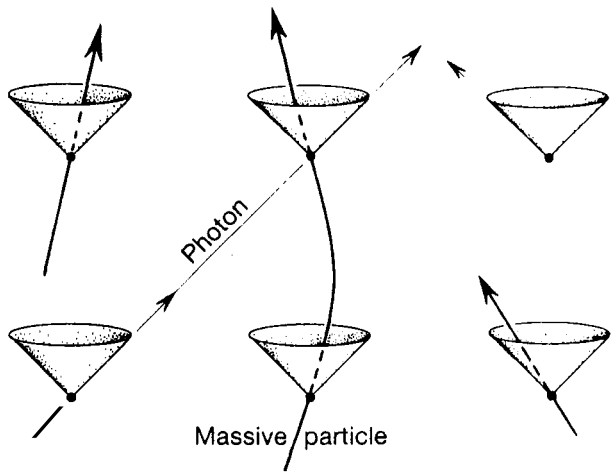


FIGURE 8. A picture of Minkowski geometry.

material particle cannot travel faster than the speed of light. No signal can travel from inside to outside the future light-cone, so that the light-cone does indeed represent the limits of causality.

There are some remarkable geometrical properties that relate to the light-cones. Let us consider two observers moving at different speeds through space-time. Unlike the case of Newtonian theory, in which the planes of simultaneity are the same for all observers, there is no absolute simultaneity in relativity. Observers moving at different speeds draw their own planes of simultaneity as different sections through space-time, as illustrated in figure 9. There is a very well-defined way of transforming from one plane to another through what is known as a *Lorentz transformation,*

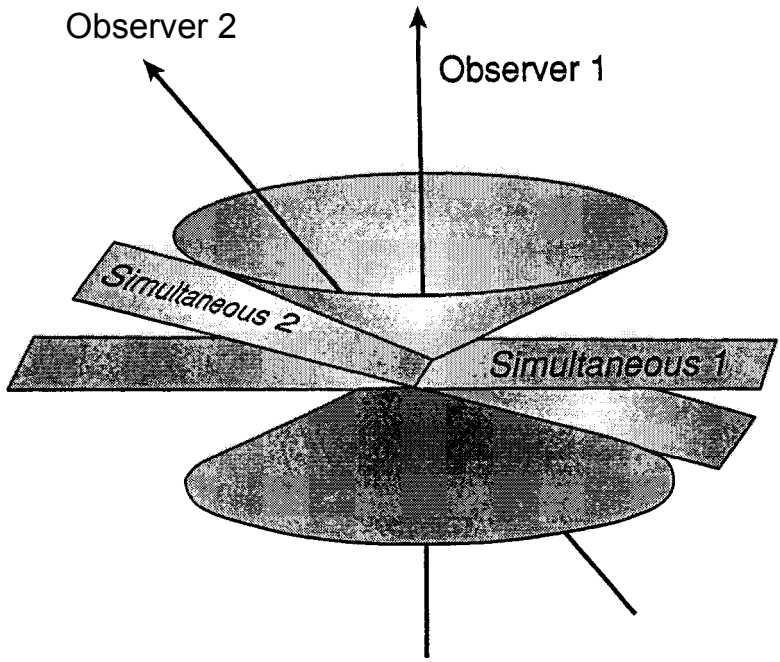


FIGURE 9. Illustrating the relativity of simultaneity according to Einstein's Special Theory of relativity, Observers 1 and 2 are moving relative to one another through space-time. Events that are simultaneous for Observer 1 are not simultaneous for Observer 2 and vice versa.

these transformations constituting what is called the *Lorentz group.* The discovery of this group was an essential feature in the discovery of Einstein's Special Theory of relativity. The Lorentz group can be understood as a group of (linear) space-time transformations, leaving a light-cone invariant.

We can also appreciate the Lorentz group from a slightly different viewpoint. As I have emphasised, the light-cones are the fundamental structures of space-time. Imagine that you are an observer located somewhere in space, looking out at the Universe. What you see are the light rays coming from the stars to your eyes. According to the space-time viewpoint, the events you observe are the intersections of the world-lines of the stars with your past light-cone, as illustrated in figure 10(a). You observe along your past light-cone the positions of the stars at particular points. These points seem to be situated on the celestial sphere that appears to surround you. Now, imagine another observer, moving at some great speed relative to you, who passes closely by you at the moment you both look out at the sky. This second observer perceives the same stars as you do, but finds them to be located in different positions on the celestial sphere —this is the effect known as *aberration.* There is a set of transformations that enables us to work out the relationship between what each of these observers sees on his or her celestial sphere. Each of these transformations is one that takes a sphere to a sphere. But it is one of a very special kind. It takes exact circles to exact circles and it preserves angles. Thus, if a pattern in the sky appears to be circular to you, then it must also appear circular to the other observer.

There is a very beautiful way of describing how this works and I illustrate it to show that there is a particular elegance in the mathematics that often underlies physics at its most fundamental level. Figure 10(c) shows a sphere with a plane drawn through its equator. We can draw figures on the surface of the sphere and then examine how they are projected to the equatorial plane from the south pole, as illustrated. This type of projection is known as

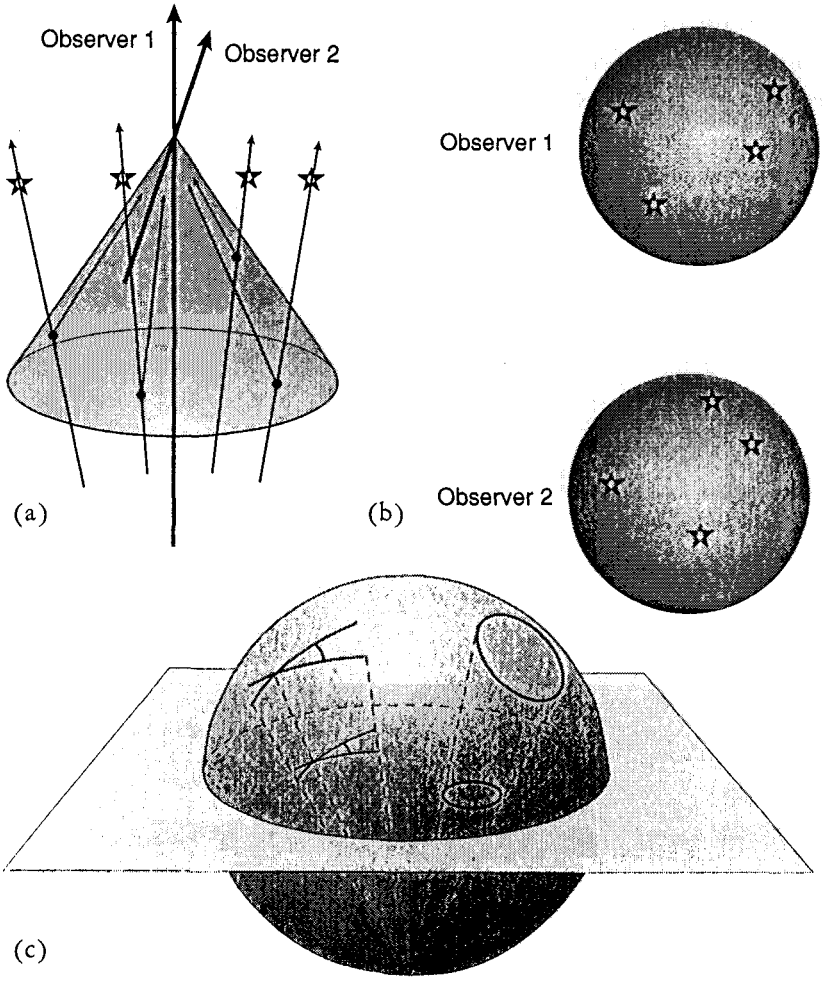


FIGURE 10. Illustrating how observations are made of the sky by Observers 1 and 2. (a) Observer 1 observes stars along the past light-cone. The points at which stars cross the light-cone are indicated by black dots. Light signals propagate from the stars to the observers along the light-cone as illustrated. Observer 2 is moving through space-time at a certain speed relative to Observer 1. (b) Illustrating the location of stars on the sky as observed by Observer 1 and Observer 2, when they are coincident at some point in space-time, (c) Illustrating the use of the Riemann sphere to locate the positions of stars on the celestial sphere for observers moving at a constant speed relative to each other.

a stereographic projection and it has some rather extraordinary properties. Circles on the sphere are projected into exact circles on the plane, and the angles between curves on the sphere are projected into exactly the same angles on the plane. As I shall discuss more fully in the second lecture, this projection allows us to label the points of the sphere by complex numbers (numbers involving the square root of -1) that can be used to label the points of the equatorial plane, together with "infinity," to give it the structure known as the "Riemann sphere."

For those who are interested, the transformation is

$$\mu \to \mu' = \frac{\alpha\mu + \beta}{\lambda\mu + \delta}$$

It is a fact well known to mathematicians that this transformation sends circles into circles and preserves angles. Transformations of this kind are known as Möbius transformations. For our present purposes, we need merely note the simple elegance of the form of the Lorentz (aberration) formula when written in terms of such a complex parameter μ

A striking point about this way of looking at these transformations is that, according to Special Relativity, the formula is very simple, whereas, in expressing the corresponding aberration transformation according to Newtonian Mechanics, the formula would be much more complicated. It often turns out that, when you get down to the fundamentals and develop a more exact theory, the mathematics turns out to be simpler, even if the formalism appears to be more complicated in the first instance. This important point is exemplified by the contrast between Galilean and Einstein's relativity.

