The Inflationary Universe

A new theory of cosmology suggests that the observable universe is embedded in a much larger region of space that had an extraordinary growth spurt a fraction of a second after the primordial big bang

By Alan H. Guth and Paul J. Steinhardt

In the past few years certain flaws in the standard big-bang theory of cosmology have led to the development of a new model of the very early history of the universe. The model, known as the inflationary universe, agrees precisely with the generally accepted description of the observed universe for all times after the first 10-30 second. For this first fraction of a second, however, the story is dramatically different. According to the inflationary model, the universe had a brief period of extraordinarily rapid inflation, or expansion, during which its diameter increased by a factor perhaps 1050 times larger than had been thought. In the course of this stupendous growth spurt all the matter and energy in the universe could have been created from virtually nothing. The inflationary process also has important implications for the present universe. If the new model is correct, the observed universe is only a very small fraction of the entire universe.

The inflationary model has many features in common with the standard big-bang model. In both models the universe began between 10 and 15 billion years ago as a primeval fireball of extreme density and temperature, and it has been expanding and cooling ever since. This picture has been successful in explaining many aspects of the observed universe, including the red-shifting of the light from distant galaxies, the cosmic microwave background radiation and the primordial abundances of the lightest elements. All these predictions have to do only with events that presumably took place after the first second, when the two models coincide.

Until about five years ago there were few serious attempts to describe the universe during its first second. The temperature in this period is thought to have been higher than 10 billion degrees Kelvin, and little was known about the properties of matter under such conditions. Relying on recent developments in the physics of elementary particles, however, cosmologists are now attempting to understand the history of the universe back to 10-45 second after its beginning. (At even earlier times the energy density would have been so great that Einstein's general theory of relativity would have to be replaced by a quantum theory of gravity, which so far does not exist.) When the standard big-bang model is extended to these earlier times, various problems arise. First, it becomes clear that the model requires a number of stringent, unexplained assumptions about the initial conditions of the universe. In addition most of the new theories of elementary particles imply that the standard model would lead to a tremendous overproduction of the exotic particles called magnetic monopoles (each of which corresponds to an isolated north or south magnetic pole).

The inflationary universe [theory] was invented to overcome these problems. The equations that describe the period of inflation have a very attractive feature: from almost any initial conditions the universe evolves to precisely the state that had to be assumed as the initial one in the standard model. Moreover, the predicted density of magnetic monopoles becomes small enough to be consistent with observations. In the context of the recent developments in elementary-particle theory the inflationary model seems to be a natural solution to many of the problems of the standard big-bang picture.

The standard big-bang model is based on several assumptions. First, it is assumed that the fundamental laws of physics do not change with time and that the effects of gravitation are correctly described by Einstein's general theory of relativity. It is also assumed that the early universe was filled with an almost perfectly uniform, expanding, intensely hot gas of elementary particles in thermal equilibrium. The gas filled all of space, and the gas and space expanded together at the same rate. When they are averaged over large regions, the densities of matter and energy have remained nearly uniform from place to place as the universe has evolved. It is further assumed that any changes in the state of the matter and the radiation have been so smooth that they have had a negligible effect on the thermodynamic history of the universe. The violation of the last assumption is a key to the inflationary-universe model.

The big-bang model leads to three important, experimentally testable predictions. First, the model predicts that as the universe expands, galaxies recede from one another with a velocity proportional to the distance between them. In the 1920's Edwin P. Hubble inferred just such an expansion law from his study of the red shifts of distant galaxies. Second, the big-bang model predicts that there should be a background of microwave radiation bathing the universe as a remnant of the intense heat of its origin. The universe became transparent to this radiation several hundred thousand years after the big bang. Ever since then the matter has been clumping into stars, galaxies, and the like, but the radiation has simply continued to expand and red-shift, and in effect to cool. In 1964 Arno A. Penzias and Robert W. Wilson of the Bell Telephone Laboratories discovered a background of microwave radiation received uniformly from all directions with an effective temperature of about three degrees K. Third, the model leads to successful predictions of the formation of light atomic nuclei from protons and neutrons during the first minutes after the big bang. Successful predictions can be obtained in this way for the abundance of helium 4, deuterium, helium 3, and lithium 7. (Heavier nuclei are thought to have been produced much later in the interior of stars.)

Unlike the successes of the big-bang model, all of which pertain to events a second or more after the big bang, the problems all concern times when the universe was much less than a second old. One set of problems has to do with the special conditions the model requires as the universe emerged from the big bang.

