Vacuum theory: a possible solution to the singularity problem of cosmology

Ya. B. Zel'dovich

Usp. Fiz. Nauk 133, 479–503 (March 1981)

The present state of the question of the vacuum energy, cosmological constant, and vacuum polarization in a curved non-Euclidean space-time, i.e., in a gravitational field, is discussed. A cosmological model of the early Universe that is a self-consistent solution of the Einstein equations with vacuum polarization on the right-hand side has many attractive features.

PACS numbers: 95.30.Sf, 98.80.Bp

CONTENTS

1. INTRODUCTION

The reader is right to think that there may be an error in the title of this paper; for does a vacuum theory exist? Can one say something meaningful about vacuum, i.e., about space containing nothing? It is clear to everyone that it is the bodies and matter filling space that should be investigated experimentally and theoretically. Modern science has achieved stupendous sucesses, first reducing the entire manifold of substances to combinations of a comparatively small number of chemical elements, and then, at the beginning of this century, to combinations of three species of elementary particle—protons, neutrons, and electrons.

These particles are coupled to one another by definite forces. In particular, the coupling of electrons in atoms, the chemical bond in molecules, and the forces that combine molecules in solids and liquids are all manifestations of the electromagnetic interaction of electrons with nuclei and also electrons and nuclei with one another. Of course, it is important that the electrons satisfy quantum mechanics and the Pauli principle.

But the theory of electromagnetism leads to the conclusion that besides the static electric field surrounding charges there also exist specific solutions in the form of fields propagating freely in space and describing electromagnetic waves (radio waves, light, x rays, gamma rays). The next decisive step was made by Einstein in 1905, who showed that light must be regarded as a flux of photons, i.e., certain particles. This conclusion was based on an analysis of experimental data and was preceded by Planck's construction of the theory of thermal radiation. The quantum constant appeared for the first time in physics in Planck's work. Soon, many experiments on the chemical effect of light, the photoelectric effect, and the scattering of light by electrons confirmed the photon theory.

Much later, at the end of the twenties, the existence of photons was proved theoretically as a consequence of the systematic application of quantum theory to the electromagnetic field.

During the last twenty years, many new types of elementary particles have been discovered. These particles are "not necessary" to describe the objects and phenomena that we encounter in everyday life, this including the most complicated electronic devices, nuclear power plants, and biological and psychological phenomena. These particles could be called redundant. I remember the title of a paper relating to this period: "Why are μ mesons necessary?" But nature is a whole, and an all encompassing theory cannot be based on only a part of knowledge circumscribed by the power of existing accelerators.

Recently, we have begun to form a true estimate of the possible part played by all the different particles in the initial stage in the evolution of the Universe. To understand the existence of protons and neutrons, we need knowledge of all species of particles.

But let us return to the subject announced in the title.

Vacuum theory—does it really need for particles to be considered?

Let us recall the story of the man instructed to sell aerated water at a charitable bazaar. He was told to ask: "Which juice would you like with your water?" When a purchaser requested water without juice, our hero asked: "Without which juice? Without raspberry or without cherry juice?" It remains to add that the hero of this story was the physicochemist Ivan Alek-

 0038-5670/81/030216-15$01.10

seevich Kablukov (1875–1942) (Corresponding and then Honorary Member of the Academy of Sciences), whose absentmindedness was legendary.

Thus, vacuum is space in which there are no protons, electrons, photons, mesons, etc., so why is it necessary to know the properties of the particles which are not present? For the taste of aerated water can hardly depend on the raspberry juice which is *not* added to it....

But this simple argument is false, and it is false because nature satisfies quantum theory.

The application of quantum theory—and not only to atoms, plasmas, and radiation but also to vacuum—is extremely important for astronomy.

The present-day rich and complicated picture of the vacuum arises as a logical consequence of experiments and theories. The picture is the inevitable result of long and consistent work of scientists.

One last comment before we turn to the subject of the paper. The conception of the vacuum cannot be simplified by a mere redefinition of words. One cannot say that the vacuum, i.e., empty space, is devoid of all properties "by definition" and that all the complexities are connected with something that should not be called "vacuum."

We must define vacuum as space without any particles. Such a definition coincides with the condition of a minimum of the energy density in the given volume of space. If the energy E of some region of space is greater than the minimal value E_{min} for this region, then E can be represented as the sum $E_{min}+\Delta$, and the addition Δ can be regarded as the energy of the field or particles present in the given volume. Hence, a state with $E > E_{min}$ should not be called "vacuum." But the actual properties of the "minimal" state which is called the "vacuum" are dictated by the laws of physics, and we cannot insist that the minimum be zero or that the simplest possible situation be as simple as we wish.

Besides quantum theory, it is necessary to take into account the general theory of relativity. The basic proposition of general relativity is that the geometry of space at a particular point depends on the physical conditions at the point itself and also in the regions surrounding it. In the light of the special theory of relativity, it is more correct to speak of the "space–time" complex. It is not only the geometry of space (angles and lengths) but also the rate of clocks—the passage of time—that depends on the physical conditions.

Quantum theory and general relativity meet in a region of research that is at once difficult but extremely attractive and fundamentally important. In this paper I should like to consider some of the problems that arise in this region, qualitatively and with a minimum of simple equations. The reader should also be warned that, of necessity, the paper makes very varying demands with regard to the difficulty of the material.

Theoretical physicists can omit Secs. 2 and 3 and the first half of Sec. 4 without loss.

Experimentalists and students may find these sections helpful, if only to remind them of what they have been taught in physics courses.

The final section contains ideas which appeared in 1979–1980; they have not yet taken their final form, they are difficult, and they may be questioned. Nevertheless, I should like to give an idea of these topical unresolved problems and to draw attention to them, even at the price of an incomplete and inchoate exposition.

2. THE UNCERTAINTY PRINCIPLE AND ZERO-POINT OSCILLATIONS

The 19th century gave us a remarkable achievement: The experimental genius and depth of Faraday's understanding married to the theoretical insight of Maxwell led to the theory of electric and magnetic fields. As we have already said, a consequence of this theory was the prediction of electromagnetic waves propagating freely in empty space, i.e., in vacuum. Part of the spectrum, visible light, was known empirically for aeons.

The long-wavelength region of radio waves was discovered or, rather, specially created and exploited by mankind through the efforts of Hertz, Popov, and Marconi. Then came quantum theory. We shall pass over the history of its development. The modern view of electromagnetic waves emphasizes their similarity to a mechanical oscillator, i.e., a mass on a spring.

If one writes down the corresponding equations (which we shall not), it is found that the magnetic field plays the part of the spring, i.e., the energy of the magnetic field is analogous to the deformation of a spring, depending on the departure from the equilibrium position.

The energy of the electric field is the analog of the kinetic energy of a moving particle.

Thus, each definite oscillation mode of the electromagnetic field is analogous to the mechanical vibration of a mass on a spring. As yet, the concept of "mode" is not fully explained; we shall return to it.

Deferring therefore the particularization, let us use the analogy with mechanical vibrations. What do we know about the vibrations of a mass on a spring? In classical theory, the value of the mass and the elasticity of the spring determine the vibration frequency; during the vibrations, the time average of the kinetic energy is equal to that of the potential energy. Both forms of energy are proportional to the square of the amplitude. The main property of classical vibrations is that their amplitude can be arbitrary. The amplitude is not determined by the equations of motion but depends on the initial conditions, which in classical theory may be specified arbitrarily. In particular, one can have (i.e., there is a corresponding solution of the equations) the case of a mass at rest in a state of equilibrium, i.e., the case of vanishing vibration energy.

Further, making the analogy between the mass on a spring and an electromagnetic wave more precise, we can anticipate a basic property of the classical equations of the electromagnetic field—the possibility of a

solution in which the electric and magnetic fields are everywhere exactly zero. Accordingly, the field energy density is also zero. Of course, we are here dealing with a space in which there are no electric charges. Thus, in classical (but not quantum!) theory the vacuum concept is indeed rather simple—there is neither field nor energy.

Now quantum mechanics appears on the scene. The momentum (i.e., the velocity multiplied by the mass) and the coordinate of the mass cannot have definite values simultaneously. The well-known uncertainty principle of Heisenberg holds. Applied to the electromagnetic field, this means that the magnetic and electric field cannot vanish simultaneously.

But quantum mechanics does more than simply violate the picture of deterministic motion. It is more than a negative theory, and it possesses positive content and predictive power. It predicts that the possible values of the total energy of an oscillator are $E_n=(n+\frac{1}{2})h\nu$ with arbitrary integral n, where h is Planck's constant and ν is the oscillator frequency. Thus, one can have only states with an energy value in the sequence

$$n=0,\quad E_0=\frac{1}{2}h\nu,\quad n=1,\quad E_1=\frac{3}{2}h\nu,\quad n=2,\quad E_2=\frac{5}{2}h\nu,\ldots.$$

If an oscillator can exchange energy with other objects, then it gives up or receives energy only in definite portions, which are multiples of $h\nu$.

In a transition $n=1\to n=0$, the oscillator gives up $h\nu$; in a transition $n=0\to n=2$, it acquires $2h\nu$, etc.

In the initial development of quantum theory, it was this fact that was regarded as fundamental.