Thus, in the Special Theory of relativity, we have a theory that is, in many ways, simpler than Newtonian mechanics. From the point of view of mathematics, and particularly from the point of

view of group theory, it is a much nicer structure. In Special Relativity, the space-time is flat and all the light cones are lined up regularly, as illustrated in figure 8. If we now go one step further to Einstein's General Relativity, that is, the theory of space-time in the presence of gravity, the picture seems at first sight rather muddied up —the light-cones are all over the place (figure 11). Now, I have been saying that, as we develop deeper and deeper theories, the mathematics becomes simpler, but look what has happened here —I had a nice elegant piece of mathematics that has become horribly complicated. Well, that sort of thing happens —you will have to bear with me for a little while until the simplicity reappears.

Let me remind you of the fundamental ingredients of Einstein's theory of gravity. One basic ingredient is called Galileo's Principle of Equivalence. In figure 12, I show Galileo leaning over from the top of the Tower of Pisa dropping large and small rocks. Whether or not he actually performed this experiment, he certainly well understood that, if the effects of air resistance are ignored, the two rocks would fall to the ground in the same time. If you happened to be sitting on one of these rocks looking at the other one as they fall together, you would observe the other rock hovering in front of you (I have shown a camcorder attached to one of the rocks to make the observation). Nowadays, with space travel, this is a very familiar phenomenon —just recently, we have seen a British-born astronaut walking in space, and, just like the big rock and the little rock, the spaceship hovers in front of the astronaut —this is exactly the same phenomenon as Galileo's Principle of Equivalence.

Thus, if you look at gravity in the right way, that is, in a falling frame of reference, it seems to disappear right in front of your eyes. This is indeed correct. But Einstein's theory does *not* tell you that gravity disappears — it only tells you that the *force* of gravity disappears. There is something left and that is the tidal effect of gravity.

Let me introduce a little bit more mathematics, but not much. We need to describe the curvature of space-time and this is de-

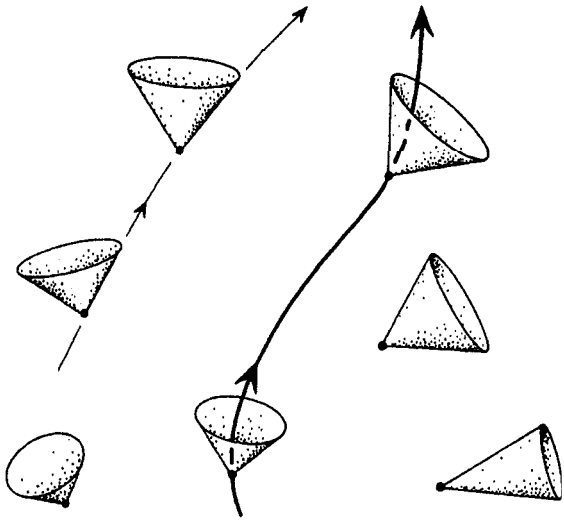


FIGURE 11. A picture of curved space-time.

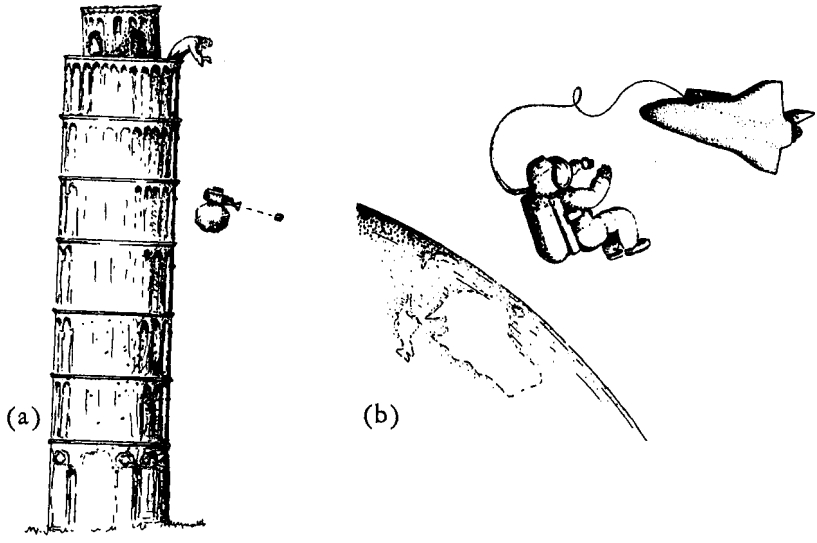


FIGURE 12. (a) Galileo dropping two rocks (and a camcorder) from the Leaning Tower of Pisa. (b) The astronaut sees the space-vehicle hover before him or her, seemingly unaffected by gravity.

scribed by an object known as a tensor which I have called *Riemann* in the following equation. It is actually called the Riemann curvature tensor but I will not tell you what it is except that it is represented by a capital *R* with a number of indices stuck on the bottom, which are indicated by the dots. The Riemann curvature tensor is made up of two pieces. One of the pieces is called the *Weyl* curvature and the other piece is called the *Ricci* curvature, and we have the (schematic) equation

$$\begin{aligned} Riemann &= Weyl + Ricci \\ R\ldots. &= C\ \ldots.\ + R..g.. \end{aligned}$$

Formally, C. . . . and *R*. . are the Weyl and Ricci curvature tensors, respectively, and g. . is the metric tensor.

The Weyl curvature measures the tidal effect. What is the "tidal" effect? Recall that, from the astronaut's point of view, it seems that gravity has been abolished, but that is not quite true. Imagine that the astronaut is surrounded by a sphere of particles, which are initially at rest with respect to the astronaut. Now, at first they will just hover there but soon they will start to accelerate because of the slight differences in the gravitational attraction of the Earth at different points in the sphere. (Notice that I am describing the effect in Newtonian language, but that is quite adequate.) These slight differences cause the original sphere of particles to become distorted into an elliptical arrangement, as illustrated in figure 13(a) .

This distortion occurs partly because of the slightly greater attraction of the Earth for those particles closer to the Earth and lesser attraction for those further away, and partly because, at the sides of the sphere, the Earth's attraction acts slightly inwards. This causes the sphere to be distorted into an ellipsoid. It is called the tidal effect for the very good reason that, if you replace the Earth by the Moon and the sphere of particles by the Earth with its oceans, then the effect of the Moon on the surface of the oceans is

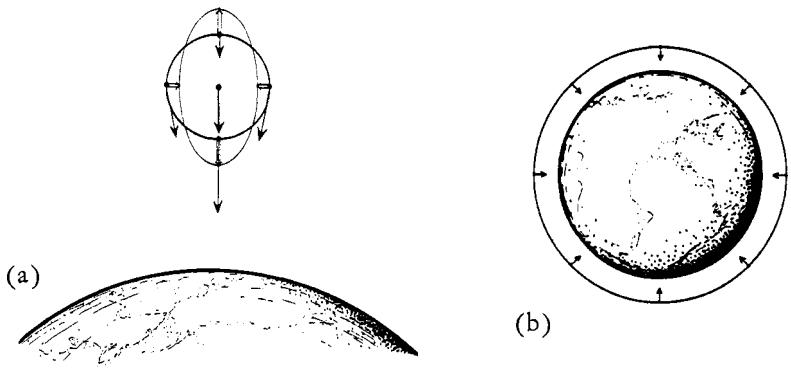


FIGURE 13. (a) The tidal effect. Double arrows show relative acceleration. (b) When the sphere surrounds matter (here on Earth), there is a net inward acceleration.

exactly the gravitational effect upon the sphere of particles —the sea surface closest to the Moon is pulled towards it, whereas the ocean surface on the other side of the Earth is, in effect, pushed away from it. The effect causes the sea surface to bulge out on either side of the Earth and is the cause of the two high tides that occur each day.

The effects of gravity, from Einstein's point of view, are simply this tidal effect. It is defined precisely by the Weyl curvature, that is, the part C. . . . of the Riemann curvature. This part of the curvature tensor is volume-preserving —that is, if you work out the initial accelerations of the particles of the sphere, the volume of the sphere and the volume of the ellipsoid into which it is distorted are initially the same.

The remaining part of the curvature is known as the *Ricci* curvature and it has a volume-reducing effect. From figure 13(b), it can be seen that if, instead of being at the bottom of the diagram, the Earth were inside the sphere of particles, the volume of the sphere of particles would be reduced as the particles accelerate inwards. The amount of this reduction in volume is a measure of the Ricci curvature. Einstein's theory tells us that the Ricci curva-

ture is determined by the amount of matter present within a small sphere about that point in space. In other words, the density of matter, appropriately defined, tells us how the particles are accelerated inwards at that point in space. Einstein's theory is almost the same as Newton's when expressed in this way.

