THE ASTROPHYSICAL JOURNAL, 196: 661–670, 1975 March 15

ON THE ORIGIN OF THE MICROWAVE BACKGROUND

F. HOYLE
Kitt Peak National Observatory

Received 1974 August 15

ABSTRACT

The background of microwave radiation is known to be remarkably uniform over the sky, although the regions giving rise to the radiation in widely separated elements of solid angle have, according to the usual cosmological theories, always been out of communication with each other. Using a new approach to the big-bang cosmologies, an explanation of this uniformity is given.

The intensity of the background appears to be related to the energy of conversion of hydrogen to helium within galaxies. Yet this circumstance is regarded as coincidental in the usual theories. Here it receives explanation.

Subject headings: cosmic background radiation — cosmology

I. THE MASS FIELD

A mass field at a general point x can be defined by a summation over all particles,

$$M(x) = \sum_a \int \tilde{G}(x, A)\epsilon_A da\,, \tag{1}$$

the point A being located at the element da of the world line of a typical particle, denoted by a. The scalar Green's function $\tilde{G}(x, A)$ will be defined in a later section in terms of a certain wave equation, while the dimensionless quantity ϵ_A is a coupling constant. Writing the coordinate displacement along da as da^i,

$$da^2 = g_{ik}da^i da^k\,, \tag{2}$$

the metric tensor of the Riemannian space being g_{ik}. The situation is illustrated in figure 1.

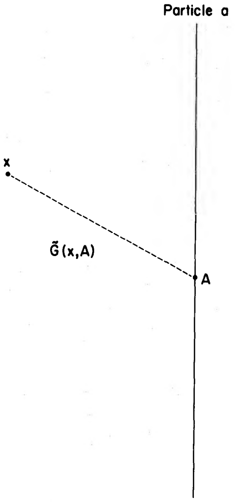

FIG. 1.—The function $G(x, A)$, taken for all elements on the paths of all particles, builds the mass field $M(x)$.

The mass $m_b(B)$ of a particle b, at a point B of its path, is then taken to be given by

$$m_b(B) = \epsilon_B M(B) = \epsilon_B \sum_a \int \tilde{G}(B, A)\epsilon_A da\,, \tag{3}$$

the field M coupling to the particle through the constant ϵ_B. We shall regard ϵ_A, ϵ_B as simple numbers, taking each of them to be either ϵ (>0) or $-\epsilon$, with ϵ a fixed number. That is to say, all our "particles" are structureless and similar to each other. An attempt can be made to represent "real" particles by taking ϵ_A, ϵ_B to be matrices, with the product $\epsilon_A \cdot \epsilon_B$ an invariant with respect to transformations in the abstract space that determines the structure of the particles. However, such a development involves problems which go beyond the scope of this paper.

As in electrodynamics, we therefore have both plus and minus contributions to the mass field. But whereas in electrodynamics the strength of the forces is so great that plus and minus charges are everywhere distributed with nearly equal densities, we contemplate here that large scale aggregates can exist some of which make only plus contributions, others making only minus contributions, as indicated schematically in figure 2. *The regions of figure 2 are to be thought of as large compared with the range of astronomical observation.*

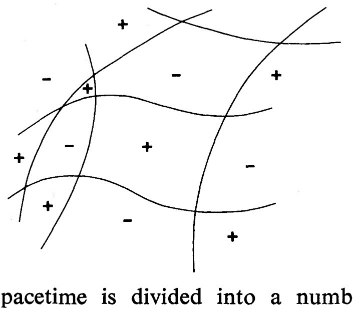

FIG. 2.—Spacetime is divided into a number of four-dimensional volumes which make plus and minus contributions to the mass field. A plus aggregate is bordered by minus aggregates, and vice versa.

Cosmological distances, as ordinarily understood, fit into a single aggregate. Our experience in astronomy is therefore confined to one sign for the contributions to the mass field.

Averaged on a scale much greater than that of practical astronomy, plus and minus contributions are taken to be equally important, so that $M(x)$ will sometimes be plus and sometimes minus, depending on the position of the point x in relation to the aggregates. Hence there will be three-dimensional surfaces in spacetime on which $M(x) = 0$, separating four-dimensional regions with $M > 0$ from regions with $M < 0$. Figure 2 can be reinterpreted as a schematic representation of such a set of surfaces.

It is worth noting that $M < 0$ does not necessarily require particle masses to be negative, since the particle couplings may take the same sign as the mass field, in which case particle masses are never negative.

The disposition of the plus and minus aggregates is not arbitrary, but must satisfy the gravitational equations—as the disposition of plus and minus charges satisfy the electrodynamic equations. The gravitational equations are determined from an action S defined by a summation over all particles

$$S = -\sum_a \int m_a(A)da\,, \tag{4}$$

by requiring $\delta S = 0$ to the first order in small quantities for an arbitrary variation $g_{ik} \rightarrow g_{ik} + \delta g_{ik}$ of the metric tensor. The details of the derivation of the gravitational equations, which will not concern us here, have been given by Hoyle and Narlikar (1974). The situation differs from the Einstein theory in that no term of the form

$$\frac{1}{16\pi G}\int R(-g)^{1/2}d^4x \tag{5}$$

appears in equation (4). Yet with an appropriate choice for the constant ϵ (taking the place of a choice for G) the present theory contains the gravitational equations of the Einstein theory, as will be seen below.

II. LOCAL MEASUREMENTS OF PHYSICAL QUANTITIES

The following remarks refer to a local situation in which the geometry is taken to have the flat-space Minkowski form.

There is no more basic way for measuring time intervals than by counting the electromagnetic oscillations of a monochromatic wave emitted in a suitably chosen atomic transition from atoms that are stationary. The transition between the hyperfine levels of the ground state of ^{133}Cs is used in practice. And, by generating standing waves inside a box with reflecting walls, static spatial displacements can be determined, simply by counting the number of standing waves between specified points.