But here we wish to draw attention to the mysterious "halves," i.e., the value $\frac{1}{2}h\nu$ of the oscillator ground-state energy.

It is not a mistake: Experiments with atoms and molecules confirm the presence of the "halves." Even at the lowest energy, having given up all the energy that it can, our mass continues to vibrate with definite energy and amplitude. It is impossible to use quantum mechanics and avoid this result. It is impossible to imagine a mass at rest in a state of equilibrium; for this would signify exactly zero velocity in a quite definite position of strict equilibrium—a monstrous violation of the uncertainty principle and in contradiction with modern theory.

By analogy, one could readily believe that the application of quantum theory to the electromagnetic field will necessarily lead to a similar result. Indeed, the electric field and the magnetic field cannot vanish simultaneously; the electromagnetic energy density cannot vanish. One can pose the question of the minimum of the energy in the same way that one can speak about the lowest (ground) state of an oscillator. It is clear however that this minimum is not zero.

To get any further, we must now make more precise what we mean by the modes of the electromagnetic waves and consider what are the quantities that occur in the expressions relating to electromagnetic waves, which are not, in fact, masses on springs. It is important that the appropriate variables, i.e., the analogs of the position and velocity of the mass, are not the magnetic and electric field at one point of space; for Maxwell's equations contain derivatives with respect to the spatial coordinates, and the evolution of fields at a given point depends on the values of the fields at other points of space.

This circumstance makes it necessary to consider individual waves, which are independent of each other.

However, before we proceed, we must discuss a mathematical question—the concept of eigenfunctions and eigenvalues. But who can draw the boundary between mathematics and physics? The question to which we now turn has immense importance for physics. Professional theoreticians should omit the section that follows as trivial.

3. EIGENFUNCTIONS

The problem to which this section is devoted arose long before quantum theory, general relativity, and even the theory of electromagnetism.

We begin by considering a mechanical system, for example, a string. The characteristic feature of a stretched string is the interaction of neighboring sections of the string. If we pull or strike the string at one point, we initiate a motion that in the course of time encompasses other sections of the string that were previously in equilibrium and not subject to an external influence.

Therefore, it is not easy to solve a problem such as, for example, the thermal ("Brownian") motion of a string; for if we ask for the probability of a definite deviation from equilibrium of a given particle of the string at a given temperature, we get the response: Do we mean the deviation of the given particle for known or for arbitrary positions of the other particles at the present time or in the past?

The particles interact, and this is the reason for the complexity of the situation. The problem is solved constructively by finding the simplest noninteracting types of vibration. Concretely, for a string such vibrations have the form

$$y_n=a_n\cos(\omega_n t+\varphi_n)\sin\frac{\pi n x}{l},\qquad (2.1)$$

where y is the deviation of the string from the equilibrium position, x is the coordinate along the string, which is fixed at the points $x=0$ and $x=l$, so that $y=0$ at these points. The quantity y_n is the amplitude of the given nth vibration, i.e., the maximal (in time and space) deviation from the equilibrium position ($y=|a_n|$ where $|\sin(\pi n x/l)=1$, i.e., at the points $x=l(k+\frac{1}{2})/\pi n$, $k=1,2,\ldots,n-1$ when $|\cos(\omega_n t+\varphi_n)|=1$); the quantity φ_n is the phase of the vibration and ω_n is the frequency. The equation of motion of the string does not fix the amplitude or phase. However, for the frequency ω_n the equation gives a definite value: $\omega_n=\pi n a/l$, where a is the wave propagation velocity of perturbations, $a^2=q/\mu$. Here, q is the tension in the string, i.e., the force [g·cm·sec^{-2}], and μ is the density, i.e., the mass per

unit length [g/cm]; therefore $[q/\mu]=[\mathrm{cm}^2/\mathrm{sec}^2]$, as it must.

How are these results obtained? We seek solutions of the form $y(x,t)=A(t)\varphi(x)$. Note that a solution of this form—with separation of the variables—is possible only for a definite set of functions $\varphi(x)$. This set can be characterized by $\varphi_n(x)=\sin(\pi nx/l)$, each number n corresponding to a function with different wavelength and different number (equal to $n-1$) of nodes between the fixed ends.

It is now time to write down the equation of motion of the string:

$$\frac{\partial^2 y}{\partial t^2}=a^2\frac{\partial^2 y}{\partial x^2}. \tag{2.2}$$

After substitution of the solution with the separated variables, we obtain

$$\frac{d^2\psi_n}{dt^2}=-\omega_n^2\psi_n. \tag{2.3}$$

Each nth vibration mode behaves as an oscillator, as a pendulum with definite frequency; see the expression given above.

It should be noted here that when the density (the mass per unit length) of the string is variable, the spatial functions have a more complicated form. Even more complicated is the situation when one considers the vibrations of a plate or a three-dimensional body (for example, a continuous elastic sphere or bell). However, despite the more complicated form of the function of the spatial coordinates x, y, z, the time dependence of each vibration mode remains harmonic, i.e., it is described by the differential equation (2.3) described above. Thus, a continuous body, like electromagnetic radiation in a definite volume, is equivalent to a system of oscillators.[1)]

Why is the possibility of describing the solution in the form of a system of independent equations for individual oscillators so important? One answer, seen immediately in the 19th century, is that if a set of particular solutions is known it is possible to construct a solution to the problem with arbitrary initial conditions. For we are concerned with a linear equation, and any sum of particular solutions is also a solution.

Different initial conditions give a different set of quantities a_n and φ_n in the general expression

$$y=\sum{}' a_n\cos(\omega_n t+\varphi_n)\sin\frac{\pi nx}{l}.$$

There is however a deeper reason for using solutions of this type.

[1)]A string with constant density has the special property that all the frequencies are multiples of a single, lowest frequency: $\omega_n=n\omega_1$. Therefore, the motion of the string as a whole for arbitrary initial conditions is strictly periodic. After a time $T_1=2\pi/\omega_1$ the initial conditions are exactly reproduced. In an arbitrary body, the ratios of the periods are transcendental numbers and the motion as a whole is not periodic. However the individual vibration modes found above are not only periodic but also harmonic. For what follows, this is all that is important. We recall that we consider a system without friction and in the linear approximation, which is important for the possibility of considering noninteracting modes.

The point is that these solutions can be numbered and ordered. They can be arranged in a sequence with increasing value of the frequency. One can find the number of solutions with frequency less than a definite value or in a given interval of frequencies.

In particular, for electromagnetic radiation in volume V the number of such solutions is

$$dN=V\frac{8\pi\nu^2\,d\nu}{c^3}.$$

It is here understood that the frequency ν is such that the corresponding wavelength $\lambda=c/\nu$ is less than the linear dimension of the container $d\sim V^{1/3}$, and we consider an interval $d\nu$ that is not too narrow, so that

$$dN=8\pi\left(\frac{V}{\lambda^3}\right)\frac{d\nu}{\nu}\gg 1$$

(despite $d\nu/\nu\ll 1$).

Accordingly, the total number of solutions with frequency less than the given ν (per unit volume) is

$$n=\frac{8\pi}{3}\left(\frac{\nu}{c}\right)^3=\frac{8\pi}{3\lambda^3}.$$

For a string, bell, and so forth there is a physical restriction, namely, the minimal wavelength of the vibrations cannot be less than the distance between the atoms. But in vacuum there is no definite minimal wavelength! Accelerator experiments study photons with an energy of about 10^{10} eV, and their wavelength is $\lambda\approx 10^{-14}$ cm.

In cosmic rays, we observe photons of even higher energy and shorter wavelength. But more important is the argument of relativistic invariance: There is not and cannot be a limit to the photon energy or wavelength because these are quantities that depend on the motion of the observer. For an oncoming observer, the energy will be higher, the wavelength shorter.

The vacuum has an infinite number of vibration modes, or, more precisely, an infinite number of vibrations per unit volume of the vacuum. Theory must take into account this fact and must be able to overcome the difficulties—computational and conceptual, i.e., "physical" associated with this fact.

4. VACUUM ENERGY DENSITY

We now turn to the above assertion ($E_0=\frac{1}{2}h\nu$), which follows from quantum theory.

In granting a modest $0.5h\nu$ to each individual wave, we soon discover with horror that when all the waves are taken together they give an infinite energy density. If we were to restrict ourselves to a definite maximal frequency ν_m, we would obtain a result of the form

$$\varepsilon=a\int_0^{\nu_m}\frac{1}{2}h\nu\cdot\nu^2\,d\nu=(ah/8)\,\nu_m^4,$$

where ε is the energy density and a is a constant ($a=kc^{-3}$, where c is the velocity of light and k is a number of order unity). In the limit $\nu_m\to\infty$, the value of ε also tends to infinity. If we set $\nu_m=\infty$ directly, we obtain a divergent integral.