This is how Einstein formulates his theory of gravity — it is expressed in terms of the tidal effects, which are measurements of the local space-time curvature. It is crucial that we have to think in terms of the curvature of four-dimensional space-time. This was shown schematically in figure 11 — we think of the lines that represent the world lines of particles and the ways in which these paths are distorted as a measurement of the curvature of space-time. Thus, Einstein's theory is essentially a geometric theory of four-dimensional space-time — it is an extraordinarily beautiful theory mathematically.

The history of Einstein's discovery of the theory of General Relativity contains an important moral. It was first fully formulated in 1915. It was not motivated by any observational need but by various aesthetic, geometric, and physical desiderata. The key ingredients were Galileo's Principle of Equivalence, exemplified by his dropping rocks of different masses (figure 12), and the ideas of non-Euclidean geometry, which is the natural language for describing the curvature of space-time. There was not a great deal on the observational side in 1915. Once General Relativity was formulated in its final form, it was realised that there were the three key observational tests of the theory. The perihelion of the orbit of Mercury is swung around, or precesses, in a way that could not be explained by the Newtonian gravitational influence of the other planets — General Relativity predicts exactly the observed precession. The paths of light rays are bent by the Sun and this was the reason for the famous eclipse expedition of 1919, led by Arthur Eddington, which found a result consistent with Einstein's prediction (figure 14). The third test was the prediction that clocks run slow in a gravitational potential — that is, a clock closer to the

ground runs slow with respect to a clock at the top of a tower. This effect has also been measured experimentally. These were never, however, very impressive tests —the effects were always very small and various different theories can give the same results.

The situation has now changed dramatically —in 1993, Russell Hulse and Joseph Taylor were awarded the Nobel prize for a most

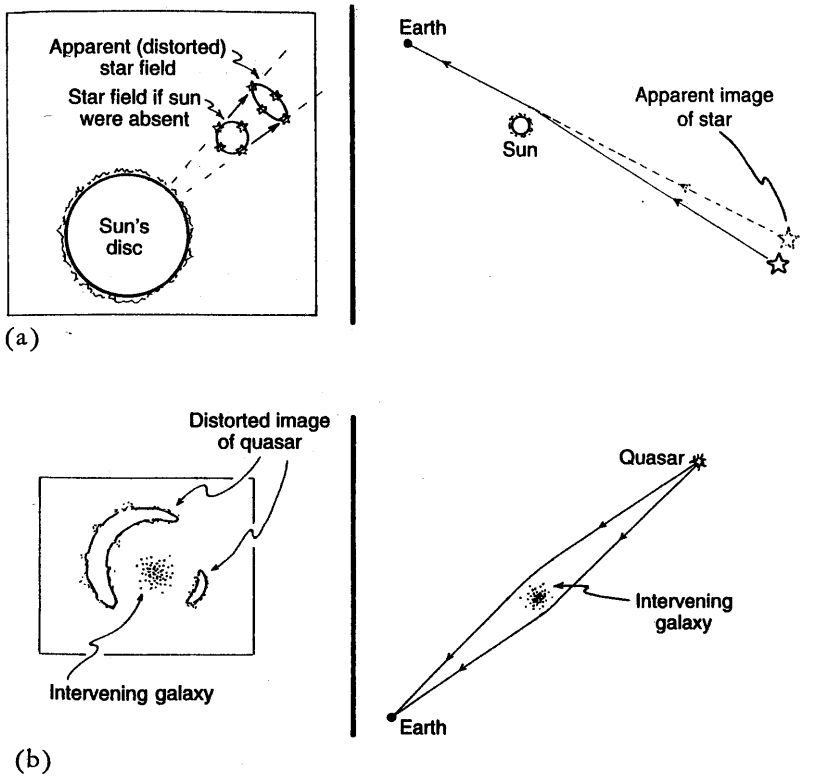
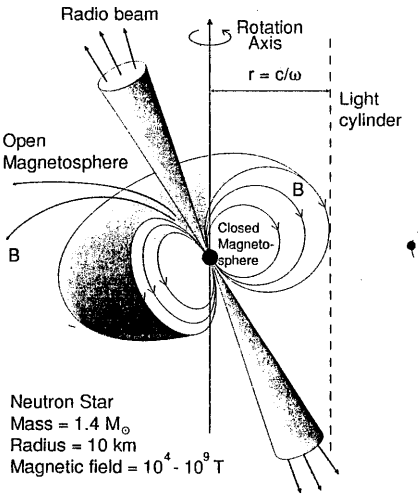


FIGURE 14. (a) A direct observational effect of light-cone tilting. The *Weyl* space-time curvature manifests itself as a distortion of the distant star field, here owing to the light-bending effect of the Sun's gravitational field. A circular pattern of stars would get distorted into an elliptical one. (b) Einstein's light-bending effect is now an important tool in observational astronomy. The mass of the intervening galaxy may be estimated by how much it distorts the image of a distant quasar.

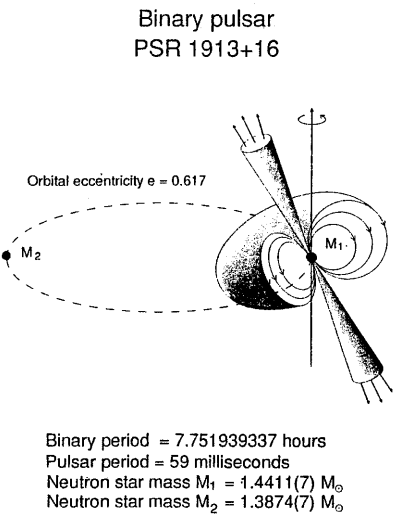
extraordinary series of observations. Figure 15 shows the binary pulsar known as PSR 1913+16 —it consists of a pair of neutron stars, each of which is an enormously dense star that has mass about that of the Sun but is only a few kilometres in diameter. The neutron stars orbit around their common centre of gravity in highly elliptical orbits. One of them has a very strong magnetic field and particles get swung round and emit intense radiation that travels to the Earth, some 30,000 light years away, where it is observed as a series of well-defined pulses. All sorts of very precise observations have been made of the arrival times of these pulses. In particular, all the properties of the orbits of the two neutron stars can be worked out as well as all the tiny corrections due to General Relativity.

There is, in addition, a feature that is completely unique to General relativity, and not present at all in the Newtonian theory of gravity. That is that objects in orbit about each other radiate away energy in the form of gravitational waves. These are like light waves but are ripples in space-time rather than ripples in the electromagnetic field. These waves take energy away from the system at a rate that can be precisely calculated according to Einstein's theory, and the rate of loss of energy of the binary neutron star system agrees very precisely with the observations, as illustrated by figure 15(c), which shows the speed-up of the orbital period of the neutron stars, measured over twenty years of observation. These signals can be timed so precisely that, over twenty years, the accuracy with which the theory is known to be correct amounts to about one part in 10^{14} This makes General Relativity the most accurately tested theory known to science.

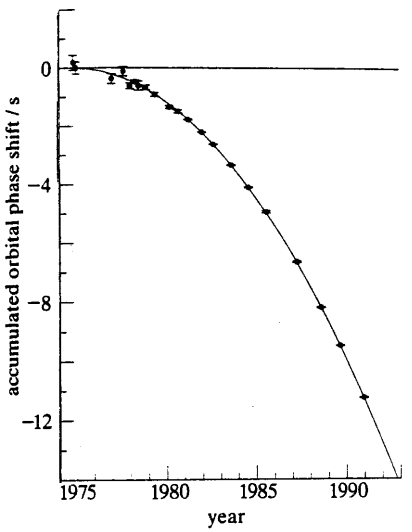
There is a moral in this story —Einstein's motivations for devoting eight or more years of his life to deriving the General Theory were not observational or experimental. Sometimes people argue, "Well, physicists look for patterns in their experimental results and then they find some nice theory that agrees with these. Maybe this explains why mathematics and physics work so well



(a)



(b)



(c)

FIGURE 15. (a) A schematic diagram illustrating the properties of those neutron stars that are radio pulsars. Radio emission is emitted along the poles of the magnetic dipole that is misaligned with respect to the rotation axis of the neutron star. Sharply defined pulses are observed when the narrow beam of radiation is swept across the line of sight to the observer. (b) A schematic representation of the binary pulsar PSR 1913 + 16. The properties of the two neutron stars have been derived from very precise timing of the arrival times of the pulses using effects that are only present in Einstein's General Relativity. (c) The change of phase of the arrival times of the pulses from the binary pulsar PSR 1913 + 16, compared with the expected change due to the emission of gravitational radiation by the binary neutron star system.

together." But, in this case, things were not like that at all. The theory was developed originally without any observational motivation —the mathematical theory is very elegant and it is physically very well motivated. The point is that the mathematical structure is just there in nature, the theory really is out there in space—it has not been imposed upon nature by anyone. That is one of the essential points of this lecture. Einstein revealed something that was there. Moreover, it was not just some minor piece of physics he discovered —it is the most fundamental thing that we have in nature, the nature of space and time.