The determination of the instantaneous velocity of a particle, or of the speed of light, involves the measurement of both a time interval and a static spatial distance. It is important to notice that provided light moves through a vacuum, and provided electromagnetic radiation from the same atomic transition is used to measure both time and spatial intervals, then inevitably the speed of light is found to be unity. It is only when the atomic transition chosen to determine intervals of time is different from the transition chosen to determine spatial intervals, or when quite arbitrary multiples of the atomically determined time and space units are used, that the speed of light can be anything other than unity.

The $2p_{10}$ to $5d_5$ transition of ^{86}Kr is used in practice to determine spatial distances, the *meter* being defined as a multiple 1650763.73 of the resulting spatial unit. The *second* is defined as a multiple 9192631770 of the time unit given by ^{133}Cs. When we say that the speed of light is 299792500 meters per second, we are simply stating the ratio of these two practical measuring scales. An infinity of such arbitrary prescriptions could indeed be devised, all leading to different numerical values for the speed of light, each expressing the arbitrariness of the procedure. When the natural space and time units given by the same monochromatic wave are used, however, the speed of light is always unity. We shall take it to be so in the following discussion.

Because the charge e and the mass m of a particle are usually first encountered through the study of classical physics, we tend to think of e and m as having an existence apart from the Planck constant $\hbar$. But the electronic charge e always occurs in quantum mechanics in the fine-structure combination, $e^2/\hbar$, and the particle mass m occurs either in the ratio $m/\hbar$ or as a ratio with respect to the mass of another particle, like m_e/m_p for the electron and proton masses. Every formula determining an experimental result can be constructed according to quantum mechanics from powers of $e^2/\hbar$ (and from dimensionless constants of similar form for the weak and strong interactions), from powers of $m_e/\hbar$, $m_p/\hbar, \ldots$, for various particles, and from either known or calculable dimensionless numbers—the latter usually involving matrix elements. Since classical physics is contained within quantum physics, e and m cannot therefore be separated from $\hbar$, except by absorbing $\hbar$ into each particle mass and into e^2 (and the other coupling constants). This can be done by writing $\hbar = 1$ in all the usual formulae. Then e^2 becomes the fine-structure constant, $e^2 = 7.297351 \times 10^{-3}$, and all particle masses have the dimensionality of an inverse length—the Compton wavelength of a particle being just the reciprocal of its mass.

It will be clear then, that vague suggestions to the effect that perhaps the speed of light might be variable from place to place, or that Planck's constant might be variable, are without meaning. Taken in a sensible way, both Planck's constant and the speed of light are unity, and they are so everywhere, provided that whatever spacetime location we are concerned with we elect to use flat Minkowski space for the local geometry.

The oscillation frequency of radiation emitted by an explicit transition is determined by the various particle masses—electron, proton, neutron—and by the structure of the atom itself, involving the dimensionless fine-structure constant and in some small degree the nuclear couplings also. Provided all dimensionless quantities are taken to be fixed numbers, including particle mass ratios, the radiation frequency is determined by these fixed numbers and by any one of the particle masses, say that of the electron, m_e. The latter may be considered to be variable with respect to the spacetime location, in accordance with the ideas of the preceding section, but the dimensionless quantities are not taken to be variable.

Intervals of time and space can therefore be considered to be measured with respect to a unit determined by m_e^{-1}. Moreover, the dimensionalities of all physical quantities can be expressed as some power of m_e, m_e^n say. As examples, pressure and energy density have $n = 4$; current density and surface tension have $n = 3$; luminosity, force, and the electromagnetic field have $n = 2$; energy, mass, and frequency have $n = 1$; length has $n = -1$.

Every experiment consists, when its procedures are analyzed, in the counting of a dimensionless number, which is always made up as a product of physical quantities and their inverses in such a way that the sum of the dimensionalities add to zero. No physical quantity with $n \neq 0$ is ever measured, except as a ratio to another quantity of the same dimensionality. Hence it follows that, so long as $m_e(x)$ is only slowly variable with respect to the spacetime position x, as would be the case if m_e were to vary only on a cosmological time scale, no local laboratory experiment can detect the variation.

It is only when a dimensionless number can be measured involving two widely separated locations that a variation of m_e is in principle detectable. For example, by observing light from a distant object it is possible to determine frequency ratios of spectrum lines emitted from similar atoms, some present in the object, others in the laboratory. This observation depends directly on the ratio of m_e in the local laboratory to m_e in the object. Such a comparison of m_e is blurred, however, by the circumstance that the large-scale spacetime geometry can also affect the observed frequency ratio—in a manner familiar from the usual cosmological studies. This apparent ambiguity between the effect of world geometry and the effect of variable particle masses can be considered with precision, however, through the concept of conformal invariance.

III. CONFORMAL INVARIANCE

Consider a scalar function of position $\Omega(x)$, with the property that Ω is never infinite and never zero. For definiteness, let Ω be positive. A transformation from the Riemannian space

$$ds^2 = g_{ik} dx^i dx^k \tag{6}$$

to the space

$$ds^{*2} = \Omega^2 g_{ik} dx^i dx^k \tag{7}$$

is known as a conformal transformation. Notice that such transformations are not to be confused with coordinate transformations. A coordinate transformation never changes the length ds associated with the displacement between neighboring points, whereas the conformal transformation from equation (6) to (7) changes the length associated with dx^i from ds to $ds^* = \Omega ds$.

Again for definiteness, take dx^i to be timelike. By choosing locally flat space, and by arranging for dx^i to be along the time axis, we can seek to determine ds, in the manner discussed in the previous section. Radiation from stationary atoms determines a time unit of the form

$$\frac{\text{dimensionless constant}}{m_e}, \tag{8}$$

the numerator here being determinable through the evaluation of the matrix element associated with the atomic transition and through known fixed constants. If the geometry is given by equation (6), the number determined in this way for a physically specified dx^i is equal to

$$\frac{m_e ds}{\text{dimensionless constant}}. \tag{9}$$

Hence, with the denominator of equation (9) known, the dimensionless product $m_e ds$ is determined. Notice that ds is not itself determined, since we do not know m_e in terms of any unit more fundamental than the electron mass itself.