This is the well-known divergence problem, the so-called "ultraviolet catastrophe" of quantum electrody-

namics or, rather, it is part of this problem.[2)] And there is no simple escape; one cannot ignore or simply reject the problem. The nonvanishing fields in the absence of photons (the fields corresponding to the "halves" $\frac{1}{2}h\nu$ for all possible ν) are observed, and they modify the motion of electrons in atoms. The famous Lamb–Retherford experiment confirms this. The point is that the most numerous short-wavelength, high-frequency "halves" have a comparatively weak influence on electrons, which move only slightly under the influence of a rapidly varying force for a short period. The theory with divergent vacuum energy gives convergent finite results for the motion of electrons, and gives corrections to the theory of atomic spectra which are confirmed experimentally. The solution to the mystery of the divergent energy cannot consist of the simple negation of the zero-point vibrations. Such negation would lead to a contradiction with modern exact experiments.

A careful examination of the part played by the zero-point vibrations (the "halves") in laboratory physics shows that it is less important than might appear at the first glance.

The electromagnetic vacuum energy is infinite, but if one considers a particular process, for example, the emission of a photon by an atom, $A^*=A+\gamma$, the infinity cancels. If the process takes place in a volume V and the vacuum energy density is denoted by ε_V, the conservation equation for the (total) energy E_{tot} has the form

$$E_{tot}=\text{const}=\varepsilon_V V+E_{A^*}=\varepsilon_V V+E_A+E_\gamma.$$

Irrespective of the value of ε_V—zero, finite, or infinite—we obtain the energy conservation law in the usual form $E_A{}^*=E_A+E_\gamma$.

For the same reason, it is obvious that the vacuum energy cannot be used in practice to turn electric motors or for illumination. A tremendous achievement at the end of the forties and beginning of the fifties was the development of a consistent method for calculating the influence of the "halves" on the motion of an electron in an atom and on the magnetic moment of the electron. I am referring to the theory of renormalization in quantum electrodynamics, for which Feynman, Schwinger, and Tomonaga received the Nobel Prize in 1965. The theory is in marvelous agreement with experiment. The corrections are about 10^{-3} or less of the main effect and the accuracy of the experiment is about 10^{-10}, i.e., 10^{-7} of the corrections.

Besides the Nobel lectures of the creators of renormalization theory,[2-4] one can recommend the lucid and intuitive paper of Weisskopf.[5] The latest measurements, which completely confirm the theory, are described, for example, in Ref. 6. We note also a phenomenon associated with the idea of the zero-point energy—the Casimir effect.[7] Suppose that metallic, i.e., electrically conducting, bodies or dielectrics are placed in vacuum. Their presence has a definite influence on the spectrum of electromagnetic oscillations and, therefore, on the zero-point energy. In this case, we are speaking of the zero-point energy of space containing bodies.

The higher the photon energy, the less influence the presence of bodies in space has on its propagation. Therefore, the change in the zero-point energy of the electromagnetic oscillations associated with the presence of the bodies (the metals or dielectrics) diverges less strongly than the zero-point energy itself.

But Casimir calculated a more subtle effect, namely, he found the dependence of the zero-point energy on the mutual position of the bodies, for example, on the distance between the plates of an uncharged capacitor. But the derivative of the energy with respect to the displacement is the force acting in the direction of the displacement. This quantity is finite and the corresponding integrals converge.

Physically, it is clear that for short wavelengths (many times shorter than the distance between the plates in the case of a capacitor) the position of the bodies is unimportant, the short waves making no contribution to the integral that determines the force.

Thus, the actually observed force, capable of measurement by a balance, actually depends on the zero-point energy of the electromagnetic oscillations in the vacuum. And the infinite value of this energy does not appear in the calculation but cancels, and the theory gives a result in agreement with experiment.

However, such a favorable situation does not occur in all phenomena, and the density of the zero-point energy does not always cancel. It should not escape notice that we made a restriction to laboratory physics earlier.

The most important manifestation of the nonzero vacuum energy density could be its influence on the gravitational force field and on the gravitational potential.

The theory of gravitation contains the energy density of a body, including the energy density of the vacuum within the body and the surrounding space.

In this case, we are not speaking of energy differences, which could be zero. At first glance, we face an ineluctable contradiction. In principle, the contradiction could perhaps be avoided by taking into account the contribution of other particles. We shall merely em-

---

[2)]Divergent integrals appear not only in the calculation of the vacuum energy but also in other problems, for example, the calculation of the corrections to the masses of elementary particles, which depend on the interaction of these particles with the electromagnetic and other fields. We shall not dwell on these questions here, despite their great importance. The development of theoretical physics during the last decade has been subject to the condition of creation of a renormalizable theory, i.e., such that it does not give inifinite answers. This principle played a decisive part in the formation of the theory which combines the weak and electromagnetic interactions. This is discussed by one of the creators of the theory, Weinberg, in his Nobel address[1] (to which we shall return later). We hope that the principle that the theory must agree with the experimental data on the total vacuum energy density will also play its part in the development of a general theory encompassing all fields.

phasize here (and return to this point later) that this fundamental possibility has not yet been realized by modern science quantitatively and exactly!

Let us return to the theory of electrons. Modern pupils (at least, those with interest) know that the energy of an electron and its momentum are related by the relativistic equation $E^2 = c^2p^2 + c^4m^2$, so that $E = \pm\sqrt{c^2p^2 + c^4m^2}$; the two signs in front of the radical must be taken seriously.

In classical theory, we are reassured by the circumstance that the momentum and energy of an electron can vary only smoothly. If in the initial state all electrons have energy $E = +\sqrt{\ }$ (we omit the radicand), then states with $E = -\sqrt{\ }$ can be simply forgotten. However, in quantum theory states with $E = -\sqrt{\ }$ cannot be eliminated.

The quantum laws of motion do not preclude the transition of an electron downward, for example, from $E = +m_ec^2$ to $E = -m_ec^2$, with the emission of two or three photons.

Some years ago Dirac celebrated his 70th birthday. Simultaneously, physicists commemorated the 50th anniversary of a remarkable idea of Dirac: Downward jumps of electrons are forbidden by the Pauli principle!

For this, Dirac introduced the idea of an infinite number of electrons that populate without exception all the states with negative energy ("the Dirac sea"). Vacancies, i.e., unfilled states, in this sea are observed as positive charges (the absence of a negative is something positive). This theory was brilliantly confirmed by the discovery of positrons—all their properties agreed with the predictions for vacancies in the Dirac sea.

An electron with positive energy can fall into an unfilled state, emitting, for example, two photons. Annihilation of an electron and positron is described in this manner.

In reality, Dirac's theory is symmetric: One could regard positorns as "elementary," introduce the concept of a sea of positrons with negative energy, and regard the electron as a vacancy in this sea.

In fact, in the modern exposition one introduces operators of creation and annihilation of electrons and positrons and formulates the theory directly in a symmetric form without recourse to the perspicuous concept of an occupied sea of states with negative energy.[3)]

This modern symmetric theory eliminates the problem of the density of the electric charge of the vacuum. In the alternative theories, the charge is $+\infty$ or $-\infty$. In the symmetric theory, the charge density is zero by virtue of the symmetry.

For the following exposition, another feature of Dirac's theory is relevant to the question in which we are interested.

Since we consider occupied states with negative energy ("sea"), a negative total energy density is also naturally obtained. This property of the theory also remains fully in the modern symmetric formulation.

Thus, there is a possibility of compensating the positive contribution to the vacuum energy density associated with the zero-point energy of the photons by the negative contribution of the electrons.

More generally, the positive contribution of bosons (particles with integral spin) could in principle be compensated by the negative contribution of the fermions (particles with half-integral spin).

Promising here is the fact that the principal, divergent terms in the integrals have the same order of divergence for the positive and negative integrals, i.e., for the bosons and fermions.

However, this is in no way a justification for complacency. The masses of the different particles are not equal. It is also necessary to take into account the interaction of the various species of particle when the vacuum energy is considered—this is no longer a surprise.

Therefore, even after canceling of the infinities, i.e., the divergent parts of the integral, it would be perfectly possible to obtain a finite and nonzero value.

Astronomical observations (see below) tell us that

$$|\rho_{vac}| < 10^{-29}\ \text{g/cm}^3,\quad |\varepsilon_{vac}| < 10^{-8}\ \text{erg/cm}^3.$$

The canceling between the different fields works miraculously, since a simple order of magnitude estimate based on dimensional considerations would give

$$\rho_{vac} = m\left(\frac{mc}{\hbar}\right)^3.$$

If m is the proton mass, we obtain $m = 1.6\cdot10^{-24}$ g, $\hbar/mc = 10^{-13}$ cm, and $\rho_{vac} = 2\cdot10^{14}$ g/cm^3.

Thus, the vacuum energy density does not exceed the fraction 10^{-43} of the simple estimate. We have here a remarkable and hitherto unexplained fact.

The question arose about 50 years ago. The general progress in the physics of elementary particles and field theory during this period excites admiration. It is hard to find in history another half-century so full of discoveries.

Nevertheless, the most important theoretical question—that of the vacuum energy density—remains unanswered. Only astronomy gives definite strong restrictions.

To prove that the problem has been recognized but not solved, we give a quotation from Weinberg's Nobel address[4]: "There is nothing impossible in this [*i.e., that the particle masses should not be superlarge; Ya. B. Z.*], but I have not been able to think of any reason why it should happen. The problem may be related to the old mystery of why quantum corrections do not produce an enormous cosmological constant[4)]; in both cases, one is

[3)]Above, we spoke of symmetry in a different sense, namely, the possibility of choosing between two alternative theories each of which is separately asymmetric (electron sea, or positron sea).