Here is a very clear case —it goes back to my original diagram concerning the relation between the world of mathematics and the physical world (figure 3). In General Relativity, we have some kind of structure that really does underlie the behaviour of the physical world in an extraordinarily precise way. The way in which these fundamental features of our world are discovered is often not by looking at the way in which nature behaves, although that is obviously very important. One has to be prepared to throw out theories that might appeal for all sorts of other reasons but that do not fit the facts. But here we have a theory that does fit the facts with extraordinary accuracy. The accuracy involved is about twice as many figures as one has in Newtonian theory; in other words, General Relativity is known to be correct to one part in 10^{14} whereas Newtonian theory was tested only to one part in 10^{7}. The improvement is similar to the increase in accuracy with which Newton's theory was known to be correct between the seventeenth century and now. Newton knew his theory was correct to about one part in 1,000, whereas now it is known to be accurate to one part in 10^{7}.

Einstein's General Relativity is just a theory, of course. What about the structure of the actual world? I said this lecture would not be botanical but, if I talk about the Universe as a whole, that is not being botanical, since I will consider only the one Universe as a whole that is given to us. There are three types of standard model

that come out of Einstein's theory and these are defined by one parameter, which is, in effect, the one denoted by k in figure 16. There is another parameter that sometimes appears in cosmological arguments that is known as the cosmological constant. Einstein regarded his introduction of the cosmological constant into his equations of General Relativity as his greatest mistake and so I shall leave it out too. If we are forced to bring it back, well, we shall have to live with it.

Assuming the cosmological constant is zero, the three types of Universe, which are described by the constant k, are illustrated in figure 16. In the diagram, k takes values 1, 0, and -1, because all the other properties of the models have been scaled away. A better way would have been to talk about the age or scale of the Universe, and then one would have a continuous parameter but, qualitatively, the three different models can be thought of as being defined by the curvature of the space sections of the Universe. If the space sections of Universe are flat, they have zero curvature and $k = 0$. If the space sections are positively curved, meaning that the Universe closes in on itself, then $k = +1$. In this case the Universe has an initial singular state, the Big Bang, which marks the beginning of the Universe. It expands to a maximum size and then recollapses to a Big Crunch. Alternatively, there is the $k = -1$ case, in which the Universe expands forever. The $k = 0$ case is the limiting boundary between the $k = 1$ and $k = -1$ cases. Beside the diagrams, I have shown the radius-time relations for these three types of Universe in figure 16(d). The radius can be thought of as some typical scale in the Universe and it can be seen that only the case $k = +1$ collapses to a Big Crunch, while the other two expand indefinitely.

I want to consider the $k = -1$ case in a little more detail — it is perhaps the most difficult of the three to come to terms with. There are two reasons for being interested in this case particularly. One reason is that, if you take the observations as they exist at the moment at their face value, it is the preferred model. According

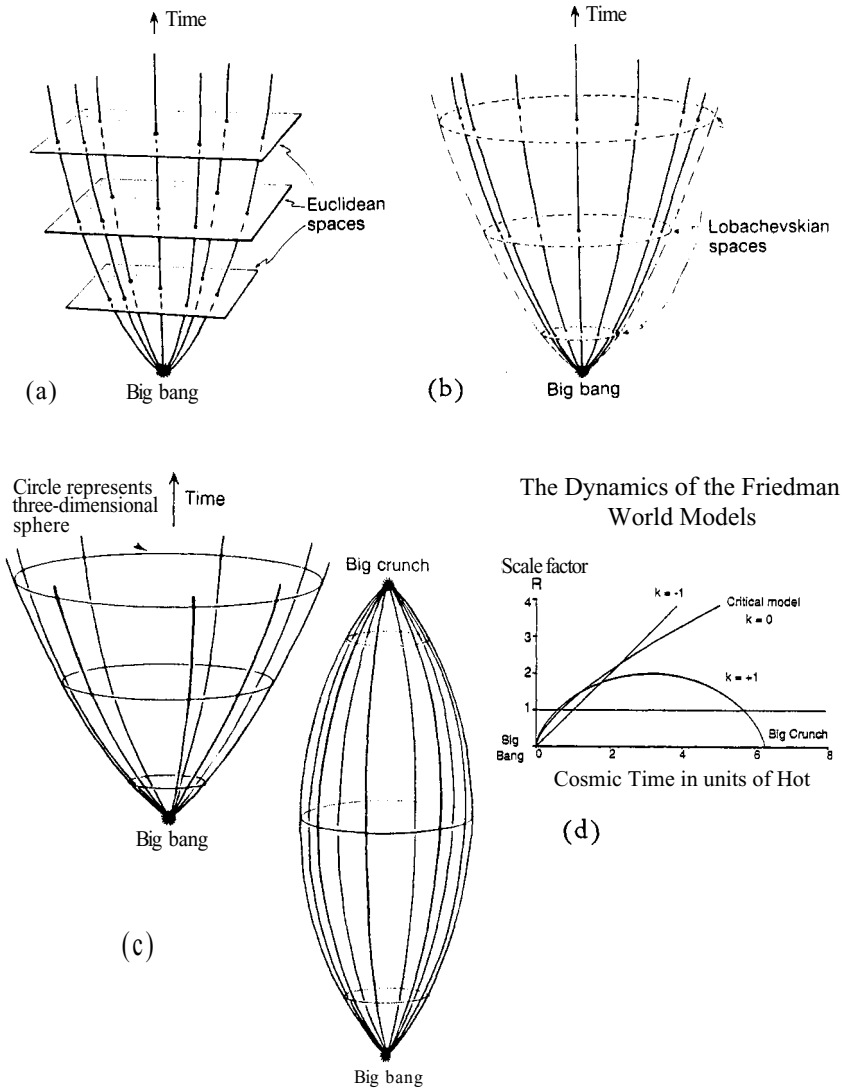


FIGURE 16. (a) Space-time picture of an expancaLig universe with Euclidean spatial sections (two space dimensions depicted), $k = 0$. (b) As in (a) but for an expanding universe with Lobachevskian spatial sections, $k = -1$. (c) As in (a), but for an expanding universe with spherical spatial sections $k = +1$. (d) The dynamics of the three different types of Friedman model.

to General Relativity, the curvature of space is determined by the amount of matter present in the Universe and there doesn't seem to be enough to close the geometry of the Universe. Now, it may be that there is a lot of dark or hidden matter, which we do not yet know about. In this case, the Universe could be one of the other models but, if there is not a lot of extra matter, much more than we believe must be present within the optical images of galaxies, then the Universe would have $k = -1$. The other reason is that it is the one I like the best! The properties of $k = -1$ geometries are particularly elegant.

What do the $k = -1$ universes look like? Their spatial sections have what is known as hyperbolic or Lobachevski geometry. To get a picture of a Lobachevski geometry it is best to look at one of Escher's prints. He made a number of prints that he called *Circle Limits,* and *Circle Limit 4* is shown in figure 17. This is Escher's description of the Universe-you see it is full of angels and devils!

A point to note is that it looks as though the picture gets very crowded towards the edge of the limit circle. This occurs because this representation of hyperbolic space is drawn on an ordinary plane sheet of paper, in other words, in Euclidean space. What you have to imagine is that all the devils are supposed to be actually exactly the same size and shape so that, if you happened to live in this Universe towards the edge of the diagram, they would look exactly the same to you as the ones in the middle of the diagram. This picture gives some impression of what is going on in Lobachevski geometry —as you walk from the centre out to the edge, you have to imagine that, because of the way the picture of the geometry has had to be distorted, the actual geometry there is exactly the same as it is in the middle, so that the geometry all about you remains the same no matter how you move.

This is perhaps the most surprising example of a well-defined geometry. But Euclidean geometry is, in its way, just as remarkable. Euclidean geometry provides a wonderful illustration of the

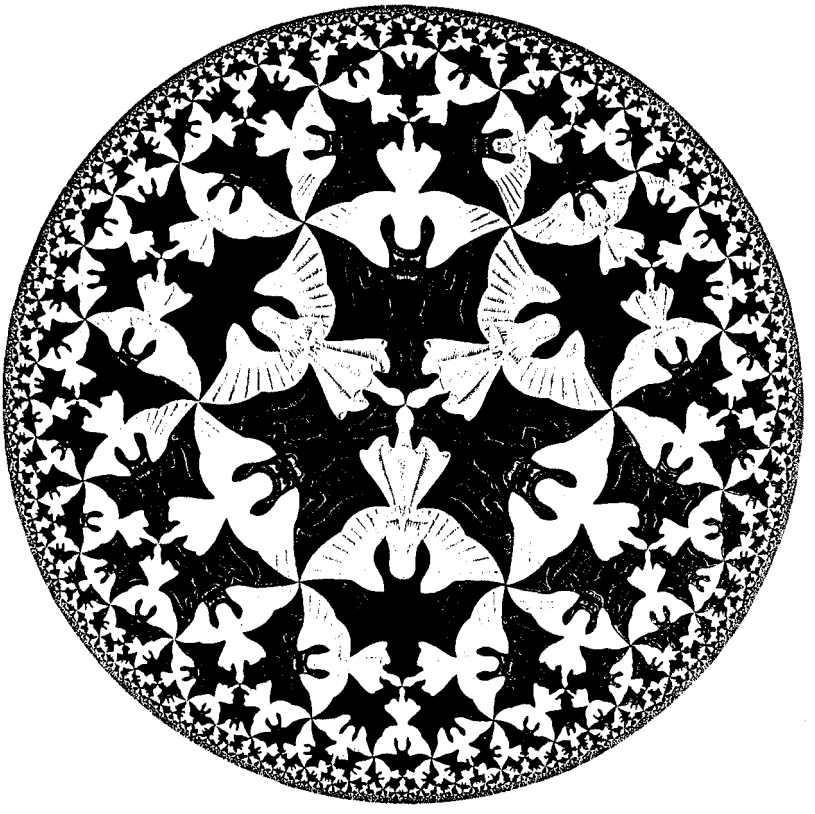


FIGURE 17. *Circle Limit 4* by M. C. Escher.