Suppose now that we *both* change the geometry from equation (6) to (7) and also change m_e to a different mass, m_e^* say. Then the experimentally determined number associated with the same physically specified displacement dx^i must be equal to

$$\frac{m_e^* ds^*}{\text{dimensionless number}}, \tag{10}$$

the denominator here being the same known constant as before. It follows that the product $m_e^* ds^*$ must be the same as $m_e ds$, so that to avoid contradiction m_e^* has to be chosen so that

$$m_e^* = \Omega^{-1} m_e\,. \tag{11}$$

Provided we associate m_e with the geometry (6) and $m_e^* = \Omega^{-1} m_e$ with the geometry (7), spacetime measurements cannot distinguish between the two geometries.

Is there any other way in which we might distinguish these two possibilities? At first sight there might seem to be ways. Suppose, for example, that we attempt to determine the electromagnetic influence of particle b on particle a. For m_e and the geometry (6) we have[1]

$$D\left(m_a g_{ik} \frac{da^k}{da}\right) = e_a F^{(b)}{}_{ik} \cdot da^k \tag{12}$$

[1] The notation D in equation (12) denotes the part of the change of $m_a g_{ik} da_k/da$ that is due to the electromagnetic field.

for an element da^k of the path of particle a. Here e_a is the charge of particle a, m_a is its mass, da is the element of length associated with da^i taken with respect to (6), and $F^{(b)}{}_{ik}$ is the electromagnetic field of particle b, determined by

$$F^{(b)ik}{}_{;k} = -4\pi e_b \int \frac{\delta_4(x - B)}{[-g(B)]^{1/2}} db^i\,, \qquad (13)$$

$$F^{(b)}{}_{ik;j} + F^{(b)}{}_{ji;k} + F^{(b)}{}_{kj;i} = 0\,, \qquad (14)$$

the covariant derivatives in equations (13) and (14) being determined with respect to g_{ik} as metric tensor, $\delta_4(x - B)$ being the four-dimensional Dirac delta function, and B being a point at the element db^i of the path of particle b. The left-hand sides of equations (13) and (14) are evaluated at a general field point x.

The corresponding equations for the geometry (7), involving the starred particle mass, $m_a{}^* = \Omega^{-1} m_a$, are

$$D\left(m_a{}^*\Omega^2 g_{ik}\frac{da^k}{da^*}\right) = e_a F^{*(b)}{}_{ik} da^k\,, \qquad (15)$$

$$F^{*(b)ik}{}_{;k} = -4\pi e_b \int \frac{\delta_4(x - B)}{[-g^*(B)]^{1/2}} db^i\,, \qquad (16)$$

$$F^{*(b)}{}_{ik;j} + F^{*(b)}{}_{ji;k} + F^{*(b)}{}_{kj;i} = 0\,, \qquad (17)$$

the covariant derivatives in equations (16) and (17) being evaluated with respect to $\Omega^2 g_{ik}$ as the metric tensor, and da^* being the length associated with da^i also taken with respect to $\Omega^2 g_{ik}$ as the metric tensor. Do equations (12) and (15) give any difference in the motion of particle a?

With $m_a{}^* = \Omega^{-1} m_a$, we have

$$m_a \frac{da^k}{da} = \Omega^2 m_a{}^* \frac{da^k}{da^*}\,, \qquad (18)$$

so we are concerned on the left-hand sides of equations (12) and (15) with varying the same quantity. At first sight, we might expect the right-hand sides of these equations to be different, since apparently different fields $F^{(b)}{}_{ik}$ and $F^{*(b)}{}_{ik}$ are involved. But it turns out from equations (13), (14), and from (16), and (17), that $F^{(b)}{}_{ik} = F^{*(b)}{}_{ik}$, and in fact the attempt to distinguish between the geometries (6) and (7) fails. The result $F^{(b)}{}_{ik} = F^{*(b)}{}_{ik}$ is expressed by saying that Maxwell's equations are conformally invariant.

The present considerations are classical, but the same situation arises also in quantum mechanics [for details, see Hoyle and Narlikar (HN) 1974].

The possibility of distinguishing between (6) and (7) can therefore depend only on gravitation. Yet an extensive consideration of gravitational effects (again for details, see HN 1974) has shown that, provided the function $\tilde{G}(x, A)$ in equation (1) satisfies the scalar wave equation

$$\Box_x \tilde{G}(x, A) + \tfrac{1}{6}R(x)\tilde{G}(x, A) = \frac{\delta_4(x - A)}{[-g(A)]^{1/2}}\,, \qquad (19)$$

gravitational effects also fail to distinguish between (6) and (7). Indeed, the adoption of (19) forces (11), $m_e{}^* = \Omega^{-1} m_e$, to hold good. Thus the adoption of equation (19) removes the need to assume (11).

The failure to distinguish between the geometry (6) associated with m_e and the geometry (7) associated with $m_e{}^*$ is total. No distinction is possible through any observation or through any experiment.

We now take the view that to have attempted to distinguish between (6) *and* (7) *was an irrelevant problem.* It makes no physical difference whether we choose (6) or (7), provided we relate the particle masses in the two geometries through equation (11)—or what is the same thing, through (1), (3), and (19). All geometries of the form (7), given by various choices for the function $\Omega(x)$, are physically equivalent to one another. We describe this situation by saying that our system of physics is *conformally invariant.*

It may be noted that (19) is similar in many respects to the wave equation satisfied by the electromagnetic potential, except that instead of being a vector, equation (19) is scalar. Like the electromagnetic wave equation, (19) has advanced and retarded solutions, say $\tilde{G}_{\rm ret}(x, A)$, $\tilde{G}_{\rm adv}(x, A)$. We regard $\tilde{G}(x, A)$ as being uniquely determined by choosing the symmetric form, given by

$$\tilde{G}(x, A) = \tfrac{1}{2}[\tilde{G}_{\rm ret}(x, A) + \tilde{G}_{\rm adv}(x, A)]\,. \qquad (20)$$

IV. THE EINSTEIN CONFORMAL FRAME

Within a classical framework, suppose we attempt to make a complete physical solution for an assigned set of particles, which we regard as constituting the universe. The number in the set can be as large as we please. For each such particle, say a, we describe the path by four functions $a^i(a)$, where a is to be a measure of length along the path taken from a particular starting point. The determination of a is to be with respect to a so-far unknown metric tensor $g_{ik}(x)$.