[4)]In gravitational theory, i.e., in the general theory of relativity and in astronomy, a nonzero vacuum energy density is usually called the "cosmological constant" or the "cosmological term in the equations." For more details, see below.

concerned with an anomalously small "super-renormalizable" term in the effective Lagrangian which has to be adjusted to zero. In the case of the cosmological constant, the adjustment must be precise to some fifty decimal places."

We mention in this connection the paper of Hawking with the intriguing title "Space-time foam."[8]

Developing ideas put forward long ago by Wheeler in a qualitative form, Hawking considers the smallest scales, at which it is necessary to take into account strong fluctuations of the metric, which do not reduce to zero-point oscillations of gravitational waves (see the beginning of Sec. 5 for more details on the metric of space). Hawking shows that one can also have fluctuations that change the topology of space-time. A perspicuous two-dimensional example of such fluctuations is the successive transition from a flat film of soap solution to a curved film (without change in topology) to a foam in which particles originally separated by a finite distance touch and previously neighboring particles are torn apart (new topology!). Hence the title of the paper. In a discussion with the present author in July 1980, Hawking said that when allowance is made for the foam structure the effective energy density of the vacuum, averaged over a large scale, may vanish. In the paper of 1978 and subsequent preprints, this idea is not expressed and there are no quantitative estimates of the residual vacuum energy. The appealing idea of self-regulation, leading to $\varepsilon_{\text{vac}}=0$, has not yet been realized.

How would a finite energy density be manifested? In relativity theory, it is necessary that this energy density be the same for any observer. This leads to the condition that the pressure (tension) is the same in all directions and equal to $p=-\varepsilon$, where p is the pressure and ε the energy density of the vacuum. As early as 1917, Einstein considered the possibility that the vacuum energy density could be nonzero. He used a different terminology and introduced the "cosmological constant" Λ, which is proportional to ε. This name emphasized that such an energy density would have its strongest influence on cosmological phenomena.

In a paper entitled "The cosmological constant and the theory of elementary particles" published in this journal, I gave a detailed review of the state of the problem as posed decades ago.[9] Recently, the $\varepsilon_{\text{vac}}\neq 0$ question has again arisen (Zel'dovich, Syunyaev[10]) in connection with reports of a neutrino mass.

The period between Ref. 9 and Ref. 10 has seen not only remarkable discoveries and the construction of deep theories. Under the influence of real advances, a general psychological shift has taken place in the family of physicists.

The criterion of the simplicity of nature has been replaced by the criterion of unity and symmetry of nature. A physics in which there existed only protons, neutrons, electrons, massless photons, and massless neutrinos was maximally simple; as Mayakovskii said "simple as mooing."

It would seem that if there are no direct indications that the neutrino has a nonvanishing mass it is natural to assume that $m_\nu\equiv 0$—such a theory is simpler. Today, experimental indications that $m_\nu\neq 0$ appear at the same time as theoretical papers with various predictions of a possible neutrino mass.

We shall not here discuss these particular papers. Let us merely point out a common tendency: If (or as long as) it has not been proved that the neutrino mass vanishes, at present it is natural to assume that it does not vanish. The stability of the proton has been proved experimentally only in definite limits—it is natural to assume that the proton can nevertheless decay with a small probability that does not contradict the experiments.

We can approach the question of the vacuum energy density in a similar spirit. We cannot now rule out the possibility that theory and observations will give some very small but nonvanishing value of ε_{vac} and Λ (see above).

The psychological considerations that we have just given can be bolstered by more technical arguments.

Physicists assumed that sensible theories do not give dimensionless numbers differing too strongly from unity. The only aesthetically acceptable exception was zero. Given the order of magnitude estimate of ε_{vac} constructed above ($m^4c^5/\hbar^3$) by means of the proton mass, it would seen that $\varepsilon_{\text{vac}}\equiv 0$ is the only acceptable solution.

But in reality the unification of all the forces of nature, including gravitation, necessarily leads to the appearance of numbers very different from unity. The first example of such a number is $Gm_p^2/\hbar c=10^{-37}$, where G is the Newtonian gravitational constant. In modern theories unifying the strong and weak interactions there is an X boson, which is 10^{15} times more massive than the proton. Having such numbers at our disposal, we can readily construct a formula for ε_{vac} that is not zero, has the correct dimensions, and does not contradict experiments. For example,

$$\varepsilon_{\text{vac}}=\frac{Gm_p^4c^4}{\hbar^4}=10^{-40}\ \text{g/cm}^3.$$

Theories have recently appeared that establish a definite symmetry between bosons and fermions.

I am referring to the theory of supergravity and supersymmetry. In such a theory, the number of species of bosons and fermions may be such that the cosmological constant is automatically zero, which is a definite merit of the corresponding variant of the theory. However, the observed masses of the particles are very different. There is no doubt that in nature the symmetry holds only asymptotically, at high energies. Therefore, the conclusion of supersymmetric theories is correctly formulated as follows: The most dangerous term of the type $\int k^3dk$, which diverges as k^4, may vanish automatically. However, this does not preclude terms that diverge as $m^2\int kdk$ or finite terms of the type m^4. In addition, in the vacuum one can have field fluctuations of the type of below-barrier transitions, which cannot be described by the theory of small per-

turbations; these are the so-called instantons.

This circumstance is also ignored in supersymmetric theories when they give $\varepsilon_{vac}=0$.

Thus, despite the great value of the concept of supersymmetry, it must be stated that there remains a mystery concerning the small value of ε_{vac}, i.e., Λ.

A final comment: If ε_{vac} is obtained by the almost complete cancellation of large positive and negative quantities, it may have either positive or negative value. This is a difference from ordinary matter and ordinary fields (excitations of the vacuum), which always make only a positive contribution.

5. EMPTY CURVED SPACE

The reader will be familiar with the general theory of relativity—if not the mathematical details, at least the ideas. Lobachevskiĭ in Russia and Bolyai in Hungary were the first to point out that space need not necessarily satisfy the laws of Euclidean geometry (Gauss developed similar ideas, but was afraid of publishing them).

The next idea, developed by Riemann, was that of geometry which varies from place to place. It was then necessary to consider what causes the geometry to change and what influence such geometry has on the motion of bodies, the propagation of light, and so forth. However, in Riemann's time physics was not yet ready to answer these questions.

Before the program was realized, the electromagnetic field had been studied, atomic theory developed, and the special theory of relativity, which links space and time, had been introduced. It was only after the development of the relativistic theory of the electromagnetic field that the prerequisites for the creation of a relativistic theory of gravitation were created.

At the end of this development, in 1915, Einstein created the general theory of relativity. This theory is based on the following assumptions: 1) the density and pressure of matter make space-time curved and 2) motion in the curved space describes motion under the influence of gravitational forces.

The general theory of relativity is a theory of gravitation, or rather, using the emphasis achieved by the definite article in English, it is *the* theory of gravitation, logically closed, and in outstanding agreement with all experiments. Instead of the curvature of space discussed in the 19th century (until the present century, time was regarded as absolute and dependent on nothing), we now speak of the curvature of the space-time complex. The very rate of time—the ticking of clocks, the vibrations of atoms, or the aging of man—depends on the gravitational potential. In the simplest case, the difference in the flow of time is the measure of the gravitational potential. The gravitational red shift, i.e., the loss of energy by a photon that leaves the sun and overcomes a gravitational barrier, depends on the potential. We shall here use the concepts of quantum theory. A photon uses a part of its energy E_γ to leave the sun, while the photon energy is related to its frequency by $E_\gamma=h\nu$. If E_γ on the earth is smaller than E_γ on the sun at the start of the journey, then ν_{obs} (the frequency observed on the earth) is less than ν_{em} (the frequency emitted on the sun). But the wavelength is $\lambda=c/\nu$, the wavelength increases, the spectral lines are shifted from the blue to the red, and hence the name "red shift."

The frequency is lowered, and this means that an observer on the earth can with full justice say that all the processes on the sun-take place somewhat more slowly, and time flows differently. This explains why one does not simply speak of "spatial curvature" but rather "space-time curvature."

Space-time is curved in regions occupied by matter, but it is also curved in the surrounding regions—gravitation is a long-range force, and the elasticity of space-time causes a rearrangement of it outside bodies as well.

We now return to the theme of our paper, the behavior of the vacuum.

We should now repeat all our previous discussions about the energy of the zero-point oscillations of the particles and waves, the Dirac sea, and so forth. But the waves (for example, electromagnetic) and wave functions of the particles (electrons, for example) are now distorted.

We must repeat all the calculations with distorted waves. From the start, it is clear that at short distances, for the shortest wavelengths, the distortion produced by the curvature is small. This means that the most disagreeable infinities associated with the infinite number of short wavelengths cancel for each separate field if we are interested only in the difference between the cases of curved and flat space-time.[5)]

On the other hand, this difference is of the same sign for fermions and bosons. Therefore, canceling of the energy density in flat space does not indicate canceling of the difference for curved and flat spaces. Our view is that this difference is nonzero and observable. Let us analyze the difference.