relation between mathematics and physics. The geometry is a part of mathematics, but the Greeks also thought of it as a description of the way the world is. Indeed, it turns out to be an extraordinarily accurate description of the way the world actually is — not utterly accurate because Einstein's theory tells us that space-time is slightly curved in various ways, but it is an extraordinarily accurate description of the world nevertheless. People used to worry about whether or not other geometries were possible. In particular, they worried about what is known as *Euclid's fifth postulate.* This can be reformulated as the statement that, if there is a

line in a plane and there is a point outside that line, then there is a unique parallel to this line through that point. People used to think that maybe this could be proved from the other more obvious axioms of Euclidean geometry. It turns out that it is not possible, and from this the notion of non-Euclidean geometry arose.

In non-Euclidean geometries, the angles of a triangle do not add up to 180°. This is another example where you think that things must become more complicated because, in Euclidean geometry, the angles of a triangle do add up to 180°. But then, in the non-Euclidean geometry, if you take the sum of the angles of a triangle away from 180° you find that this difference is proportional to the area of the triangle. In Euclidean geometry, the area of a triangle is a complicated thing to write down in terms of angles and lengths. In non-Euclidean, Lobachevskian, geometry, there is this wonderfully simple formula, due to Johann Heinrich Lambert, which enables the area of the triangle to be found. In fact, Lambert derived his formula before non-Euclidean geometry was discovered and I have never quite understood that!

There is another very important point here, which concerns the real numbers. These are absolutely fundamental to Euclidean geometry. They were essentially introduced by Eudoxus in the fourth century B.C. and they are still with us. They are the numbers that describe all our physics. As we shall see later, complex numbers are needed too, but they are based upon real numbers.

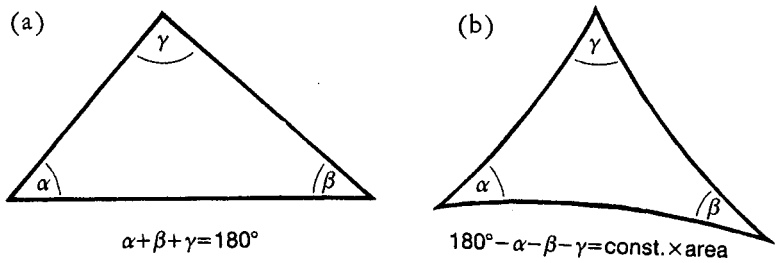


FIGURE 18. (a) A triangle in Euclidean space. (b) A triangle in Lobachevskian space.

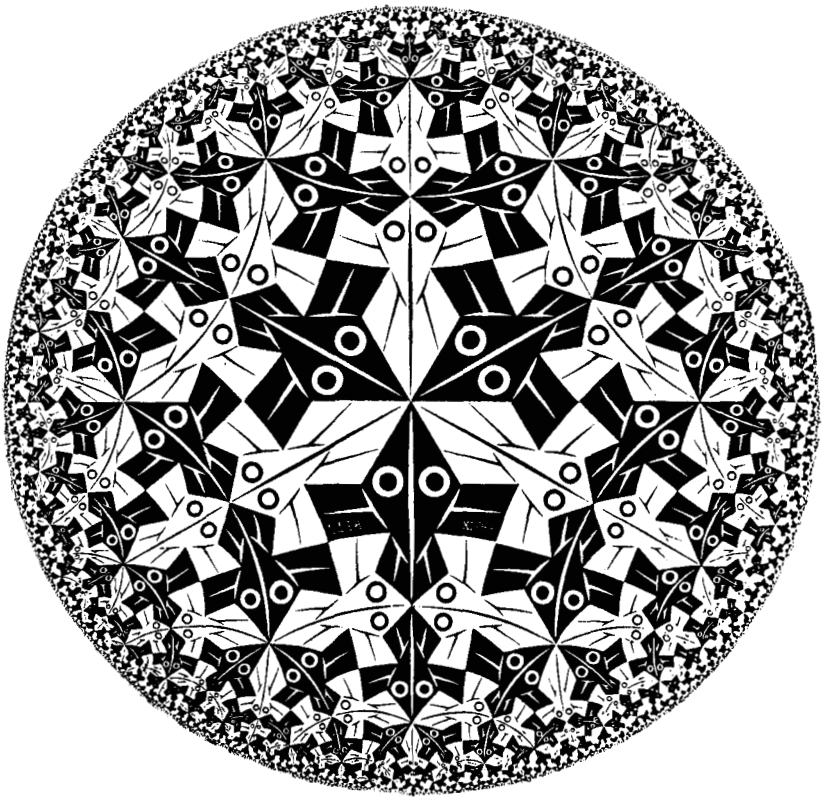


FIGURE 19. *Circle Limit 1* by M. C. Escher.

Let us look at another of the Escher prints to see how the Lobachevski geometry works. Figure 19 is even nicer than figure 18 for understanding this geometry because the "straight lines" are more obvious. They are represented by arcs of circles that meet the boundary at right angles, So, if you were a Lobachevskian person, and lived in this geometry, what you would think of as a straight line would be one of these arcs. You can see these clearly in figure 19 —some of them are straight lines through the centre of the diagram, but all the others are curved arcs. Some of these "straight lines" are shown in figure 20. In that diagram, I have

indicated a point that is not on the straight line crossing the diagram. Lobachevskian people can draw two (and more) separate lines parallel to the diagonal through that point, as I have indicated. Thus, the parallel postulate is violated in this geometry. Furthermore, you can draw triangles and work out the sums of angles of the triangles in order to work out their areas. This may give you some taste for the nature of hyperbolic geometry.

Let me give another example. I said that I like hyperbolic, Lobachevskian geometry the best. One of the reasons is that its group of symmetries is exactly the same as the one that we have already encountered, namely, the Lorentz group —the group of Special Relativity, or the symmetry group of the light-cones of relativity. To see that it is, I have drawn a light-cone in figure 21 but with some extra bits drawn on. I have had to suppress one of the space dimensions in order to depict it in three-dimensional space.

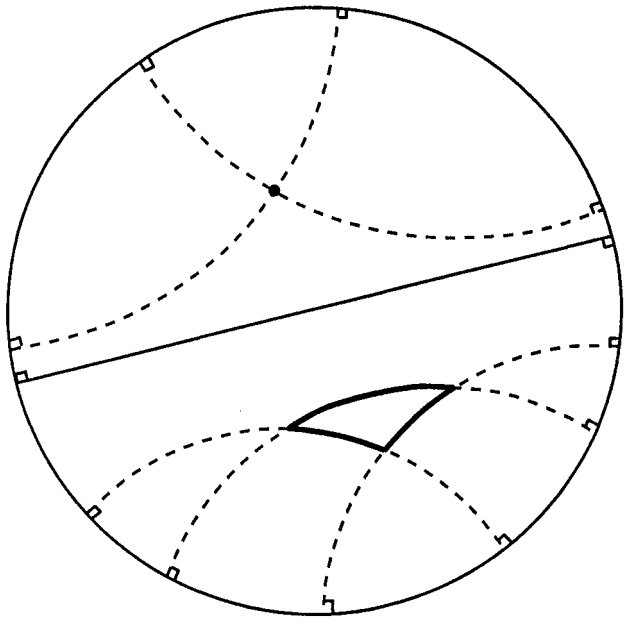


FIGURE 20. Illustrating the geometry of Lobachevskian (hyperbolic) space.