The first problem is the difficulty of explaining the large-scale uniformity of the observed universe. The large-scale uniformity is most evident in the microwave background radiation, which is known to be uniform in temperature to about one part in 10,000. In the standard model the universe evolves much too quickly to allow this uniformity to be achieved by the usual processes whereby a system approaches thermal equilibrium. The reason is that no information or physical process can propagate faster than a light signal. At any given time there is a maximum distance, known as the horizon distance, that a light signal could have traveled since the beginning of the universe. In the standard model the sources of the microwave background radiation observed from opposite directions in the sky were separated from each other by more than 90 times the horizon distance when the radiation was emitted. Since the regions could not have communicated, it is difficult to see how they could have evolved conditions so nearly identical.

The puzzle of explaining why the universe appears to be uniform over distances that are large compared with the horizon distance is known as the horizon problem. It is not a genuine inconsistency of the standard model; if the uniformity is assumed in the initial conditions, the universe will evolve uniformly. The problem is that one of the most salient features of the observed universe—its large-scale uniformity—cannot be explained by the standard model; it must be assumed as an initial condition.

Even with the assumption of large-scale uniformity, the standard big-bang model requires yet another assumption to explain the nonuniformity observed on smaller scales. To account for the clumping of matter into galaxies, clusters of galaxies, superclusters of clusters, and so on, a spectrum of primordial inhomogeneities must be assumed as part of the initial conditions. The fact that the spectrum of inhomogeneities has no explanation is a drawback in itself, but the problem becomes even more pronounced when the model is extended back to 10-45 second after the big bang. The incipient clumps of matter develop rapidly with time as a result of their gravitational self-attraction, and so a model that begins at a very early time must begin with very small inhomogeneities. To begin at 10-45 second the matter must start in a peculiar state of extraordinary but not quite perfect uniformity. A normal gas in thermal equilibrium would be far too inhomogeneous, owing to the random motion of particles. This peculiarity of the initial state of matter required by the standard model is called the smoothness problem.

Another subtle problem of the standard model concerns the energy density of the universe. According to general relativity, the space of the universe can in principle be curved, and the nature of the curvature depends on the energy density. If the energy density exceeds a certain critical value, which depends on the expansion rate, the universe is said to be closed: space curves back on itself to form a finite volume with no boundary. (A familiar analogy is the surface of a sphere, which is finite in area and has no boundary.) If the energy density is less than the critical density, the universe is open: space curves but does not turn back on itself, and the volume is infinite. If the energy density is just equal to the critical density, the universe is flat: space is described by the familiar Euclidean geometry (again with infinite volume).

The ratio of the energy density of the universe to the critical density is a quantity cosmologists designate by the Greek letter Ù (omega). The value Ù = 1 (corresponding to a flat universe) represents a state of unstable equilibrium. If Ù was ever exactly equal to 1, it would remain exactly equal to 1 forever. If Ù differed slightly from 1 an instant after the big bang, however, the deviation from 1 would grow rapidly with time. Given this instability, it is surprising that Ù is measured today as being between .1 and 2. (Cosmologists are still not sure whether the universe is open, closed, or flat.) In order for Ù to be in this rather narrow range today, its value a second after the big bang had to equal 1 to within one part in 1015. The standard model offers no explanation of why Ù began so close to 1 but merely assumes the fact as an initial condition. This shortcoming of the standard model, called the flatness problem, was first pointed out in 1979 by Robert H. Dicke and P. James E. Peebles of Princeton University.

The successes and drawbacks of the big-bang model we have considered so far involve cosmology, astrophysics, and nuclear physics. As the big-bang model is traced backward in time, however, one reaches an epoch for which these branches of physics are no longer adequate. In this epoch all matter is decomposed into its elementary-particle constituents. In an attempt to understand this epoch cosmologists have made use of recent progress in the theory of elementary particles. Indeed, one of the important developments of the past decade has been the fusing of interests in particle physics, astrophysics and cosmology. The result for the big-bang model appears to be at least one more success and at least one more failure.

Perhaps the most important development in the theory of elementary particles over the past decade has been the notion of grand unified theories, the prototype of which was proposed in 1974 by Howard M. Georgi and Sheldon Lee Glashow of Harvard University. The theories are difficult to verify experimentally because their most distinctive predictions apply to energies far higher than those that can be reached with particle accelerators. Nevertheless, the theories have some experimental support, and they unify the understanding of elementary-particle interactions so elegantly that many physicists find them extremely attractive.

The basic idea of a grand unified theory is that what were perceived to be three independent forces—the strong, the weak, and the electromagnetic—are actually parts of a single unified force. In the theory a symmetry relates one force to another. Since experimentally the forces are very different in strength and character, the theory is constructed so that the symmetry is spontaneously broken in the present universe.