One part of the effect is completely masked (see the paper Ref. 11 of Ginzburg, Kirzhnits, and Lyubushin). This is the part of the energy density and the pressure which arises in the zero-point oscillations and in the Dirac sea and is strictly proportional to the energy density and pressure of the ordinary matter that curves space. In empty space, this part is zero. It is clear that our general notion of gravitation is not changed by this part of the vacuum polarization; Newton's law for weak fields and the equations of general relativity remain valid. To get a clear idea, we shall suppose that for each gram of matter there is, say 0.01 g mass equivalent of energy density due to these effects, i.e., vacuum polarization.

We substitute this contribution in the gravitational equations and obtain $(Gm+0.01Gm)/r^2$ for the Newtoni-

[5)]Thus, we consider a situation similar to the Casimir effect; see above.

an acceleration. But this result can be interpreted as a change in the gravitational constant:

$$Gm + 0.01Gm = G\cdot 1.01m = G'm.$$

It is G' that we observe and measure, so that for the whole of macroscopic physics the "old" unobservable value of G is unimportant, and, these problems being interrelated, we do not face the problem of establishing what is the real contribution of the vacuum, which we took above arbitrarily, for illustration, to be 0.01.

A similar procedure was used for the first time in the fifties in quantum electrodynamics in connection with the electric charge: A free charge produces a vacuum polarization charge, perturbing the motion of charged particles (electrons, etc.) in the Dirac sea of states with negative energy, which are everywhere present in the vacuum. With each electric charge there is associated a transformation $e \to e'$ like the transformation $G \to G'$ for the gravitational constant. This procedure is called "charge renormalization." It can be carried out even if the ratio e/e' is infinite. The theoreticians developed schemes for calculating all observable effects using only the observed value e'. This was a great success in the fifties! However, it is necessary to emphasize a difference between electrodynamics and the theory of gravitation. In electrodynamics, one can study the interaction of two elementary particles at a very short distance. It is possible to probe the distribution of the cloud of vacuum charge responsible for the difference between e and e'. This cloud changes the energy of the atomic levels by a measurable amount. For the hydrogen atom, the difference is about $2\cdot 10^{-6}$ of the electron binding energy in the atom, but modern methods have made it possible to measure this quantity.

We can study the gravitational interaction only at the macroscopic level, and therefore the experimental investigation of the gravitational vacuum polarization is at present outside the scope of the possible. We must content ourselves with an analysis of the theoretical conclusions.

Thus, the vacuum energy in flat space is very small; we know this from astrophysical data. The part of the vacuum energy in curved space proportional to the energy of the matter producing the curvature is manifested in a change in the previously unknown gravitational constant, and in this sense is unobservable.

The notion of a possible change in the gravitational constant due to vacuum polarization does not change the form of the equations but changes their meaning. Sakharov[12] has conjectured that the gravitational constant is entirely determined by vacuum polarization.

The equations of gravitation can be perspicuously interpreted as a manifestation of the elasticity of space (we recall that whenever, for brevity, we speak of "space," we mean the four-dimensional "space-time" complex).

The first half of general relativity consists of considering the motion of particles in curved space-time. Curvature influences the motion of the particles. Developing these ideas mathematically, we find the equations of motion of the particles, the equations of motion of a fluid, and Maxwell's equations.

In accordance with Newton's third law—the law of the equality of action and reaction—it is natural to expect a reaction of the particles and fields on space. When the rails act on a railway car, bending its trajectory, the car acts on the rails with a definite force.

The second part of general relativity is analogous to considering the behavior of the rails: Besides the forces exerted by the car, it is necessary to take into account the elasticity of the rails and their connection to the rail ties and the embankment. It can be said that Einstein's equations describe the elasticity of space. The possibility mooted by Sakharov[12] is that this elasticity could be entirely determined by vacuum polarization effects, i.e., be similar to the Casimir effect.[6)] We write Einstein's equations in the form

$$T_i^k = \frac{c^4}{8\pi G}\left(R_i^k - \frac{1}{2}\delta_i^k R\right).$$

The influence of matter on the metric is determined by the energy–momentum tensor T_i^k (in particular, the component T_0^0 is simply the energy density, with dimensions erg/cm^3).

On the right-hand side in the brackets we have the curvature, which has dimensions cm^{-2}. The coefficient $c^4/8\pi G$ is the elasticity of the vacuum.

This coefficient is large in the cgs system of units, $c^4/8\pi G = 5\cdot 10^{45}$ g·cm·sec^{-4}. In itself, this means nothing. However, it is important that if we express the elasticity in terms of quantities that characterize the elementary particles—Planck's constant $\hbar$, the velocity of light c, and the proton mass—we obtain a quantity many times smaller than the one given above. We obtain the correct dimensions by taking the expression $m^2c^3/\hbar = 10^{11}$ (in the cgs system), which is almost 10^{35} times smaller than we need. It would seem that the hypothesis is unrealistic, though at the present time elementary particles with a mass 10^{15} times greater than the proton's are seriously considered. The hypothesis that reduces the elasticity of the vacuum, and thus the theory of gravitation itself, to vacuum polarization is again attracting the interest of theoreticians. In support of this statement, we may mention the lecture of the well-known American physicist Adler at the P.N. Lebedev Physics Institute in October 1980.

However, independently of the *explanation* of the equations of general relativity, there are questions relating to their *change* in a situation when the components of the curvature tensor R^i_{lkm} are large. We recall that the curvature is completely characterized by this four-index tensor. The quantities R_{ik} and R are definite sums of the components R^i_{lkm}. The vanishing or a small value of R_{ik} and R does not yet mean that the terms R^i_{lkm} themselves are small.

[6)] *Translator's Footnote.* A volume of Sakharov's scientific papers, including some previously unpublished material on this subject, is currently being prepared by the publishing house Marcel Dekker.

The point is that there are other parts of the vacuum polarization which depend on the space-time curvature differently; they are not proportional to the combination $R_{ik}-\frac{1}{2}g_{ik}R$, which, in its turn is in accordance with general relativity proportional to the density and pressure of the matter.[7)] In what follows, we shall refer to these terms as the *true vacuum polarization*. A characteristic feature of the true vacuum polarization is that it remains nonzero even in complete vacuum, i.e., in a space devoid of ordinary matter. The pioneering work in this direction was done in the United States by DeWitt and Parker. The properties of the vacuum energy density and vacuum pressure produced by the true vacuum polarization are radically different from the energy and pressure of ordinary matter.

To the best of my knowledge, it was in the joint paper by Pitaevskii and myself,[14] published in 1971, that this feature of the true vacuum polarization was clearly pointed out.

The point is that in the true vacuum polarization *the energy density need not be greater than the pressure.*

In the language of the specialists this is called a "violation of the dominant energy condition for the true vacuum polarization." In the following section, we shall discuss the important consequences of this property.

6. THE DOMINANT ENERGY CONDITION AND SINGULARITIES

We begin with simple examples. The pressure of an ideal gas is $p_g=nkT$, where n is the number of particles in unit volume, k is Boltzmann's constant, T is the temperature, and the subscript g denotes gas. The mean kinetic energy of atoms is $3kT/2$, and therefore the density of the kinetic (thermal) energy is $e_{th,g}=3nkT/2>p_g$.

But in our treatment involving gravitation, we must also take into account the rest mass of each atom and its energy equivalent mc^2. Therefore, the total energy density is

$$\varepsilon_{tot,g}=n\left(mc^2+\frac{3}{2}kT\right).$$

In a nonrelativistic gas $kT\ll mc^2$ and $\varepsilon_{tot}\gg p$, i.e., ε_{tot} is appreciably greater than p. For hydrogen at 1 000 000 °K, the pressure is $2\cdot10^{-7}$ of the energy density (including the rest mass).

Thermal radiation gives a further example with a much larger pressure/energy ratio: It is well known that (the subscript r stands for radiation) $\varepsilon_r=7.5\cdot10^{-15}T^4$ (T is measured in degrees Kelvin, and ε in erg/cm) and $p=\varepsilon/3$, but p is still less than ε. The limiting case is that of radiation propagating in both directions along only one line (instead of random three-dimensional thermal radiation). In this case, the pressure p_{xx} on the surface perpendicular to the x axis is equal to ε, but even in this case it does not exceed ε. Energy dominance still holds!

One will evidently expect this to remain true in the most exotic situations in superdense nuclear matter, in neutron stars, and so forth.

Why is so much attention devoted to the dominant energy condition?

The British school—Hawking, Penrose, Ellis, and many other brilliant scientists—have shown that a singularity, i.e., an infinity, is unavoidable in the classical general theory of relativity. This must mean for cosmology that Einstein's classical theory is invalid for the description of the actual beginning of the expansion of the Universe, i.e., for the hot big bang characteristic of modern theory. The singularity does not make the entire theory absurd, but it is a pale beyond which we cannot advance.

In a very crude approximation, the singularity can be regarded as an infinite matter density, which is associated with infinite curvature in general relativity. Such a situation can be encountered either in the case of unlimited contraction, for example, in collapse, or in expansion from an initial state of infinite density. The equations of general relativity are symmetric with respect to reversal of the time, so that mathematically the problems of collapse and expansion are similar. The singularity in the case of the collapse of a star is not particularly dangerous: It is hidden by a black hole. If gravitational collapse occurs in a star, the last electromagnetic waves and neutrinos are emitted by the matter long before the singularity is formed. Being more precise, one should say that the collapse is dangerous for someone who falls into the black hole but harmless for a distant "external" observer.