At our disposal in this enterprise we have electromagnetic equations like (13), (14), for every charged particle; we have gravitational equations based on the action (4); and we have equations (1), (3), and (19) for determining the particle masses. Note that dynamical equations like (12) are contained within the gravitational equations, and so do not need to be considered separately. Our aim is to obtain information about the metric functions $g_{ik}(x)$ and also information about the functions $a^i(a)$ describing the paths of the particles.

From what was said in the preceding section we must expect to fail in an attempt to determine a unique metric tensor. We can only hope to obtain $g_{ik}(x)$ to within a conformal transformation. That is to say, the equations at our disposal cannot serve to distinguish between $g_{ik}(x)$ and $\Omega^2(x)g_{ik}(x)$. This ambiguity will show itself through the whole scheme of physical equations being self-consistent with respect to an infinite family of conformally related geometries.

Suppose we have determined a particular metric tensor $g_{ik}(x)$ with respect to which all our equations are consistent, and write $m_e(x)$ for the function we have found to represent the electron mass. Then we know that, provided $\Omega(x)$ is neither infinite nor zero, $g^*_{ik}(x) = \Omega^2(x)g_{ik}(x)$, $m_e^*(x) = \Omega^{-1}(x)m_e(x)$, must also be consistent with our equations. Choose

$$\Omega(x) = (\text{nonzero constant})^{-1}m_e(x)\,, \tag{21}$$

in which case $m_e^*(x)$ becomes just the constant appearing in equation (21). Hence even if our first consistent solution for the universe did not lead to an electron mass independent of the position of x, by a conformal transformation we can arrive at such a solution. This particular solution will be referred to as the Einstein conformal frame, because when particle masses do not depend on the position of x, *the gravitational equations based on (4) reduce to the Einstein equations* (again for details, see HN 1974).

At this stage it may be wondered just what has been achieved by the whole of the above discussion. If all conformal frames are physically equivalent, and if one of them is the Einstein theory, then nothing physically different from the Einstein theory has apparently been achieved. Yet one crucially important difference has in fact been achieved. The conformal transformation (21), required to pass to the Einstein frame from a general frame in which $m_e(x)$ is variable, *cannot be used at points x where $m_e(x) = 0$*. This means that although the Einstein frame can be used consistently within any one of the regions of figure 2, it cannot be used to pass from one such region to another.

The usual mysteries concerning the so-called origin of the universe begin now to dissolve. In the usual cosmological discussions based on the Einstein frame there is no means for getting beyond the boundary of "our aggregate," which has therefore come to be regarded as a metaphysical "origin" for the universe. The metric collapses in the Einstein frame at points x such that $m_e(x) = 0$, because $\Omega(x)$ given by (21) is then zero, so that $g^*_{ik} = \Omega^2(x)g_{ik}(x) = 0$ at all such points. Yet there is no requirement for $g_{ik}(x)$ to be zero. Conformal frames exist that can carry us smoothly across the zero surfaces of figure 2.

In order to understand the universe, we need to connect all the regions of figure 2 into a consistent whole. We shall see below, for example, how it is possible to understand the existence of the microwave background in terms of such a connection. In this we must avoid, for the reasons just stated, the use of the Einstein frame.

V. THE FORM OF THE SOLUTION NEAR A SPACELIKE SURFACE OF ZERO MASS

In this section we shall see how the usual cosmological symmetry postulates of homogeneity and isotropy can be replaced by an interpretation of the concept of being "near" to a spacelike surface of zero mass. We shall find the geometrical structure in such a locality to be of the form which is usually referred to as the Einstein–de Sitter model ($k = 0$ in the usual notation). To make easier the relation to the usual discussions, we begin by working in the Einstein conformal frame. This will be permissible because in these initial considerations we shall not attempt to cross a surface of zero mass.

By taking geodesics normal to the zero surface, coordinates can be set up which enable the line element to be expressed in the form

$$ds^2 = dt^2 + g_{\alpha\beta}dx^\alpha dx^\beta\ ;\quad \alpha, \beta = 1, 2, 3\,, \tag{22}$$

the spatial coordinates being x^1, x^2, x^3, and the time coordinate, $x^4 = t$, being measured along the geodesics. As in the usual discussion, the energy-momentum tensor is represented in terms of a smooth fluid low-pressure approximation,

$$T^{ik}(x) = \rho(x)\frac{da^i}{da}\frac{da^k}{da}\,, \tag{23}$$

$a^i(a)$ being the path of a typical element of the fluid. Close enough to the zero surface, and over sufficiently small ranges of x^1, x^2, x^3, the fluid density ρ can be considered to be a function of t only, while the elements of the fluid can be taken to follow parallel paths with da^i/da as the tangent vector.

By the concept of closeness to the zero surface we also mean that variations of the metric functions $g_{\alpha\beta}$ are much more markedly dependent on t than on x^1, x^2, x^3. Important consequences then follow from neglecting the weak variations of $g_{\alpha\beta}$ with respect to x^1, x^2, x^3.

It can be shown that

$$R_{4\alpha} - \tfrac{1}{2}g_{4\alpha}R = 0\ ;\quad \alpha = 1, 2, 3\,. \tag{24}$$

Hence the gravitational equations, which take the usual form, $R_{ik} - \frac{1}{2}g_{ik}R = -\kappa T_{ik}$, in the Einstein frame, require $T_{4\alpha} = 0$; $\alpha = 1, 2, 3$, whence it is easily shown that

$$\frac{da^\alpha}{da} = 0\ ;\quad \alpha = 1, 2, 3\,. \tag{25}$$

The gravitational equations require the path of a typical fluid element to be normal to the zero surface.