A spontaneously broken symmetry is one that is present in the underlying theory describing a system but is hidden in the equilibrium state of the system. For example, a liquid described by physical laws that are rotationally symmetric is itself rotationally symmetric: the distribution of molecules looks the same no matter how the liquid is turned. When the liquid freezes into a crystal, however, the atoms arrange themselves along crystallographic axes and the rotational symmetry is broken. One would expect that if the temperature of a system in a broken-symmetry state were raised, it could undergo a kind of phase transition to a state in which the symmetry is restored, just as a crystal can melt into a liquid. Grand unified theories predict such a transition at a critical temperature of roughly 1027 degrees.

One novel property of the grand unified theories has to do with the particles called baryons, a class whose most important members are the proton and the neutron. In all physical processes observed up to now the number of baryons minus the number of antibaryons does not change; in the language of particle physics the total baryon number of the system is said to be conserved. A consequence of such a conservation law is that the proton must be absolutely stable; because it is the lightest baryon, it cannot decay into another particle without changing the total baryon number. Experimentally the lifetime of the proton is known to exceed 1031 years.

Grand unified theories imply that baryon number is not exactly conserved. At low temperature, in the broken-symmetry phase, the conservation law is an excellent approximation, and the observed limit on the proton lifetime is consistent with at least many versions of grand unified theories. At high temperature, however, processes that change the baryon number of a system of particles are expected to be quite common.

One direct result of combining the big-bang model with grand unified theories is the successful prediction of the asymmetry of matter and antimatter in the universe. It is thought that all the stars, galaxies, and dust observed in the universe are in the form of matter rather than antimatter;

their nuclear particles are baryons rather than antibaryons. It follows that the total baryon number of the observed universe is about 1078. Before the advent of grand unified theories, when baryon number was thought to be conserved, this net baryon number had to be postulated as yet another initial condition of the universe. When grand unified theories and the big-bang picture are combined, however, the observed excess of matter over antimatter can be produced naturally by elementary-particle interactions at temperatures just below the critical temperature of the phase transition. Calculations in the grand unified theories depend on too many arbitrary parameters for a quantitative prediction, but the observed matter-antimatter asymmetry can be produced with a reasonable choice of values for the parameters.

A serious problem that results from combining grand unified theories with the big-bang picture is that a large number of defects are generally formed during the transition from the symmetric phase to the broken-symmetry phase. The defects are created when regions of symmetric phase undergo a transition to different broken-symmetry states. In an analogous situation, when a liquid crystallizes, different regions may begin to crystallize with different orientations of the crystallographic axes. The domains of different crystal orientation grow and coalesce, and it is energetically favorable for them to smooth the misalignment along their boundaries. The smoothing is often imperfect, however, and localized defects remain.

In the grand unified theories there are serious cosmological problems associated with pointlike defects, which correspond to magnetic monopoles, and surfacelike defects, called domain walls. Both are expected to be extremely stable and extremely massive. (The monopole can be shown to be about 1016 times as heavy as the proton.) A domain of correlated broken-symmetry phase cannot be much larger than the horizon distance at that time, and so the minimum number of defects created during the transition can be estimated. The result is that there would be so many defects after the transition that their mass would dominate the energy density of the universe and thereby speed up its subsequent evolution. The microwave background radiation would reach its present temperature of three degrees K. only 30,000 years after the big bang instead of 10 billion years, and all the successful predictions of the big-bang model would be lost. Thus any successful union of grand unified theories and the big-bang picture must incorporate some mechanism to drastically suppress the production of magnetic monopoles and domain walls.

The inflationary-universe model appears to provide a satisfactory solution to these problems. Before the model can be described, however, we must first explain a few more of the details of symmetry breaking and phase transitions in grand unified theories.

All modern particle theories, including the grand unified theories, are examples of quantum field theories. The best-known field theory is the one that describes electromagnetism. According to the classical (nonquantum) theory of electromagnetism developed by James Clerk Maxwell in the 1860's, electric and magnetic fields have a well-defined value at every point in space, and their variation with time is described by a definite set of equations. Maxwell's theory was modified early in the 20th century in order to achieve consistency with the quantum theory. In the classical theory it is possible to increase the energy of an electromagnetic field by any amount, but in the quantum theory the increases in energy can come only in discrete lumps, the quanta, which in this case are called photons. The photons have both wavelike and particlelike properties, but in the lexicon of modern physics they are usually called particles. In general the formulation of a quantum field theory begins with a classical theory of fields, and it becomes a theory of particles when the rules of the quantum theory are applied.