But in cosmology the singularity becomes a problem; for it does not occur at the end but at the very beginning of the evolution of the Universe. An infinite density at the beginning of evolution is the common lot of all matter which fills the Universe at present. One sometimes says that the Universe is a giant black (or rather white) hole through which every existing thing has passed.

However, one of the conditions, or axioms, on which the theorems proving the inescapability of singularities is based, is the dominant energy condition $\varepsilon>p$, or, in more general form, $\varepsilon\geqslant T_\alpha^\beta$, where α and β are spatial indices. The entity T_α^β is a generalization of the concept of pressure to the case when the stresses are nonisotropic: T_α^β is the component of the force directed along axis α on unit surface s^β such that the normal to the surface is directed along axis β. The concept of T_α^β includes the existence of shear stresses as well as pressure. In the isotropic case, when Pascal's law holds, $T_\alpha^\beta=\delta_\alpha^\beta p$ is the definition of the pressure.

In principle, it is readily seen that if the matter filling space does not satisfy any definite conditions the singularity theorems are impossible. Indeed, in the framework of general relativity one can invert the prob-

[7)] Here, the quantities g_{ik}, R_{ik}, R, which depend on the coordinates and the time, describe the metric and curvature of space-time. We refer the reader to the book The Classical Theory of Fields by Landau and Lifshitz,[13] which contains a detailed and pedagogical exposition of the foundations of the general theory of relativity.

lem. We can write down a metric describing contraction of a world with the contraction then replaced smoothly by expansion, for example,

$$ds^2 = dt^2 - a^2(t)\,[dx^2 + dy^2 + dz^2],$$

or

$$ds^2 = dt^2 - a^2(t)\,[dr^2 + \sin^2 r\,(d\theta^2 + \sin^2\theta\, d\varphi^2)],$$

with function $a(t)$ of the form $a(t) = \sqrt{k^2t^2 + r_0^2}$ or $a(t) = r_0 \cosh kt$. The functions $a(t)$ here are chosen such that $a(t) \to \infty$ as $t \to -\infty$ and $a(t) \to \infty$ as $t \to +\infty$, and at $t = 0$ the function $a(t)$ has a definite value of r_0. The equations of general relativity enable one to determine in an elementary manner the energy and the pressure (as a function of the time) of the matter filling the Universe. Moreover, one can verify that during the evolution the energy and pressure satisfy the first law of thermodynamics, $d(a^3\varepsilon) = -p\,d(a^3)$, which corresponds to the well-known $dE = -p\,dV$ when it is borne in mind that the element of volume is proportional to a^3.

Now the general equations for the above metrics can be simplified (see, for example, Ref. 13 or Ref. 25) and have the form

$$\frac{1}{a}\frac{d^2a}{dt^2} = -\frac{4\pi G}{3c^2}(\varepsilon + 3p),$$
$$\left(\frac{1}{a}\frac{da}{dt}\right)^2 = \frac{8\pi G}{3c^2}\varepsilon + \frac{kc^2}{a^2},$$

where $k = 0$ for the first (flat) metric and $k = -1$ for the second metric (for a closed model). There exist however solutions with $k = +1$ as well (hyperbolic model). Here, the difference between the three variants is unimportant. In all cases it is an elementary matter for an arbitrary smooth function $a(t)$ to find ε and p, which remain finite throughout the entire interval $-\infty < t < +\infty$.

We note that the smooth transition from contraction to expansion necessarily requires a point of maximal contraction somewhere in the interval. At this point, $a(t)$ has a minimum: $da/dt = 0$. By itself, this still does not determine ε, but the important thing is that at the minimum of a the second derivative is positive! Hence, the sum $\varepsilon + 3p$ must be negative.

Accordingly, if we assume that the sum is always positive, $\varepsilon + 3p \geq 0$, then a smooth transition from contraction to expansion is impossible, and a singularity $a \to 0$ is unavoidable. This is the simplest form of expression of the dominant energy condition for an isotropic homogeneous Universe.

The condition naturally contains ε and p, the sums of the contribution of the matter and the vacuum, including the true vacuum polarization, i.e., the contribution of the curved space (vacuum). For matter, we naturally have $\varepsilon > 0$ and $p > 0$. If the matter pressure started to decrease on contraction, the matter would be unstable and would decay into two phases. Therefore, hopes for the construction of a realistic solution without singularity are associated with assumptions about the behavior of the true vacuum polarization, which is not related to the dominant energy condition. This does not contradict the circumstance that ultimately the true vacuum polarization arises from the contributions of the various known fields (photon, electron, etc.), each of which satisfies the dominant energy condition.

In calculating the true vacuum polarization, we must subtract some contributions from others, for example, the energy of the sea of negative-energy electrons from the zero-point energy of the photons. The necessity for subtraction can be clearly seen from the fact that the cosmological constant, i.e., the true vacuum polarization of flat space, is small or even zero. For any collection of real photons, the dominant energy condition is satisfied, as for real electrons and positrons.

The violation of the dominant energy condition can be readily seen in the very first example of the true vacuum polarization of flat space-time: It is perfectly possible that the vacuum energy density ε_{vac} is negative, and then $p_{vac} = -\varepsilon_{vac}$ is positive, so that the pressure is greater than the energy density, as we wanted to show. But in the case of flat space-time, we know that the absolute magnitudes of ε_{vac} and p_{vac} are small. The cosmological constant certainly does nof influence the behavior of the solution in the early stage in the evolution of the Universe, when the densities were appreciably higher than the contemporary $\rho = 10^{-29}$–10^{-31} g/cm^3. We are interested in the true vacuum polarization (with emphasis on the "true") in strongly curved space-time, near a singularity, and of a kind which makes it possible to avoid a singularity. There is now no formal discrepancy, no rigorous theorem, and no fundamental objections to prevent our hoping for the existence of a solution free of a singularity. This question is the subject of Sec. 7. Here, we shall only briefly mention one further fundamental effect associated with violation of the dominant energy condition in the true vacuum polarization, namely, the possibility of spontaneous production of ordinary particles by gravitational forces.

Here too there was a theorem of Hawking,[15] which forbade the production of particles by a gravitational field; the theorem uses the dominant energy condition. The essence of the proof is as follows. Gravitational forces release energy in a given volume, doing mechanical work when a volume contracts and the contraction is opposed by pressure, or when a volume expands in opposition to tension; the two effects can also be combined—contraction in one direction and expansion in another.[8)]

But suppose we begin with vacuum—with zero energy density and zero tension and pressure. If the dominant energy condition holds, the pressure and tension remain zero as long as the energy density is zero.

The release of energy is impossible—as impossible as it was for Baron Münchhausen in German folklore to pull himself out of water by his hair....

But if the dominant energy condition is violated then space-time curvature can first create pressure and

[8)]Consider a magnetic field, which has a tension along a line of force and a pressure at right angles to it. One can readily obtain a motion which pumps energy and increases the field by stretching a plasma long the field, compressing it at right angles to the field.

tension, and then the pressure and tension can cause the release of energy. This was the line of argument advanced by Pitaevskiĭ and myself in Ref. 14 (I gave further examples in the book Magic Without Magic in honor of Wheeler's 60th birthday[16]). Suppose the distortion of space-time is small and characterized by the small parameter δ. Then the pressure and tension of the true vacuum polarization is of order δ. The deformation is also proportional to δ. The work is the product of the pressure and/or tension and the deformation, so that it is proportional to δ^2. Destroying the true vacuum polarization, one can make the account balance: A direct calculation of the energy of the produced particles shows that it is proportional to δ^2 in complete agreement with the above simple considerations. Here, in the production of the particles we have directly used the violation of the dominant energy principle, since for small δ the small p is proportional to δ and, therefore, greater than ε, which is proportional to δ^2.

The possibility of producing particles by a gravitational field, i.e., curved space-time, is now beyond doubt. Hawking's beautiful theory describing the production of particles by small black holes—the so-called black hole evaporation—is the best example.

7. COSMOLOGY AND VACUUM POLARIZATION. SOLUTION WITHOUT SINGULARITY

We now draw near the end of our story. All that we have learnt above leads to a possibility of understanding the solutions currently proposed to that most intriguing mystery of the initial stage in the evolution of the Universe.

There is a remarkable cosmological solution—a law determining the structure and rate of expansion of the Universe—with an exceptionally high symmetry. This is the solution proposed long ago by the Dutch scientist de Sitter in 1917.

One of the variants of de Sitter's model is a flat Universe with Euclidean geometry of the three-dimensional space, i.e., the sections $t=\text{const}$ of the four-dimensional space-time. But the scale of this space increases exponentially:

$$ds^2 = c^2\,dt^2 - dl^2, \quad dl^2 = e^{2Ht}(dx^2+dy^2+dz^2)$$
$$= a^2(t)(dx^2+dy^2+dz^2), \; a = e^{Ht}, \; H = \text{const}.$$

Therefore, the space-time manifold is not flat.