Next, it can be shown that by a suitable choice of spatial coordinates, say $x^1 = x$, $x^2 = y$, $x^3 = z$, the off-diagonal components of $g_{\alpha\beta}$ can be made zero, and the diagonal components can be made equal, $g_{11}(t) = g_{22}(t) = g_{33}(t) = -Q^2(t)$ say. Close enough to the zero surface the line element can therefore be expressed in the form

$$ds^2 = dt^2 - Q^2(t)[dx^2 + dy^2 + dz^2]\,. \tag{26}$$

Moreover, the gravitational equations also lead to $Q^2(t) \propto t^{4/3}$, and (26) takes the familiar form of the line element of the Einstein–de Sitter cosmological model. This result has been derived from the supposition that *close enough* ρ and $g_{\alpha\beta}$ vary more steeply in the direction normal to the zero surface than they do in directions parallel to the surface. Since we are not here making a uniformity postulate applicable to the whole universe, our results do *not* apply everywhere, as in the usual Friedmann models. Far enough away from the zero surface, things may be very different from the Einstein–de Sitter model.

The point of view to be taken at this stage is the following:

Although the region over which the Einstein–de Sitter model applies is only a small element of the whole universe, it nevertheless encompasses everything which the astronomer observes, even with the largest telescope.

To be able to approach and to cross a zero surface we must avoid the use of the Einstein frame. A suitable conformal transformation of (26) is needed. The choice $\Omega = Q^{-1}$ leads to a particular simple geometry, namely the Minkowski flat-space form

$$ds^{*2} = d\tau^2 - (dx^2 + dy^2 + dz^2)\,, \tag{27}$$

in which the new time coordinate τ is defined by

$$\tau = \int_0^t \frac{dt}{Q(t)}\,. \tag{28}$$

In this Minkowski conformal frame the electron mass $m_e{}^*$ is a function of τ. Sufficiently near the zero surface, $m_e{}^*$ can be expanded in powers of τ,

$$m_e{}^* = A\tau + B\tau^2 + \cdots\,. \tag{29}$$

The gravitational equations based on (4) turn out to require $A = 0$. Hence the leading term in the expression for $m_e{}^*$ is quadratic in τ, requiring particle masses to be of the same sign on both sides of the zero surface. The coefficient B can be related to the particle density and is positive.

It will be recalled that zero surfaces arise from the properties of the mass field, $M(x)$. This mass field changes sign at the zero surface. Particle masses, on the other hand, maintain the same sign, because the coupling of particles to the mass field changes across such surfaces. This behavior of the mass field and of the particle masses is illustrated in figure 3.

The situation we have arrived at for the Minkowski conformal frame is illustrated in figure 4. The particle trajectories are everywhere normal to $\tau = 0$, and the particle density is constant within the local region to which the above considerations apply. The local region contains the range of astronomical observation, and it also extends across the zero surface. It is this extension that permits discussion of the problem of the origin of the microwave background.

From here on, we drop the starred notation from (29), denoting the electron mass in the Minkowski frame by m_e. If we have occasion to refer to the electron mass in the Einstein frame, we shall denote it by $m_e{}^*$. That is to say, from here on we shall reverse the above starred and unstarred notations, writing

$$ds^2 = d\tau^2 - (dx^2 + dy^2 + dz^2)$$

in the Minkowski frame.

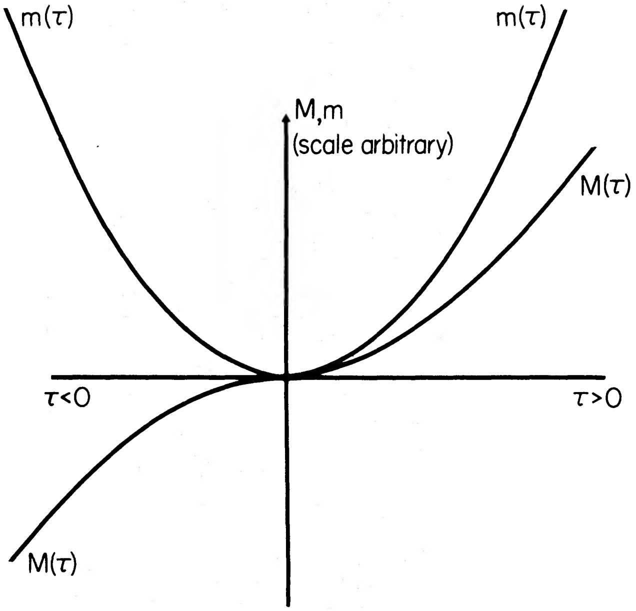

FIG. 3.—Although both the mass field M and the particle mass m vary quadratically with the time, the mass field changes sign at $\tau = 0$, whereas the particle mass is positive for all τ. This behavior is due to the coupling constant between the field and the particles, which is positive for $\tau > 0$ and negative for $\tau < 0$.

VI. THE ORIGIN OF THE MICROWAVE BACKGROUND

The sense of electromagnetic propagation will be taken on both sides of the zero surface to be in the sense of increasing τ. Radiation generated on one side then propagates away from the zero surface, and on the other side propagates toward the zero surface. Since experience on "our side" is that propagation occurs in the sense of the expansion of the universe, our side must have $\tau > 0$, our sense is away from the zero surface. Radiation generated on the "other side" goes toward the zero surface, as is illustrated in figure

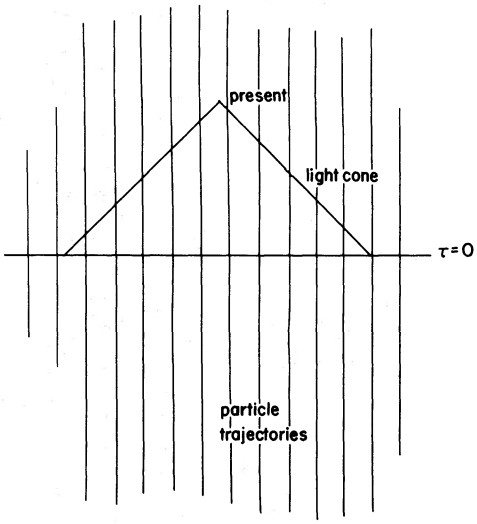

FIG. 4.—The range of astronomical observation is confined to the backward light cone taken from the present moment back only to $\tau = 0$.