As we have already mentioned, an essential ingredient of grand unified theories is the phenomenon of spontaneous symmetry breaking. The detailed mechanism of spontaneous symmetry breaking in grand unified theories is simpler in many ways than the analogous mechanism in crystals. In a grand unified theory spontaneous symmetry breaking is accomplished by including in the formulation of the theory a special set of fields known as Higgs fields (after Peter W. Higgs of the University of Edinburgh). The symmetry is unbroken when all the Higgs fields have a value of zero, but it is spontaneously broken whenever at least one of the Higgs fields acquires a nonzero value. Furthermore, it is possible to formulate the theory in such a way that a Higgs field has a nonzero value in the state of lowest energy density, which in this context is known as the true vacuum. At temperatures greater than about 1027 degrees thermal fluctuations drive the equilibrium value of the Higgs field to zero, resulting in a transition to the symmetric phase.

We have now assembled enough background information to describe the inflationary model of the universe, beginning with the form in which it was first proposed by one of us (Guth) in 1980. Any cosmological model must begin with some assumptions about the initial conditions, but for the inflationary model the initial conditions can be rather arbitrary. One must assume, however, that the early universe included at least some regions of gas that were hot compared with the critical temperature of the phase transition and that were also expanding. In such a hot region the Higgs field would have a value of zero. As the expansion caused the temperature to fall it would become thermodynamically favorable for the Higgs field to acquire a nonzero value, bringing the system to its broken-symmetry phase.

For some values of the unknown parameters of the grand unified theories this phase transition would occur very slowly compared with the cooling rate. As a result the system could cool to well below 1027 degrees with the value of the Higgs field remaining at zero. This phenomenon, known as supercooling, is quite common in condensed-matter physics; water, for example, can be supercooled to more than 20 degrees below its freezing point, and glasses are formed by rapidly supercooling a liquid to a temperature well below its freezing point.

As the region of gas continued to supercool, it would approach a peculiar state of matter known as a false vacuum. This state of matter has never been observed, but it has properties that are unambiguously predicted by quantum field theory. The temperature, and hence the thermal component of the energy density, would rapidly decrease and the energy density of the state would be concentrated entirely in the Higgs field. A zero value for the Higgs field implies a large energy density for the false vacuum. In the classical form of the theory such a state would be absolutely stable, even though it would not be the state of lowest energy density. States with a lower energy density would be separated from the false vacuum by an intervening energy barrier, and there would be no energy available to take the Higgs field over the barrier.

In the quantum version of the model the false vacuum is not absolutely stable. Under the rules of the quantum theory all the fields would be continually fluctuating. As was first described by Sidney R. Coleman of Harvard, a quantum fluctuation would occasionally cause the Higgs field in a small region of space to “tunnel” through the energy barrier, nucleating a “bubble” of the broken-symmetry phase. The bubble would then start to grow at a speed that would rapidly approach the speed of light, converting the false vacuum into the broken-symmetry phase. The rate at which bubbles form depends sensitively on the unknown parameters of the grand unified theory; in the inflationary model it is assumed that the rate would be extremely low.

The most peculiar property of the false vacuum is probably its pressure, which is both large and negative. To understand why, consider again the process by which a bubble of true vacuum would grow into a region of false vacuum. The growth is favored energetically because the true vacuum has a lower energy density than the false vacuum. The growth also indicates, however, that the pressure of the true vacuum must be higher than the pressure of the false vacuum, forcing the bubble wall to grow outward. Because the pressure of the true vacuum is zero, the pressure of the false vacuum must be negative. A more detailed argument shows that the pressure of the false vacuum is equal to the negative value of its energy density (when the two quantities are measured in the same units).

The negative pressure would not result in mechanical forces within the false vacuum, because mechanical forces arise only from differences in pressure. Nevertheless, there would be gravitational effects. Under ordinary circumstances the expansion of the region of gas would be slowed by the mutual gravitational attraction of the matter within it. In Newtonian physics this attraction is proportional to the mass density, which in relativistic theories is equal to the energy density divided by the square of the speed of light. According to general relativity, the pressure also contributes to the attraction; to be specific, the gravitational force is proportional to the energy density plus three times the pressure. For the false vacuum the contribution made by the pressure would overwhelm the energy-density contribution and would have the opposite sign. Hence the bizarre notion of negative pressure leads to the even more bizarre effect of a gravitational force that is effectively repulsive. As a result the expansion of the region would be accelerated and the region would grow exponentially, doubling in diameter during each interval of about 10-34 second.