De Sitter's solution has a very high symmetry. It corresponds to expansion, but the expansion law gives an equal relative increment of all spatial distances for each equal small interval of time, i.e., $\Delta a/a = H\Delta t$, and H is constant, so that the entire picture does not change with time. The absolute value of a does not have physical meaning for the flat infinite three-dimensional space, so that a change in a does not change the properties of the solution.

Using the formulas of the preceding section, we can readily see that the de Sitter solution holds when $\varepsilon = -p = 3H^2c^2/8\pi G$ on the right-hand side of the equation.

One would like to use de Sitter's solution to describe the beginning of the Universe. This desire is because of the circumstance that if the scale factor $a = e^{Ht}$ is extrapolated to $t=-\infty$, it will describe the infinite past. Earlier, the suggestion was made that it should be used to describe the early phase,[17] it being assumed that at this period the Universe was filled with a hot and dense plasma. But the idea that a plasma, i.e., hot matter, could have negative pressure appears strange. During the last few years, the hypothesis of the de Sitter solution has been bruited again: Initially (and implicitly) Gurovich and Starobinskiĭ[18] in 1979, and then in 1979 and 1980 Starobinskiĭ[19,20] explicitly (and with consideration of all stages of the transition from the de Sitter solution to the ordinary Friedmann solution) proposed the de Sitter model with right-hand side of the Einstein equations corresponding to true vacuum polarization.

It is important to note that: 1) in the de Sitter solution, the symmetry of the metric determines the necessary symmetry of the energy density and pressure of true vacuum polarization; 2) "real" or "normal" matter or radiation must be absent in the initial stage—it would spoil the solution. In Starobinskiĭ's solution, the true vacuum polarization is not the cosmological constant of flat space-time, although it has the same property as the cosmological constant, $p/\varepsilon = -1$, because of the symmetry of the curvature in the de Sitter solution.

The true vacuum polarization is proportional to the fourth power of H, which characterizes the rate of change of the scale factor during the initial stage of the expansion; the true vacuum polarization is also proportional to the number N of elementary-particle species.

Therefore, Einstein's equations have the form[9)]

[9)] For the mathematical reader, we note that in the de Sitter metric all components of the fourth-rank curvature tensor R^i_{klm} can be expressed symmetrically in terms of the local values of g_{ik} and are proportional to H^2. Therefore, R_{ik} is simply equal to $k_1H^2g_{ik}$, the curvature scalar R is k_2H^2, and, therefore, the left-hand side $R_{ik} - \frac{1}{2}g_{ik}R$ of Einstein's equations is $k_3H^2g_{ik}$, where k_1, k_2, k_3 are known dimensionless numbers. For the given metric, these relations are invariant; they do not depend on the coordinate mesh specified in space-time. The de Sitter metric given above describes part of the surface of a four-dimensional hyperboloid embedded in a flat five-dimensional Minkowski space. The de Sitter metric is the four-dimensional (pseudo) analog of Lobachevskiĭ space of constant curvature. The word "pseudo" is used because one of the differentials (cdt) has a sign opposite to the other three.

It follows from dimensional considerations that at large curvature, when the particle mass can be ignored, the true vacuum polarization can be expressed in terms of the square of the curvature tensor and the second derivatives of this tensor. The calculation of the true vacuum polarization does not depend on the gravitational constant, and can be calculated in a given metric. The metric is characterized by the single dimensional quantity H, so that up to a numerical factor the form of the expression on the right-hand side given below in the text follows uniquely. The symmetry of the de Sitter solution has the consequence that T_{ik} is proportional to g_{ik}. Therefore, the left- and right-hand sides of Einstein's equations are proportional to g_{ik}. Solving one of the equations, we automatically ensure solution of all ten equations. One can add that the de Sitter metric is conformally flat, and it can be written in the form $\eta^{-2}[d\eta^2 - dx^2 - dy^2 - dz^2]$. Therefore, production of real particles does not commence until deviations of the metric from the de Sitter metric begin.

$$R_{ik}-\frac{1}{2}g_{ik}R=\frac{8\pi G}{c^4}T_{ik(\mathrm{TVP})}\approx GNH^4\frac{\hbar}{c^7}g_{ik}.$$

They have a nontrivial solution (besides $H=0$) for a definite

$$H\sim\frac{1}{\sqrt{GN\hbar/c^3}}.$$

Here, we have systematically omitted dimensionless factors, such as squares and cubes of the number π, but we have retained N, which, as we have noted above, is apparently of the order of hundreds.

For the H given above and N many times greater than unity, we are justified in using the above equations, and the equations and their solution are correct.

The applicability of the classical treatment of general relativity is governed by

$$H_{\mathrm{Pl}}=\frac{1}{t_{\mathrm{Pl}}}=\frac{1}{\sqrt{G\hbar/c^5}};$$

here, t_{Pl} is called the Planck unit of time and in order of magnitude is 10^{-43} sec.

As soon as he had introduced the constant h, Planck recognized the fundamental importance of this constant for the whole of physics, extending far beyond the theory of thermal radiation.

Two great constants—the Newtonian gravitational constant $G=6.7\cdot10^{-8}$ cm$^3\cdot$sec$^{-2}\cdot$g^{-1} and the velocity of light $c=3\cdot10^{10}$ cm/sec—were already known several centuries before Planck's work.

Planck advanced the far reaching idea that the three constants G, c, and h are sufficient to determine the natural units of mass, distance, and time. From the modern point of view, classical general relativity applies only to a metric that changes little in the time t_{Pl} and over a distance $l_{\mathrm{Pl}}=ct_{\mathrm{Pl}}$. Applied to the de Sitter metric, this gives the condition $H<1/t_{\mathrm{Pl}}$.

Returning to Starobinskii's solution, we note that it can be regarded at the level of classical general relativity, using classical ideas of space and time, and therefore we quantize all the remaining fields, i.e., the electromagnetic, electron, etc. The large number N of quantized fields makes it possible to preserve the classical notions of space and time. More precisely, we may add that the small ripples on the space-time metric identified with gravitational waves can be *quantized*, but this does not spoil the *classical* picture of the averaged metric. It is important to note that for large N we have $H<1/t_{\mathrm{Pl}}$, and this is the condition of applicability of classical theory.

We now consider gravitational waves in the proposed metric without singularity. Starobinskiĭ[19] considers the fate of these gravitational waves. Initially, they are at the level of zero-point oscillations, but then they grow in amplitude, forming real gravitons, and then classical gravitational waves. It is possible that these waves will be observed by satellites before the end of this century. We are as anxious to have experimental data on gravitational waves of cosmological origin as we are to know whether the proton decays and what are the masses of the various species of neutrino. For this it would be worth living 20 or 30 years more!

The de Sitter solution is unstable, and Starobinskiĭ[20] has investigated in detail how the law $a(t)\sim e^{Ht}$ of expansion of empty (apart from true vacuum polarization) space breaks down with the course of time. He shows that after a period of pulsations $a(t)\propto t^{2/3}[b+r\cos\varphi(t)]$ due to the production of real particles a very hot neutral plasma arises and the expansion law is replaced by the normal Friedmann law for the case with radiation dominance:

$$a(t)\propto\sqrt{t}.$$

Then follows a period which is considered in detail in the review of Ref. 21 by Dolgov and the present author. Because of baryon nonconservation and a small baryon–antibaryon asymmetry of the theory, a baryon excess is obtained. This idea is currently very popular. It should be noted here that the possibility of baryon nonconservation was pointed out for the first time by Weinberg in 1964 in the lectures of Ref. 22. It is based on the circumstance that there is no massless vector field that could be coupled to the baryon charge in the way that the electromagnetic field is coupled to the electric charge. In the same lectures, Weinberg writes that the baryon asymmetry of the Universe could be due to baryon nonconservation. But Weinberg had in mind a steady state Universe (Hoyle, Bondi) with continuous creation of matter. In Weinberg's well-known book The First Three Minutes[23] the law of baryon conservation is included among the fundamental laws of nature, and his idea of 1964 is not mentioned. The idea of combining baryon nonconservation with the theory of a hot Universe is due to Sakharov.[24] Protons and neutrons are effectively stable (lifetime $>10^{30}$ years) at low temperature, but the process of variation of the baryon charge can proceed fairly rapidly (in a time shorter than 10^{-6} sec) at high temperature due to dissociation of baryons into quarks. The present state of the question is considered in the already quoted review.[21] Finally, once the temperature has sunk below 1 GeV $=10^{13}$ °K the well-known scenario of the evolution of the Universe commences. We recall the main events: 1) nucleosynthesis of helium-4 and deuterium in the primordial plasma; 2) the era of radiation-dominated plasma consisting of photons and neutrinos with a low-density, ionized gas as a small admixture; 3) decoupling of the radiation from the matter after the electrons and protons have combined to form neutral hydrogen atoms; 4) growth of perturbations leading to the formation of galaxies, stars, and everything else; see Weinberg's book[23] or the more detailed and mathematical Refs. 25 and 26. All these parts of the scenario are now as well known as the fact that the earth is round and that the earth and the other planets revolve around the sun. Weighty arguments relating to a possible rest mass of the neutrinos change the theory quantitatively. There is a change in the picture of the formation of structure in the Universe (the formation of clusters of galaxies, etc.). The words written above about the level of our knowledge concerning the complete scenario of the evolution sound too confident. Has it not been said that in cosmology (or, quite generally, in astrophysics) that "one is frequently in error but never in doubt?"