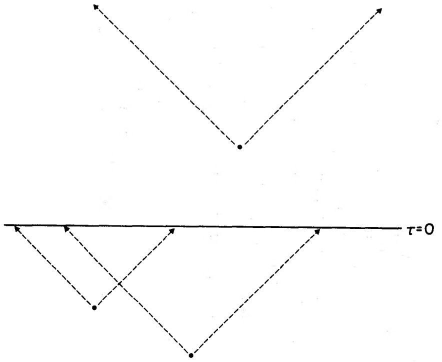

FIG. 5.—Radiation from sources at $\tau > 0$ propagates away from the surface, while radiation from sources at $\tau < 0$ propagates toward $\tau = 0$, where it is thermalized. The side $\tau > 0$ is "our side," and $\tau < 0$ is the "other side."

5. Near $\tau = 0$, such radiation is strongly absorbed and reemitted, and so becomes thermalized, because e^2/m_e becomes large as m_e decreases, so the Thomson cross section becomes large. Indeed, as $m_e \rightarrow 0$ absorption processes are formally divergent—in practice, quantum mechanical cross sections probably tend to a constant saturation level.

The suggestion of the present paper is that radiation generated on the other side becomes thermalized near $\tau = 0$ and then becomes the microwave background.

To obtain an estimate of the energy density of the thermalized radiation, we shall suppose that galaxies existed at $\tau < 0$, and that their stellar content was similar to the stellar content of the galaxies now observed by astronomers at $\tau > 0$. Writing τ_0 (>0) for the present epoch, the energy density of starlight produced on our side over the time range $0 < \tau < \tau_0$ has been estimated to be about 10^{-14} ergs cm^{-3}, less than the energy density of the microwave background by a factor of about 50. Thus with the situation on the other side taken to have been similar to that on our side, the energy density of stellar radiation produced over the time range $-\tau_0 < \tau < 0$ would have been some 50 times less than the energy density of the microwave background. This discrepancy is appropriate, however, because galaxies on the other side would also have produced starlight at times earlier than $-\tau_0$, and such earlier contributions could be large, as we shall now see.

Any physical quantity of dimensionality n that is constant in the Einstein frame varies like $m_e{}^n$ in the Minkowski frame. The luminosities of galaxies are usually taken to be approximately constants in the Einstein frame. Since luminosity has dimensionality $n = 2$, luminosities thus behave like $m_e{}^2$ in the Minkowski frame—i.e., like τ^4. Consequently the energy density of starlight propagated on the other side toward $\tau = 0$ will be dominated by emission at the largest $|\tau|$—i.e., by the earliest emission. Indeed the ratio of the emission over the range $\tau_1 < \tau < 0$ to that over $-\tau_0 < \tau < 0$ is evidently $(|\tau_1|/\tau_0)^5$, so that the emission is considerably enhanced even though $|\tau_1|$ may be only moderately larger than τ_0. In fact, for $|\tau_1|$ rather more than $2\tau_0$ we have $(|\tau_1|/\tau_0)^5 \simeq 50$, the required enhancement.

It is of interest to relate this estimate for $|\tau_1|$ to time measured in the Einstein frame, since we are better used to intervals expressed in the Einstein frame. Writing t_0 for the t-value corresponding to τ_0, present-day observational studies of the distances and of the redshifts of galaxies on our side suggest that t_0 is about 15×10^9 years. Working for the moment on our side, what then is the value of t_1 corresponding to $|\tau_1| \simeq 2\tau_0$? Since $Q(t) \propto t^{2/3}$, the definition (28) gives $\tau \propto t^{1/3}$. Hence

$$\frac{t_1}{t_0} = \left[\frac{|\tau_1|}{\tau_0}\right]^3 \simeq (50)^{3/5} \simeq 10\,. \tag{30}$$

Thus with $t_0 = 15 \times 10^9$ years we get $t_1 \simeq 150 \times 10^9$ years. Since the situation is symmetric about $\tau = 0$, we accordingly require, in the familiar t-scale of the Einstein frame, that galaxies existed on the other side backward from $t = 0$ for about 150 billion years. Normal production of starlight then explains the observed energy density of the microwave background.

VII. SOME TOPICS OF DISCUSSION

The explanation of the origin of the microwave background given in the preceding section raises many questions and problems. Some are difficult or outside the range of present knowledge. For example, the following questions:

Why do we happen to live near a zero surface?
What is the universe like when one does not live near a zero surface?
What causes the sign change of the contributions to the mass field?
Why are there plus and minus aggregates?

These questions are outside the scope of the present paper.

A more accessible question concerns the intensity of the microwave background. Could the energy density have been even larger than it is observed to be? This would have been so if normally emitting galaxies on the other side had extended over the t-scale backward from $t = 0$ for more than 150 billion years. Would we expect this to have been the case? As a possible negative answer to this question, it should be noted that the behavior $m_e \propto \tau^2$, on which the above estimates depend, is valid only *sufficiently close* to the zero surface. At $|t|$ of order 150 billion years the condition of being sufficiently close may well break down.

It is also the case that energy production by hydrogen to helium conversion cannot continue effectively for much more than this, because of the general exhaustion of the supply of hydrogen. One percent of the hydrogen in a galaxy like our own is consumed in a range of $|t|$ of about 5 billion years.