This period of accelerated expansion is called the inflationary era, and it is the key element of the inflationary model of the universe. According to the model, the inflationary era continued for 10-32 second or longer, and during this period the diameter of the universe increased by a factor of 1050 or more. It is assumed that after this colossal expansion the transition to the broken-symmetry phase finally took place. The energy density of the false vacuum was then released, resulting in a tremendous amount of particle production. The region was reheated to a temperature of almost 1027 degrees. (In the language of thermodynamics the energy released is called the latent heat; it is analogous to the energy released when water freezes.) From this point on the region would continue to expand and cool at the rate described by the standard big-bang model. A volume the size of the observable universe would lie well within such a region.

The horizon problem is avoided in a straightforward way. In the inflationary model the observed universe evolves from a region that is much smaller in diameter (by a factor of 1050 or more) than the corresponding region in the standard model. Before inflation begins the region is much smaller than the horizon distance, and it has time to homogenize and reach thermal equilibrium. This small homogeneous region is then inflated to become large enough to encompass the observed universe. Thus the sources of the microwave background radiation arriving today from all directions in the sky were once in close contact; they had time to reach a common temperature before the inflationary era began.

The flatness problem is also evaded in a simple and natural way. The equations describing the evolution of the universe during the inflationary era are different from those for the standard model, and it turns out that the ratio Ù is driven rapidly toward 1, no matter what value it had before inflation. This behavior is most easily understood by recalling that a value of Ù = 1 corresponds to a space that is geometrically flat. The rapid expansion causes the space to become flatter just as the surface of a balloon becomes flatter when it is inflated. The mechanism driving Ù toward 1 is so effective that one is led to an almost rigorous prediction: The value of Ù today should be very accurately equal to 1. Many astronomers (although not all) think a value of 1 is consistent with current observations, but a more reliable determination of Ù would provide a crucial test of the inflationary model.

In the form in which the inflationary model was originally proposed it had a crucial flaw: under the circumstances described, the phase transition itself would create inhomogeneities much more extreme than those observed today. As we have already described, the phase transition would take place by the random nucleation of bubbles of the new phase. It can be shown that the bubbles would always remain in finite clusters disconnected from one another, and that each cluster would be dominated by a single largest bubble. Almost all the energy in the cluster would be initially concentrated in the surface of the largest bubble, and there is no apparent mechanism to redistribute energy uniformly. Such a configuration bears no resemblance to the observed universe.

For almost two years after the invention of the inflationary-universe model it remained a tantalizing but clearly imperfect solution to a number of important cosmological problems. Near the end of 1981, however, a new approach was developed by A. D. Linde of the P. N. Lebedev Physical Institute in Moscow and independently by Andreas Albrecht and one of us (Steinhardt) of the University of Pennsylvania. This approach, known as the new inflationary universe, avoids all the problems of the original model while maintaining all its successes.

The key to the new approach is to consider a special form of the energy-density function that describes the Higgs field. Quantum field theories with energy-density functions of this type were first studied by Coleman, working in collaboration with Erick J. Weinberg of Columbia University. In contrast to the more typical case…, there is no energy barrier separating the false vacuum from the true vacuum; instead the false vacuum lies at the top of a rather flat plateau. In the context of grand unified theories such an energy-density function is achieved by a special choice of parameters. As we shall explain below, this energy-density function leads to a special type of phase transition that is sometimes called a slow-rollover transition.

The scenario begins just as it does in the original inflationary model. Again one must assume the early universe had regions that were hotter than about 1027 degrees and were also expanding. In these regions thermal fluctuations would drive the equilibrium value of the Higgs fields to zero and the symmetry would be unbroken. As the temperature fell it would become thermodynamically favorable for the system to undergo a phase transition in which at least one of the Higgs fields acquired a nonzero value, resulting in a broken-symmetry phase. As in the previous case, however, the rate of this phase transition would be extremely low compared with the rate of cooling. The system would supercool to a negligible temperature with the Higgs field remaining at zero, and the resulting state would again be considered a false vacuum.

The important difference in the new approach is the way in which the phase transition would take place. Quantum fluctuations or small residual thermal fluctuations would cause the Higgs field to deviate from zero. In the absence of an energy barrier the value of the Higgs field would begin to increase steadily; the rate of increase would be much like that of a ball rolling down a hill of the same shape as the curve of the energy-density function, under the influence of a frictional drag force. Since the energy-density curve is almost flat near the point where the Higgs field vanishes, the early stage of the evolution would be very slow. As long as the Higgs field remained close to zero, the energy density would be almost the same as it is in the false vacuum. As in the original scenario, the region would undergo accelerated expansion, doubling in diameter every 10-34 second or so. Now, however, the expansion would cease to accelerate when the value of the Higgs field reached the steeper part of the curve. By computing the time required for the Higgs field to evolve, the amount of inflation can be determined. An expansion factor of 1050 or more is quite plausible, but the actual factor depends on the details of the particle theory one adopts.