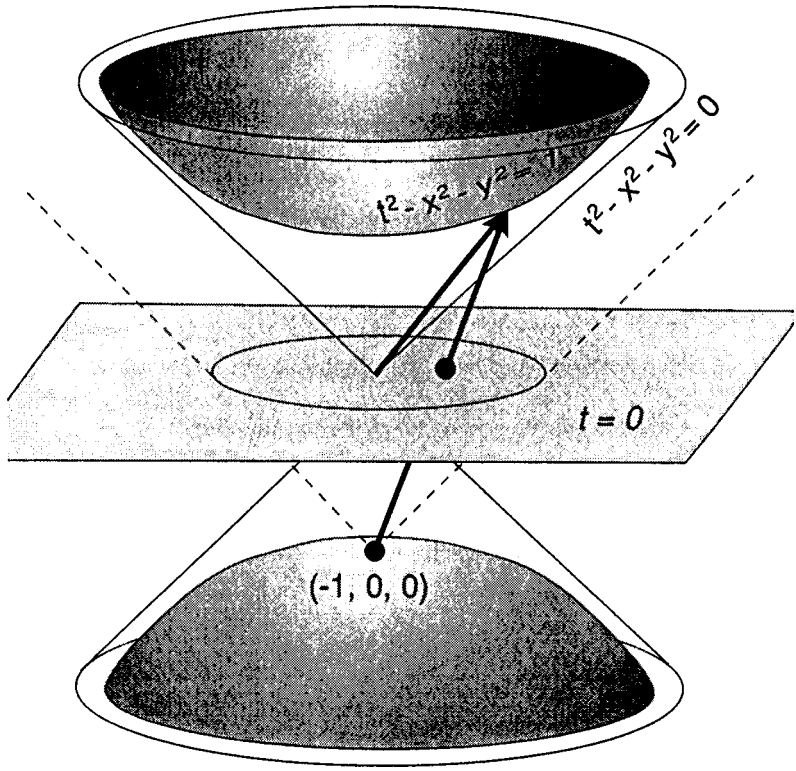


FIGURE 21. Illustrating Lobachevskian space-time geometry. The circle drawn on the plane $t = 0$ is the Poincaré disc.

The light-cone is described by the usual equation shown on the diagram:

$$t^2 - x^2 - y^2 = 0$$

The bowl-shaped surfaces shown above and below are located at "unit distance" from the origin in this Minkowskian geometry. ("Distance" in Minkowskian geometry is actually *time*—the proper time that is physically measured by moving clocks.) Thus, these surfaces represent the surface of a "sphere" for the Minkowskian geometry. It turns out that the intrinsic geometry of the

"sphere" is actually Lobachevskian (hyperbolic) geometry. If you consider an ordinary sphere in Euclidean space, you can rotate it around, and the group of symmetries is that of rotating the sphere around. In the geometry of figure 21, the group of symmetries is the group of symmetries associated with the surface shown in the diagram, in other words, with the Lorentz group of rotations. This symmetry group describes how space and time transform when a particular point in space-time is fixed—rotating the space-time about in different ways. We now see, with this representation, that the group of symmetries of Lobachevskian space is essentially just the same as the Lorentz group.

We can carry out a Minkowskian version of the stereographic projection, as illustrated in figure 10(b), The equivalent of the south pole is now the point at $(-1, 0, 0)$ and we project points from the upper bowl-shaped surface to the flat surface at $t = 0$, which is the equivalent of the equatorial plane in figure 10(b). In this procedure, we project all the points on the upper surface to the plane at $t = 0$. The projected points all lie inside a disc in the plane at $t = 0$, and this disc is sometimes referred to as the Poincaré disc. This is precisely how Escher's circle limit diagrams come about—the entire hyperbolic (Lobachevskian) surface has been mapped onto the Poincaré disc. Furthermore, this mapping does all the things that the projection of figure 10(b) does—it preserves angles and circles and it all comes out geometrically in a very nice way. Well, perhaps I am getting carried away here by my enthusiasm—I am afraid that is what mathematicians do when they get stuck into something.

The intriguing point is that, when you get carried away by something like the geometry of the above problem, the analysis and the results have an elegance that sustains them, while analyses that do not possess this mathematical elegance peter out. There is something particularly elegant about hyperbolic geometry. It would be awfully nice, at least to the likes of me, if the Universe were built that way too. Let me say that I have various other rea-

sons for believing this. Many other people do not like these open, hyperbolic universes —they frequently prefer closed universes, such as those illustrated in figure 16(c), which are nice and cosy. Well, actually, the closed universes are pretty big still. Alternatively, many people like flat world models as in figure 16(a) because there is a certain type of theory of the early Universe, the *inflationary theory,* which suggests that the geometry of the Universe should be flat. I should say that I do not really believe these theories.

These three standard types of model of the Universe are what are known as the Friedman models and they are characterised by the fact that they are very, very symmetrical. They are initially expanding models but at any moment the Universe is perfectly uniform everywhere. This assumption is built into the structure of the Friedman models and it is known as the *cosmological principle.* Wherever you are, the Friedman universe looks the same in all directions. It turns out that our actual Universe is like this to a remarkable degree. If Einstein's equations are right, and I have shown that the theory agrees with observation to a quite remarkable degree, then we are led to take the Friedman models seriously. All these models have this awkward feature, known as the Big Bang, where everything goes wrong, right at the beginning. The Universe is infinitely dense, infinitely hot, and so on —something has gone badly wrong with the theory. Nonetheless, if you accept that this very hot, dense phase took place, you can make predictions about what the thermal content of the Universe should be today and one of these expectations is that there should be a uniform background of black-body radiation all about us at the present day. Precisely this type of radiation was discovered by Arno Penzias and Robert Wilson in 1965. The most recent observations of the spectrum of this radiation, which is known as the Cosmic Microwave Background Radiation, by the COBE satellite show that it has precisely a black body spectrum with quite extraordinary precision.

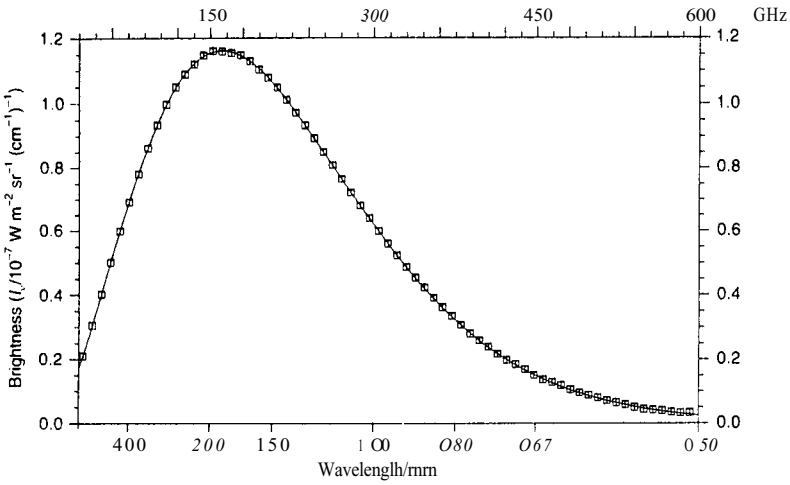


FIGURE 22. The precise agreement between the COBE measurements of the spectrum of the cosmic microwave background radiation and the expected "thermal" nature of the Big Bang's radiation.

All cosmologists interpret the existence of this radiation as evidence that our Universe went through a hot, dense phase. This radiation is thus telling us something about the nature of the early Universe —it is not telling us everything, but something like the Big Bang did take place. In other words, the Universe must have been very like the models illustrated in figure 16.

There is one other very important discovery made by the COBE satellite and this is that, although the Cosmic Microwave Background Radiation is remarkably uniform and its properties can all be accounted for very beautifully mathematically, the Universe is not quite perfectly uniform. There are tiny but real irregularities in the distribution of the radiation over the sky. In fact, we expect that these tiny irregularities must be present in the early Universe —we are here to observe the Universe and we are certainly not just a uniform smudge. The Universe is probably more like the pictures illustrated in figure 23. To show how open-minded I am, I am using as examples both an open and a closed Universe.

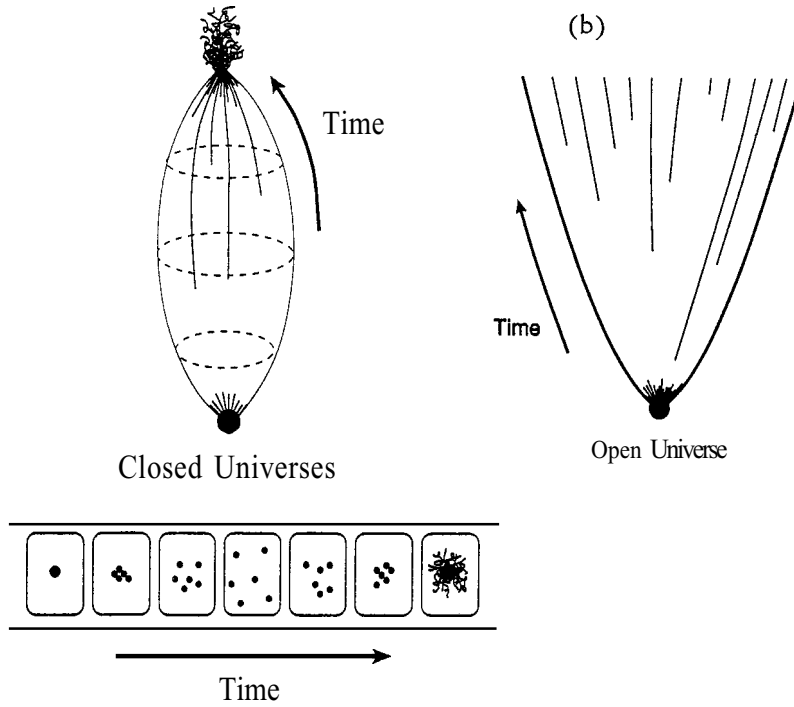


FIGURE 23. (a) The evolution of a closed world model with the formation of black holes as objects of various types reach the end points of their evolution. It can be seen that there is expected to be a horrible mess at the Big Crunch. The sequence of events is also shown as a "film-strip." (b) The evolution of the open model with the formation of black holes.