An observer on the other side at $|\tau| \sim \tau_0$ could experience an astrophysical situation much like ours, but the cosmological situation would be critically different, since light from distant galaxies would be blueshifted, not redshifted as it is on our side. To study this difference, consider first the redshift on our side. Working in the Minkowski frame, a galaxy at distance r has redshift $z = \Delta\lambda/\lambda$ given by

$$1 + z = \frac{m_e(\tau)}{m_e(\tau - r)} = \left(\frac{\tau}{\tau - r}\right)^2 , \qquad (31)$$

where $\tau(>0)$ is the epoch at which observation is made. The flux S from the galaxy is expressed in flat space by

$$S = \frac{L(\tau - r)}{4\pi r^2} , \qquad (32)$$

where $L(\tau - r)$ is the luminosity at time $\tau - r$, the moment at which light, observed at time τ $(>r)$, started its journey. Taking all the observed galaxies to have the same luminosity in the Einstein frame, we have $L(\tau - r) \propto m_e^2 \propto (\tau - r)^4$ in the Minkowski frame. Consequently the flux satisfies the proportionality

$$S \propto \frac{(\tau - r)^4}{r^2} . \qquad (33)$$

Eliminating r in equation (33) with the help of (31) gives

$$S \propto \frac{1}{(1 + z)[(1 + z)^{1/2} - 1]^2} , \qquad (34)$$

which is the well-known relationship between S and z for the Einstein–de Sitter model.

Defining the blueshift on the other side by $z = \Delta\nu/\nu$, $\Delta\nu$ being the frequency increase in the light from a distant galaxy taken for a spectrum line of local frequency ν, the corresponding formulae, with τ now <0, are

$$1 + z = \frac{m_e(\tau - r)}{m_e(\tau)} = \frac{(\tau - r)^2}{\tau^2} , \qquad (35)$$

$$S = \frac{L(\tau - r)}{4\pi r^2} , \quad S \propto \frac{(\tau - r)^4}{r^2} , \qquad (36)$$

$$S \propto \frac{(1 + z)^2}{[(1 + z)^{1/2} - 1]^2} . \qquad (37)$$

It is interesting that for $z \ll 1$ the flux is essentially proportional to z^{-2}, just as it is on our side, but that for $z > 3$ the flux actually increases with z.

This difference with respect to redshift and blueshift means that the two sides of a zero surface behave very much like the "bounce" which is sometimes postulated for the closed Friedmann model ($k = +1$ in the usual notation). But no successful mathematical model has been given for such a bounce, whereas here we have a complete mathematical scheme, at any rate within the limitations imposed by our initial assumption of the existence of plus and minus aggregates. Nor is there any need in the present case for an endless repetition of cycles of expansion (redshift phase) and contraction (blueshift phase) such as is postulated for the Friedmann model. Here we simply have the two sides of a zero surface. Far removed from a zero surface the universe could be very different from the restricted properties of an oscillatory cosmology.

Maxwell's equations admit electromagnetic propagation in either time sense, so the forms of propagation used above can simply be assumed to hold good. Alternatively, we can follow the point of view that, taken microscopically, both the advanced and retarded solutions of Maxwell's equations are generated equally. Then by introducing the Wheeler-Feynman concept of a response of the universe it is possible to obtain net propagation in one particular time sense. The physical condition required to obtain propagation forward in time is that the future light cone be perfectly absorbing (see HN 1974). In the situation contemplated here, this is certainly the case for propagation on the other side, since $m_e \to 0$ as $\tau \to 0$ ensures complete absorption at $\tau = 0$. On our side, $\tau > 0$, we require the future light cone to be also totally absorbing. This will be the case provided radiation emitted along our future light cone eventually reaches another surface of zero mass, as we may expect it to do at the boundary of our aggregate. Indeed, surfaces of zero mass, occurring extensively in the universe, provide an ideal means for controlling the sense of propagation of electromagnetic fields—and perhaps for other fields as well.

It should be a general rule, applicable through the whole universe, that the sense of electromagnetic propagation never reverses along any timelike line.

VIII. GALAXIES AND STARS ON THE OTHER SIDE

It is possible that the microwave background is by no means the only aspect of our experience to indicate the existence of the "other side." Galaxy formation on our side may do so. Some stars on our side may even be connected with stars on the other side. To begin an approach to this idea, it is useful to ask what happens to galaxies and to stars as they approach $\tau = 0$ from the other side.

Although we are used to thinking in terms of the Einstein frame, and although the present question might be answered by working in the Einstein frame (we are not concerned here with actually crossing the zero surface), there are important advantages in persisting with the Minkowski frame. We are well used to the geometry of the Minkowski frame, so much so that often enough when we claim to work in the Einstein frame we still tend to think geometrically in terms of flat space. This tendency will be avoided by taking Minkowski space as our basic conformal frame. Radiation is also more easily considered, since radiation propagates in the Minkowski frame without change of frequency—the frequency changes only when radiation interacts with matter, and in this respect our usual ideas continue to hold good.

Against these advantages, gravitation behaves peculiarly in the Minkowski frame, not only because the particle masses change with time, but because the gravitational "constant" G also changes. In the Einstein frame G is indeed constant; but being of dimensionality $n = -2$, G varies like m_e^{-2}, i.e., like τ^{-4}, in the Minkowski frame (this variation is not to be confused with the Dirac form of cosmology in which G is taken to be variable even in the Einstein frame). The treatment of gravitational problems can therefore be quite awkward and unfamiliar when considered in the Minkowski frame. Yet this is no handicap in dealing with any known gravitational problem. If the solution is known already in the Einstein frame, any physical quantity appearing in the solution can immediately be transformed to the Minkowski frame. We simply note the dimensionality n of the quantity and use an additional time dependence τ^{2n} in the Minkowski frame.

As an example, the radius of a main-sequence star stays approximately constant in the Einstein frame, at any rate as long as the sky remains dark in the sense of Olbers' well-known paradox. Since "radius" has dimensionality $n = -1$, the radius of a main-sequence star behaves like τ^{-2} in the Minkowski frame. The radius of a galaxy behaves in the same way.