So far the description of the phase transition has been slightly oversimplified. There are actually many different broken-symmetry states, just as there are many possible orientations for the axes of a crystal. There are a number of Higgs fields, and the various broken-symmetry states are distinguished by the combination of Higgs fields that acquire nonzero values. Since the fluctuations that drive the Higgs fields from zero are random, different regions of the primordial universe would be driven toward different broken-symmetry states, each region forming a domain with an initial radius of roughly the horizon distance. At the start of the phase transition the horizon distance would be about 10-24 centimeter. Once the domain formed, with the Higgs fields deviating slightly from zero in a definite combination, it would evolve toward one of the stable broken-symmetry states and would inflate by a factor of 1050 or more. The size of the domain after inflation would then be greater than 1026 centimeters. The entire observable universe, which at that time would be only about 10 centimeters across, would be able to fit deep inside a single domain.

In the course of this enormous inflation any density of particles that might have been present initially would be diluted to virtually zero. The energy content of the region would then consist entirely of the energy stored in the Higgs field. How could this energy be released? Once the Higgs field evolved away from the flat part of the energy-density curve, it would start to oscillate rapidly about the true-vacuum value. Drawing on the relation between particles and fields implied by quantum field theory, this situation can also be described as a state with a high density of Higgs particles. The Higgs particles would be unstable, however: they would rapidly decay to lighter particles, which would interact with one another and possibly undergo subsequent decays. The system would quickly become a hot gas of elementary particles in thermal equilibrium, just as was assumed in the initial conditions for the standard model. The reheating temperature is calculable and is typically a factor of between two and 10 below the critical temperature of the phase transition. From this point on, the scenario coincides with that of the standard big-bang model, and so all the successes of the standard model are retained.

Note that the crucial flaw of the original inflationary model is deftly avoided. Roughly speaking, the isolated bubbles that were discussed in the original model are replaced here by the domains. The domains of the slow-rollover transition would be surrounded by other domains rather than by false vacuum, and they would tend not to be spherical. The term “bubble” is therefore avoided. The key difference is that in the new inflationary model each domain inflates in the course of its formation, producing a vast essentially homogeneous region within which the observable universe can fit.

Since the reheating temperature is near the critical temperature of the grand-unified-theory phase transition, the matter-antimatter asymmetry could be produced by particle interactions just after the phase transition. The production mechanism is the same as the one predicted by grand unified theories for the standard big-bang model. In contrast to the standard model, however, the inflationary model does not allow the possibility of assuming the observed net baryon number of the universe as an initial condition; the subsequent inflation would dilute any initial baryon-number density to an imperceptible level. Thus the viability of the inflationary model depends crucially on the viability of particle theories, such as the grand unified theories, in which baryon number is not conserved.

One can now grasp the solutions to the cosmological problems discussed above. The horizon and flatness problems are resolved by the same mechanisms as in the original inflationary-universe model. In the new inflationary scenario the problem of monopoles and domain walls can also be solved. Such defects would form along the boundaries separating domains, but the domains would have been inflated to such an enormous size that the defects would lie far beyond any observable distance. (A few defects might be generated by thermal effects after the transition, but they are expected to be negligible in number.)

Thus with a few simple ideas the improved inflationary model of the universe leads to a successful resolution of several major problems that plague the standard big-bang picture: the horizon, flatness, magnetic-monopole, and domain-wall problems. Unfortunately the necessary slow-rollover transition requires the fine tuning of parameters; calculations yield reasonable predictions only if the parameters are assigned values in a narrow range. Most theorists (including both of us) regard such fine tuning as implausible. The consequences of the scenario are so successful, however, that we are encouraged to go on in the hope we may discover realistic versions of grand unified theories in which such a slow-rollover transition occurs without fine tuning.