In the closed Universe, the irregularities will develop to form real observable structures —stars, galaxies, and the like —and, after a while, black holes will form, through the collapse of stars, through the accumulation of mass at the centres of galaxies, and so on. These black holes all have singular centres, much like the Big Bang in reverse. However, it is not as simple as that. According to the picture we have developed, the initial Big Bang is a nice, symmetrical, uniform state but the end point of the closed model is a horrible mess —all the black holes finally coming together and producing an incredible jumble at the final Big Crunch, as in

figure 23(a). The evolution of this closed model is illustrated schematically by the film strip shown in figure 23(b). In the case of an open universe model, the black holes are still formed as well; there is still an initial singularity and singularities formed at the centres of the black holes.

I emphasise these features of the standard Friedman models to show that there is a great difference between what we seem to see in the initial state and what we expect to find in the remote future. This problem is connected with the fundamental law of physics known as the Second Law of Thermodynamics.

We can understand this law in simple everyday terms. Imagine a glass of wine perched on the edge of the table. It might fall off the table, smash to pieces, and the wine spill all over the carpet (figure 24). There is nothing in Newtonian physics that tells us that the reverse process cannot happen. However, that is never observed —you never see wine glasses reassembling themselves and the wine being sucked up out of the carpet and into the reassembled glass. So far as the detailed laws of physics are concerned, the one direction of time is just as good as the other. To understand this difference, we need the Second Law of Thermo-

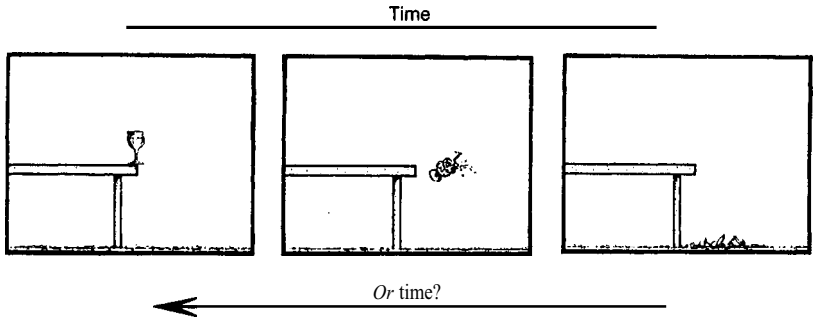


FIGURE 24. The laws of mechanics are time-reversible; yet the time-ordering of such a scene from the right frame to the left is something that is never experienced, whereas that from the left frame to the right would be commonplace.

dynamics, which tells us that the entropy of the system increases with time. This quantity called entropy is lower when the glass is on the table as compared with when it is shattered on the floor. According to the Second Law of Thermodynamics, the entropy of the system has increased. Roughly speaking, entropy is a measure of the disorder of a system. To express this concept more precisely, we have to introduce the concept of a *phase space.*

A phase space is a space of an enormous number of dimensions and each point of this multidimensional space describes the positions and momenta of all the particles that make up the system under consideration. In figure 25, we have selected a particular point in this huge phase space that represents where all the particles are located and how they are moving. As the system of particles evolves, the point moves to somewhere else in the phase space and I have shown it wiggling about from one point in phase space to another.

This wiggly line represents the ordinary evolution of the system of particles. There is no entropy there yet. To get entropy, we

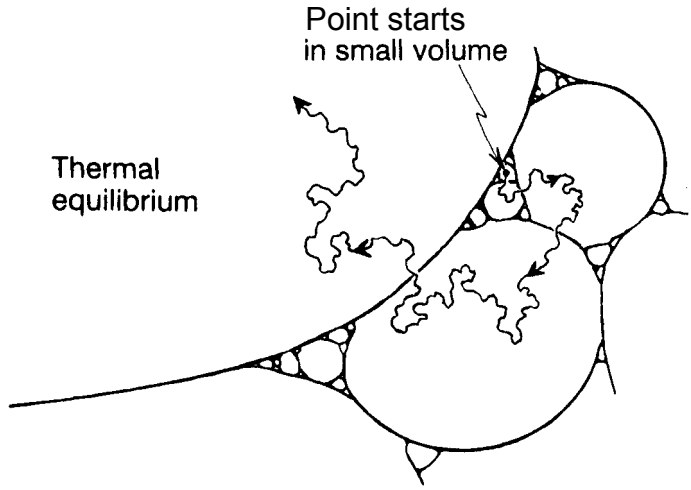


FIGURE 25. The Second Law of Thermodynamics in action: as time evolves, the phase-space point enters compartments of larger and larger volumes. Consequently, the entropy continually increases.

have to draw little bubbles around regions by lumping together those different states that you cannot tell apart. That may seem a bit obscure —what do you mean by "Cannot tell apart"? Surely that depends upon those who are looking and how carefully they look? Well, it is one of the slightly tricky questions of theoretical physics to say exactly what you mean by entropy. Essentially, what is meant is that you have to group states together according to what is known as "coarse-graining," that is, according to those things that you cannot tell apart. You take all those that, say, lie in this region here and lump them together, you look at the volume of that region, take the logarithm of the volume, and multiply it by the constant known as Boltzmann's constant, and that is the entropy. What the Second Law of Thermodynamics tells us is that the entropy increases. What it is telling you is actually something rather silly —all it is saying is that, if the system starts off in a little tiny box and it is allowed to evolve, it moves into bigger and bigger boxes. It is very likely that this happens because, if you look at the problem carefully, the bigger boxes are absolutely stupendously huger than the neighbouring little boxes. So, if you find yourself in one of the big boxes, there is absolutely no chance of getting back into a smaller box. That is all there is to it. The system just wanders about in phase space getting into bigger and bigger boxes. That is what the Second Law is telling us. Or is it?

Actually, that is only half the explanation. It tells us that, if we know the state of the system now, we can tell the most likely state in the future. But it tells us the completely wrong answer if we try to use the same argument backwards. Suppose the glass is sitting on the edge of the table. We can ask, "What is the most likely way by which it could have got there?" If you use the argument we have just given backwards, you would conclude that the most likely thing is that it started as a great mess on the carpet and then picked itself up off the carpet and reassembled itself on the table. This is clearly not the correct explanation —the correct ex-

planation is that someone put it there. And that person put it there for some reason, which was in turn due to some other reason, and so on. The chain of reasoning goes back and back to lower and lower entropy states in the past. The correct physical curve is illustrated in figure 26 —the entropy goes down and down and down in the past.

The reason why the entropy increases in the future is explained by moving into larger and larger boxes — why it goes down in the past is something completely different. There must have been something that pulled it down in the past. What pulled it down in the past? As we go into the past, the entropy gets smaller and smaller until eventually we end up at the Big Bang.

There must have been something very, very special about the Big Bang, but exactly what that was is a controversial issue. One popular theory, which I said I did not believe but which a lot of people are very keen on, is the idea of the inflationary universe. The idea is that the Universe is so uniform on the large scale because of something that was supposed to have taken place in the

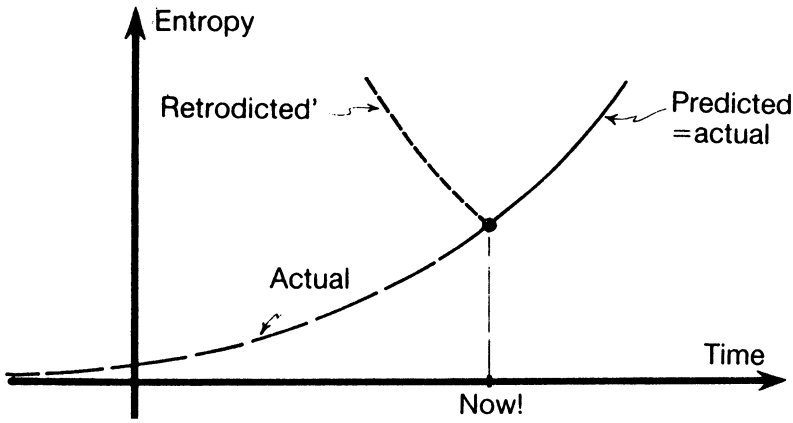


FIGURE 26. If we use the argument depicted in figure 25 in the reverse direction in time, we "retrodict" that the entropy should also increase in the *past,* from its value now. This is in gross contradiction with observation.

very earliest phases of the expansion of the Universe. It is supposed that an absolutely enormous expansion took place when the Universe was only about 10^{-36} seconds old and the idea is that, no matter what the Universe looked like in these very early stages, if you expand it by a huge factor of about 10^{60}, then it will look flat. In fact, this is one reason why these people like the flat Universe.

But, as it stands, the argument does not do what it is supposed to do —what you would expect in this initial state, if it were randomly chosen, would be a horrendous mess and, if you expand that mess by this huge factor, it still remains a complete mess. In fact, it looks worse and worse the more it expands (figure 27).