However the theories of nucleosynthesis and of the formation of the equilibrium microwave background radio radiation do indeed remain an unshakable foundation of the theory of the hot Universe. But until recently there has remained the feeling that the picture is incomplete.

Until the ideas associated with the de Sitter solution and vacuum polarization arose, troublesome questions kept arising. What is the beginning? What was there before the expansion began 20 billion years ago?

We now have a solution without an abrupt beginning extending from $t=-\infty$.

The "a" has been used advisedly, since the solution may not be the true solution—there could be other preferable solutions. But even a solution is a giant step forward. In this solution, the theory of the hot Universe is no longer associated with an arbitrary act of creation in a singularity. One important qualitative argument against the theory of an expanding universe may have been already resolved. An important detail of the new conception is the circumstance that the de Sitter law of expansion solves the problem of causality in its stride. Any two points or particles (at present widely separated) were, in the distant de Sitter past, at a very small, exponentially small distance. They could be causally connected in the past, and this makes it possible, at least in principle, to explain the homogeneity of the Universe on large scales.

Of course, there still remain many varied questions about different stages in the evolution of the Universe. The most difficult question concerns the instability of the de Sitter solution. One can say that this instability in the new stage reproduces, under other conditions, the instability that ordinary matter would have if the pressure in it were negative. If in the de Sitter solution there is a finite probability (per unit volume and unit time) for transition to another state, can one continue this solution to the region $t=-\infty$? Does it solve the problem of a world that exists for ever but in different forms? Further, the de Sitter solution is not unique: There are three solutions corresponding to flat, closed, and open universes (see any book on cosmology). The choice between these variants on the basis of the data of astronomical observations was the favorite problem of cosmologists for many years. The type of universe corresponding to this choice then remains unchanged during the expansion. However, the problem of choosing the variant remains difficult and is still unsolved.

Could not fundamental theory justify the choice of the flat variant, which, very probably, corresponds to modern observational data when allowance is made for the neutrino mass and indirect arguments about the growth of perturbations?

It is very important to develop the theory of density perturbations in the de Sitter model with true vacuum polarization. This is one further type of perturbation of the initial metric in addition to gravitational waves. The subsequent behavior of these two types of perturbation is very different.

In the late stages, gravitational waves are damped and difficult to observe. Density perturbations grow because of the gravitational instability, and it is because of these perturbations that galaxies are formed from a weakly inhomogeneous gas.

Could it not be that these perturbations arise from quantum zero-point oscillations of the initial de Sitter metric similarly to gravitational waves? What spectrum and what amplitude of the density waves can be expected? Does the theory agree with what is known about the present spatial structure of the Universe and the amplitude of the fluctuations in the temperature of the microwave background?

Other questions concern the physics of the high-temperature period. There was a time when the temperature considerably exceeded the values corresponding to the rest masses of the known particles. As yet, an exact theory of such a plasma does not exist. Was there a phase transition, as suggested by the Soviet physicists Kirzhnits and Linde[27] (see also Ref. 28), and were the filaments and walls then formed analogous to the structures that arise when a liquid crystallizes? What is a plasma with free quarks?

The period closest to ours is characterized by exact knowledge of the fundamental laws determining the behavior of the considered particles. But the problems relating to this period are mathematically complicated. Three-dimensional hydrodynamics and radiative transfer of heat are the problems that must be solved if we are to study in detail the formation of galaxies and stars. The answers concerning the structure of the Universe, i.e., its inhomogeneity, are statistical, which makes comparison of them with observations difficult. In addition, it is necessary to know much more about the nuetrino mass than is currently known. It is here worth mentioning that the Hungarian scientists Marx and Salai were the first who began to consider actively the part played by a neutrino rest mass in cosmology back in the sixties.

We recall that in 1974 astronomical indications of the existence of hidden mass were given independently by the Soviet astronomers Éĭnasto, Kaasik, and Saar and the Americans Peebles, Ostriker, and Yahil.

Marx and Salai insisted on explaining the hidden mass (the excess mass compared with the sum of the masses of the stars and gas) of galaxies and clusters of galaxies by assuming that the clusters of galaxies and individual large galaxies are surrounded by clouds (halos) consisting of neutrinos. It is obvious that such a picture is possible only if the neutrino rest mass is nonzero and the neutrinos move with velocities much less than the velocity of light. At the present time, the part played by massive neutrinos in the evolution of perturbations, both before and after recombination of the plasma, is at the center of attention.

This very brief review shows that there is no danger of unemployment for theoreticians occupied with astronomical problems.

Penetration to the great secret of the beginning of the

Universe is, perhaps, the most exciting event in the development of the natural sciences. One is lucky to be alive at such a time and witness the dramatic moment when human knowledge reaches maturity.

I thank A. D. Dolgov, L. B. Okun', A. A. Starobinskii, and M. Yu. Khlopov for discussions, valuable comments, and assistance. I should especially like to thank V. L. Ginzburg, who read the first two drafts of the paper, for valid and well intentioned critical comments.

[1]S. Weinberg, Rev. Mod. Phys. **52**, 121 (1980).
[2]R. P. Feynman, Usp. Fiz. Nauk **91**, 29 (1967) [Russian translation of Nobel prize lecture].
[3]J. Schwinger, Usp. Fiz. Nauk **91**, 49 (1967) [Russian translation of Nobel prize lecture].
[4]S.-I. Tomonaga, Usp. Fiz. Nauk **91**, 61 (1967) [Russian translation of Nobel prize lecture].
[5]V. F. Weisskopf, Shift of the levels of atomic electrons, Phys. Rev. **75**, 1240 (1949) [Russian translation published in book by IL, Moscow (1950)].
[6]J. H. Field, E. Picasso, and F. Combley, Usp. Fiz. Nauk **127**, 553 (1979) [Sov. Phys. Usp. **22**, 199 (1979)].
[7]Yu. S. Barash and V. L. Ginzburg, Usp. Fiz. Nauk **116**, 5 (1975) [Sov. Phys. Usp. **18**, 305 (1975)].
[8]S. Hawking, Nucl. Phys. **B144**, 349 (1978).
[9]Ya. B. Zel'dovich, Usp. Fiz. Nauk **95**, 209 (1968) [Sov. Phys. Usp. **11**, 381 (1968)].
[10]Ya. B. Zel'dovich and R. A. Syunyaev, Pis'ma Astron. Zh. **6**, 451 (1980) [Sov. Astron. Lett. **6**, 249 (1980)].
[11]V. L. Ginzburg, D. A. Kirzhnits, and A. A. Lyubushin, Zh. Eksp. Teor. Fiz. **60**, 451 (1971) [Sov. Phys. JETP **33**, 242 (1971)].
[12]A. D. Sakharov, Dokl. Akad. Nauk SSSR **177**, 70 (1967) [Sov. Phys. Dokl. **12**, 1040 (1968)].
[13]L. D. Landau and E. M. Lifshitz, Teoriya polya, Nauka, Moscow (1973); English translation: The Classical Theory of Fields, 4th ed., Pergamon Press, Oxford (1975).
[14]Ya. B. Zel'dovich and L. P. Pitaevskij, Commun. Math. Phys. **23**, 185 (1971).
[15]S. Hawking, Commun. Math. Phys. **18**, 301 (1970).
[16]Ya. B. Zel'dovich, in Magic without Magic: John Archibald Wheeler (ed. G. R. Klauder), Freeman, San Francisco (1972), p. 277.
[17]L. E. Gurevich, Astrophys. Space Sci. **38**, 67 (1975).
[18]V. Ts. Gurovich and A. A. Starobinskiĭ, Zh. Eksp. Teor. Fiz. **77**, 1683 (1979) [Sov. Phys. JETP **50**, 844 (1979)].
[19]A. A. Starobinskiĭ, Pis'ma Zh. Eksp. Teor. Fiz. **30**, 719 (1979) [JETP Lett. **30**, 682 (1979)].
[20]A. A. Starobinskiĭ, Phys. Lett. **B91**, 99 (1980).
[21]A. D. Dolgov and Ya. B. Zel'dovich, Usp. Fiz. Nauk **130**, 559 (1980) [Rev. Mod. Phys. **53**, 1 (1981)].
[22]S. Weinberg, in: Lectures on Particles and Fields (eds. S. Deser and K. Ford), New York (1964), p. 482.
[23]S. Weinberg, The First Three Minutes, Basic Books Publ., New York (1977).
[24]A. D. Sakharov, Pis'ma Zh. Eksp. Teor. Fiz. **5**, 32 (1967) [JETP Lett. **5**, 24 (1967)].
[25]Ya. B. Zel'dovich and I. D. Novikov, Stroenie i évolyutsiya vselennoĭ (Structure and Evolution of the Universe), Moscow (1975).
[26]P. J. E. Peebles, Physical Cosmology, Princeton University Press (1971) [Russian translation published by Mir, Moscow (1975)].
[27]D. A. Kirzhnits and A. D. Linde, Phys. Lett. **B42**, 471 (1972).
[28]Ya. B. Zel'dovich, Mon. Not. R. Astron. Soc. **192**, 246 (1980).

Translated by Julian B. Barbour