The successes already discussed offer persuasive evidence in favor of the new inflationary model. Moreover, it was recently discovered that the model may also resolve an additional cosmological problem not even considered at the time the model was developed: the smoothness problem. The generation of density inhomogeneities in the new inflationary universe was addressed in the summer of 1982 at the Nuffield Workshop on the Very Early Universe by a number of theorists, including James M. Bardeen of the University of Washington, Stephen W. Hawking of the University of Cambridge, So-Young Pi of Boston University, Michael S. Turner of the University of Chicago, A. A. Starobinsky of the L. D. Landau Institute of Theoretical Physics in Moscow and the two of us. It was found that the new inflationary model, unlike any previous cosmological model, leads to a definite prediction for the spectrum of inhomogeneities. Basically the process of inflation first smoothes out any primordial inhomogeneities that might have been present in the initial conditions. Then in the course of the phase transition inhomogeneities are generated by the quantum fluctuations of the Higgs field in a way that is completely determined by the underlying physics. The inhomogeneities are created on a very small scale of length, where quantum phenomena are important, and they are then enlarged to an astronomical scale by the process of inflation.

The predicted shape for the spectrum of inhomogeneities is essentially scale-invariant; that is, the magnitude of the inhomogeneities is approximately equal on all length scales of astrophysical significance. This prediction is comparatively insensitive to the details of the underlying grand unified theory. It turns out that a spectrum of precisely this shape was proposed in the early 1970's as a phenomenological model for galaxy formation by Edward R. Harrison of the University of Massachusetts at Amherst and Yakov B. Zel'dovich of the Institute of Physical Problems in Moscow, working independently. The details of galaxy formation are complex and are still not well understood, but many cosmologists think a scale-invariant spectrum of inhomogeneities is precisely what is needed to explain how the present structure of galaxies and galactic clusters evolved.

The new inflationary model also predicts the magnitude of the density inhomogeneities, but the prediction is quite sensitive to the details of the underlying particle theory. Unfortunately the magnitude that results from the simplest grand unified theory is far too large to be consistent with the observed uniformity of the cosmic microwave background. This inconsistency represents a problem, but it is not yet known whether the simplest grand unified theory is the correct one. In particular the simplest grand unified theory predicts a lifetime for the proton that appears to be lower than present experimental limits. On the other hand, one can construct more complicated grand unified theories that result in density inhomogeneities of the desired magnitude. Many investigators imagine that with the development of the correct particle theory the new inflationary model will add the resolution of the smoothness problem to its list of successes.

One promising line of research involves a class of quantum field theories with a new kind of symmetry called supersymmetry. Supersymmetry relates the properties of particles with integer angular momentum to those of particles with half-integer angular momentum; it thereby highly constrains the form of the theory. Many theorists think supersymmetry might be necessary to construct a consistent quantum theory of gravity, and to eventually unify gravity with the strong, the weak and the electromagnetic forces. A tantalizing property of models incorporating supersymmetry is that many of them give slow-rollover phase transitions without any fine tuning of parameters. The search is on to find a supersymmetry model that is realistic as far as particle physics is concerned and that also gives rise to inflation and to the correct magnitude for the density inhomogeneities.

In short, the inflationary model of the universe is an economical theory that accounts for many features of the observable universe lacking an explanation in the standard big-bang model. The beauty of the inflationary model is that the evolution of the universe becomes almost independent of the details of the initial conditions, about which little if anything is known. It follows, however, that if the inflationary model is correct, it will be difficult for anyone to ever discover observable consequences of the conditions existing before the inflationary phase transition. Similarly, the vast distance scales created by inflation would make it essentially impossible to observe the structure of the universe as a whole. Nevertheless, one can still discuss these issues, and a number of remarkable scenarios seem possible.

The simplest possibility for the very early universe is that it actually began with a big bang, expanded rather uniformly until it cooled to the critical temperature of the phase transition and then proceeded according to the inflationary scenario. Extrapolating the big-bang model back to zero time brings the universe to a cosmological singularity, a condition of infinite temperature and density in which the known laws of physics do not apply. The instant of creation remains unexplained. A second possibility is that the universe began (again without explanation) in a random, chaotic state. The matter and temperature distributions would be nonuniform, with some parts expanding and other parts contracting. In this scenario certain small regions that were hot and expanding would undergo inflation, evolving into huge regions easily capable of encompassing the observable universe. Outside these regions there would remain chaos, gradually creeping into the regions that had inflated.

Recently there has been some serious speculation that the actual creation of the universe is describable by physical laws. In this view the universe would originate as a quantum fluctuation, starting from absolutely nothing. The idea was first proposed by Edward P. Tryon of Hunter College of the City University of New York in 1973, and it was put forward again in the context of the inflationary model by Alexander Vilenkin of Tufts University in 1982. In this context “nothing”; might refer to empty space, but Vilenkin uses it to describe a state devoid of space, time and matter. Quantum fluctuations of the structure of space-time can be discussed only in the context of quantum gravity, and so these ideas must be considered highly speculative until a working theory of quantum gravity is formulated. Nevertheless, it is fascinating to contemplate that physical laws may determine not only the evolution of a given state of the universe but also the initial conditions of the observable universe.