So the argument by itself does not explain why the Universe is so uniform. We need a theory that tells you what the Big Bang was really like. We do not know what that theory is, but we know that it has to involve a combination of large-scale and small-scale physics. It has to involve quantum physics as well as classical physics. Furthermore, I would claim that the theory must also have as one of its implications that the Big Bang was as uniform as we observe it to be. Maybe such a theory will end up producing a hyperbolic, Lobachevskian universe, like the picture I prefer, but I shall not insist upon that.

Let us return to the pictures of the closed and open Universes again (figure 28). In addition, I have included a picture of the formation of a black hole, which will be well known to the ex-

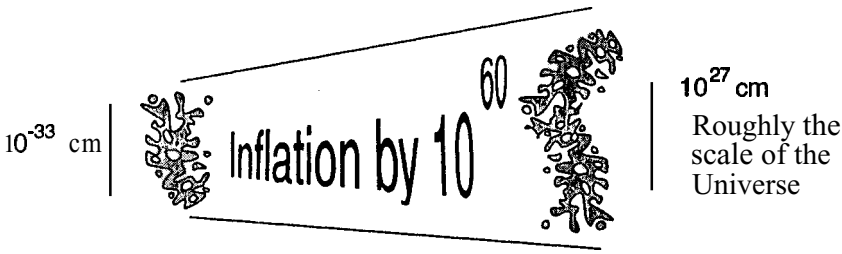


FIGURE 27. Illustrating the problem of the inflation of irregularities in the early Universe.

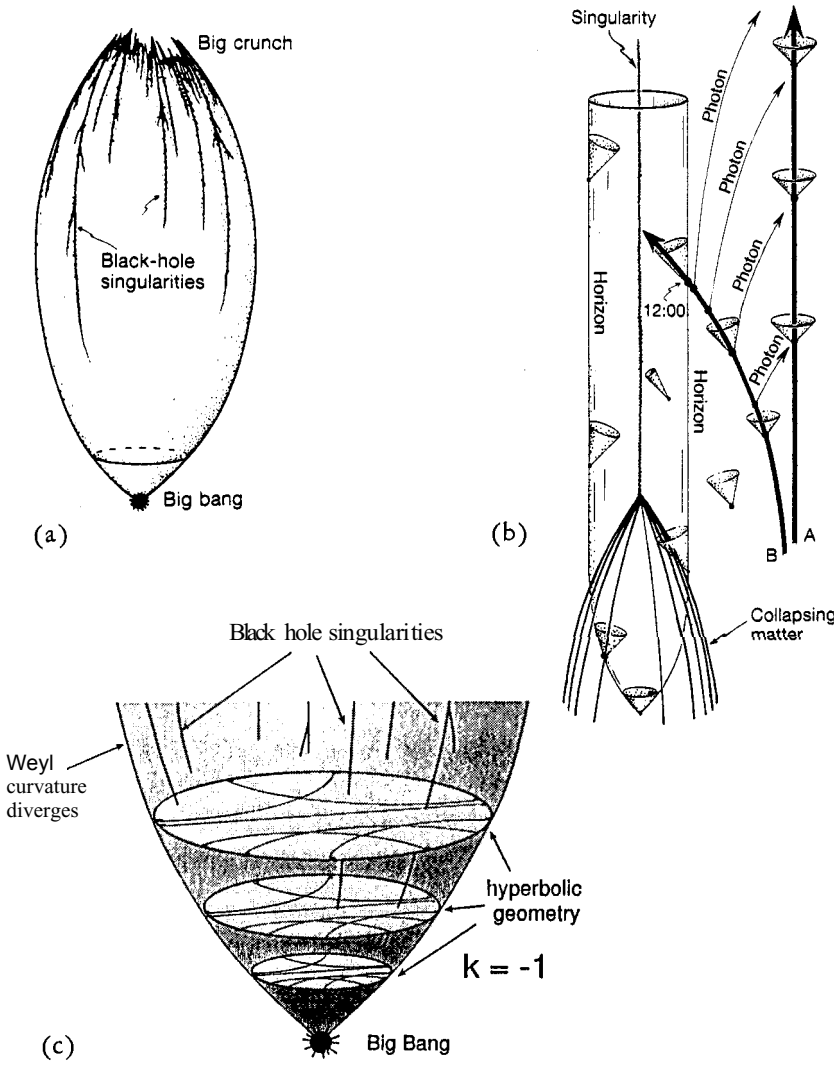


FIGURE 28. (a) The entire history of a closed universe that starts from a uniform low-entropy Big Bang with $Weyl = 0$ and ends with a high-entropy Big Crunch — representing the congealing of many black holes — with $Weyl \to \infty$. (b) A space-time diagram depicting collapse to a black hole. (c) The history of an open universe, again starting from a uniform low-entropy Big Bang with $Weyl = 0$.

perts. Matter collapsing into a black hole produces a singularity and that is what the dark lines on the space-time diagrams of the Universe represent. I want to introduce a hypothesis that I call the *Weyl curvature hypothesis.* This is not an implication of any known theory. As I have said, we do not know what the theory is, because we do not know how to combine the physics of the very large and the very small. When we do discover that theory, it should have as one of its consequences this feature that I have called the Weyl curvature hypothesis. Remember that the Weyl curvature is that bit of the Riemann tensor that causes distortions and tidal effects. For some reason we do not yet understand, in the neighbourhood of the Big Bang, the appropriate combination of theories must result in the Weyl tensor being essentially zero, or rather be constrained to be very small indeed.

That would give us a Universe like that shown in figure 28(c) and not like that in figure 28(a). The Weyl curvature hypothesis is time-asymmetrical and it applies only to the past type singularities and not to the future singularities. If the same rule of vanishing Weyl tensor that I have applied in the past also applied to the future of the Universe, in the closed model, you would end up with a dreadful looking Universe with as much mess in the past as in the future (figure 29). This looks nothing like the Universe we live in.

What is the probability that, purely by *chance,* the Universe had an initial singularity looking even remotely as it does? The probability is less than one part in $10^{10^{123}}$ Where does this estimate come from? It is derived from a formula by Jacob Beckenstein and Stephen Hawking concerning black hole entropy and, if you apply it in this particular context, you obtain this enormous answer. It depends how big the Universe is and, if you adopt my own favorite Universe, the number is, in fact, infinite.

What does this say about the precision that must be involved in setting up the Big Bang? It is really very, very extraordinary. I have illustrated the probability in this cartoon of the Creator,

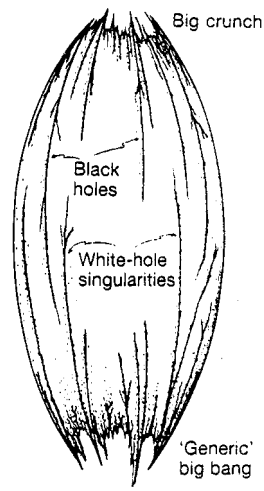


FIGURE 29. If the constraint $Weyl=0$ is removed, then we have a high-entropy Big Bang also, with $Weyl\to\infty$ there. Such a universe would be riddled with white holes, and there would be no Second Law of Thermodynamics, in gross contradiction with experience.

finding a very tiny point in that phase space that represents the initial conditions from which our Universe must have evolved if it is to resemble remotely the one we live in (figure 30). To find it, the Creator has to locate that point in phase space to an accuracy of one part in $10^{10^{123}}$. If I were to put one zero on each elementary particle in the Universe, I still could not write the number down in full. It is a stupendous number.

I have been talking about precision —how mathematics and physics fit together with extraordinary accuracy. I have also talked about the Second Law of Thermodynamics, which is often thought of as a rather floppy law —it concerns randomness and chance, and yet there is something very precise hiding underneath this law. As applied to the Universe, it has to do with the precision with which the initial state was set up. This precision must be something to do with the union of quantum theory and general relativity, a theory we do not have.

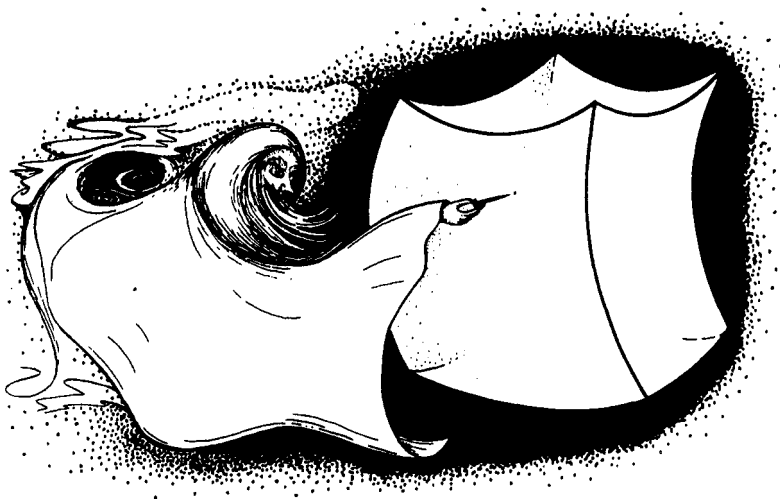


FIGURE 30. In order to produce a universe resembling the one in which we live, the Creator would have to aim for an absurdly tiny volume of the phase space of possible universes —about $1/10^{10^{123}}$ of the entire volume, for the situation under consideration. (The pin and the spot aimed for are not drawn to scale!)