As for the structure of the universe as a whole, the inflationary model allows for several possibilities. (In all cases the observable universe is a very small fraction of the universe as a whole; the edge of our domain is likely to lie 1035 or more light-years away.) The first possibility is that the domains meet one another and fill all space. The domains are then separated by domain walls, and in the interior of each wall is the symmetric phase of the grand unified theory. Protons or neutrons passing through such a wall would decay instantly. Domain walls would tend to straighten with time. After 1035 years or more smaller domains (possibly even our own) would disappear and larger domains would grow.

Alternatively, some versions of grand unified theories do not allow for the formation of sharp domain walls. In these theories it is possible for different broken-symmetry states in two neighboring domains to merge smoothly into each other. At the interface of two domains one would find discontinuities in the density and velocity of matter, and one would also find an occasional magnetic monopole.

A quite different possibility would result if the energy density of the Higgs fields were described by a [different type of] curve… As in the other two cases, regions of space would supercool into the false-vacuum state and undergo accelerated expansion. As in the original inflationary model, the false-vacuum state would decay by the mechanism of random bubble formation: quantum fluctuations would cause at least one of the Higgs fields in a small region of space to tunnel through the energy barrier… In contrast to the original inflationary scenario, the Higgs field would then evolve very slowly…to its true-vacuum value. The accelerated expansion would continue, and the single bubble would become large enough to encompass the observed universe. If the rate of bubble formation were low, bubble collisions would be rare. The fraction of space filled with bubbles would become closer to 1 as the system evolved, but space would be expanding so fast that the volume remaining in the false-vacuum state would increase with time. Bubble universes would continue to form forever, and there would be no way of knowing how much time had elapsed before our bubble was formed. This picture is much like the old steady-state cosmological model on the very large scale, and yet the interior of each bubble would evolve according to the big-bang model, improved by inflation.

From a historical point of view probably the most revolutionary aspect of the inflationary model is the notion that all the matter and energy in the observable universe may have emerged from almost nothing. This claim stands in marked contrast to centuries of scientific tradition in which it was believed that something cannot come from nothing. The tradition, dating back at least as far as the Greek philosopher Parmenides in the fifth centuryB.C., has manifested itself in modern times in the formulation of a number of conservation laws, which state that certain physical quantities cannot be changed by any physical process. A decade or so ago the list of quantities thought to be conserved included energy, linear momentum, angular momentum, electric charge, and baryon number.

Since the observed universe apparently has a huge baryon number and a huge energy, the idea of creation from nothing has seemed totally untenable to all but a few theorists. (The other conservation laws mentioned above present no such problems: the total electric charge and the angular momentum of the observed universe have values consistent with zero, whereas the total linear momentum depends on the velocity of the observer and so cannot be defined in absolute terms.) With the advent of grand unified theories, however, it now appears quite plausible that baryon number is not conserved. Hence only the conservation of energy needs further consideration.

The total energy of any system can be divided into a gravitational part and a nongravitational part. The gravitational part (that is, the energy of the gravitational field itself) is negligible under laboratory conditions, but cosmologically it can be quite important. The nongravitational part is not by itself conserved; in the standard big-bang model it decreases drastically as the early universe expands, and the rate of energy loss is proportional to the pressure of the hot gas. During the era of inflation, on the other hand, the region of interest is filled with a false vacuum that has a large negative pressure. In this case the nongravitational energy increases drastically. Essentially all the nongravitational energy of the universe is created as the false vacuum undergoes its accelerated expansion. This energy is released when the phase transition takes place, and it eventually evolves to become stars, planets, human beings, and so forth. Accordingly, the inflationary model offers what is apparently the first plausible scientific explanation for the creation of essentially all the matter and energy in the observable universe.

Under these circumstances the gravitational part of the energy is somewhat ill-defined, but crudely speaking one can say that the gravitational energy is negative, and that it precisely cancels the nongravitational energy. The total energy is then zero and is consistent with the evolution of the universe from nothing.

If grand unified theories are correct in their prediction that baryon number is not conserved, there is no known conservation law that prevents the observed universe from evolving out of nothing. The inflationary model of the universe provides a possible mechanism by which the observed universe could have evolved from an infinitesimal region. It is then tempting to go one step further and speculate that the entire universe evolved from literally nothing.