Suppose on the other side at time $|\tau| = \tau_0$ that galaxies had radii, stellar content, and spacings apart from each other that were similar to what we observe on our side at the present time τ_0. Now the spacings of galaxies stay fixed in the Minkowski frame, because galaxies behave cosmologically like the particles shown in figure 4. That is to say, galaxies have paths normal to the zero surface $\tau = 0$. Hence for $|\tau| < \tau_0$ the galaxies were larger in proportion to their spacings than they are at present. Indeed for $|\tau|$ given by

$$\left(\frac{\tau_0}{\tau}\right)^2 \gtrsim \left(\frac{\text{Spacing}}{\text{Radius}}\right)_{\text{Present day}} \tag{38}$$

the galaxies on the other side were overlapping each other. The right-hand side of equation (38) has an average value of about 300, so that galaxies on the other side were overlapping for $|\tau| \lesssim \tau_0/(300)^{1/2}$.

As the galaxies expanded with decreasing $|\tau|$, but before they overlapped, the ratio of the mean interstellar separation to the radii of the stars stayed constant. After overlap, however, the mean interstellar separation stayed fixed, whereas the radii of the stars continued to increase with decreasing $|\tau|$; which prompts the questions: Did the stars overlap also? If the stars were to continue to expand like τ^{-2}, they would certainly do so, at a time $\tau < 0$ which is easily shown to satisfy the equation

$$\left(\frac{\tau_0}{\tau}\right)^2 \simeq 300\left(\frac{\text{Stellar spacing}}{\text{Stellar radius}}\right)_{\text{Present day}}, \tag{39}$$

the ratio on the right-hand side of equation (39) being that which applies for stars in galaxies at the present time τ_0. Since this ratio is about 3×10^7, taken for a typical main-sequence star, equation (39) gives $(\tau_0/\tau)^2 \simeq 10^{10}$. At first sight then, we might expect stars to overlap in the same way as galaxies, but at a much smaller value of $|\tau|$. Yet this conclusion appears doubtful, *because the sky on the other side became bright in the sense of Olbers before this stage was reached.*

In the Einstein frame, the temperatures within main-sequence stars stay approximately constant. Since "temperature" has dimensionality $n = 1$, the temperature values in the Minkowski frame therefore behaved like τ^2, implying a value of only $\sim 10^{-3}$ ° K when $(\tau/\tau_0)^2 \simeq 10^{-10}$. At such a very low temperature, the pressure within a star would be much less than the pressure of the radiation field outside the star—we know the external radiation field after thermalization gave a temperature of $\sim 3°$ K. Because stars would be incapable of maintaining their expansion under such circumstances, the argument of the previous paragraph is incorrect. Expansion may be expected to have ceased when the internal temperatures inside stars fell to $\sim 3°$ K, which occurred for

$$(\tau_0/\tau)^2 \sim 10^7\,, \tag{40}$$

well before stellar overlap could take place. Combining equation (40) with the previous paragraph, we expect stellar expansion to have ceased when

$$\frac{\text{Mean interstellar spacing}}{\text{Mean stellar radius}} \approx 10^3\,. \tag{41}$$

While the stars would be quite close together, they would not be in immediate juxtaposition.

It may be useful to relate these considerations to our everyday concept of length, say to the centimeter. A unit of 1 cm has an operational meaning at the present day—it is a certain number of wavelengths in monochromatic radiation from a certain transition of ^{86}Kr. We can determine the ratio of 1 cm, defined in this practical way, to the mean spacing of the particles of figure 4. Having done this for our immediate neighborhood at the present day, using the local smoothed cosmological particle density, we then define "1 cm" to be the same ratio, always taken, whatever τ may be, with respect to the mean spacing of the smoothed distribution of the particles of figure 4. In the Minkowski conformal frame this latter spacing is constant, and so "1 cm" defined in this way is also constant in the Minkowski frame.

The galaxies on the other side expand in the Minkowski frame from radii of order 3×10^{22} cm at time $|\tau| = \tau_0$ to about 10^{25} cm at the time of their overlap. The mean interstellar spacing within galaxies likewise increased from about 3×10^{18} cm at $|\tau| = \tau_0$ to about 10^{21} cm at the moment when the galaxies overlapped. Thereafter, the mean interstellar spacing remained at $\sim 10^{21}$ cm. The main-sequence stars themselves increased in radius from $\sim 10^{11}$ cm at $|\tau| = \tau_0$ to $\sim 10^{18}$ cm at the stage where their internal temperatures fell to $\sim 3°$ K. The pressure of the external radiation field was then comparable to interior pressures, and expansion in the sense discussed above then ceased.

With $|\tau|$ continuing to decrease, the particles within the stars would take on a curious kind of equilibrium. With the internal temperature staying fixed and the radius fixed, the internal pressure would remain constant. The pressure of the external radiation also remained constant. So did the gravitational force on a particle. Hence it seems possible that stellar condensations may have persisted to $\tau = 0$. It is true that particles would have been evaporated from the extended stellar surfaces, particularly as long as the external radiation field remained unthermalized. The quanta impinging on the stars would be capable of endowing particles with higher and higher speeds as their masses continued to decline. Yet it seems doubtful that thermal evaporation could smooth the stellar condensations into a more or less uniform distribution of particles. At the stage where stellar expansion ceased, the stage determined by (40), the Thomson cross section had already increased by a factor $\sim 10^{14}$, due to the decline of the electron mass. Particles evaporated from the surface of a star would not be free therefore to move into the space outside, because of frequent deflections of their motions from the scattering of radiation.[2] Accordingly, it seems likely that matter reached $\tau = 0$ still with *substantial* fine-scale variations of density.

With equation (41) remaining approximately valid, even to $\tau = 0$, local particle densities would be greater than the smoothed density by a considerable factor, by a factor probably less than, but perhaps comparable with,

$$\left(\frac{\text{Stellar spacing}}{\text{Stellar radii}}\right)^3 \simeq 10^9 . \qquad (42)$$

Such islands of high density entering our side at $\tau > 0$ would very likely become stars again as the particle masses then increased with increasing τ. Hence it seems possible that many stars on our side are fossil relics of stars which existed formerly on the "other side."

IX. PROBLEMS OF CHEMICAL COMPOSITION

The possibility that matter may pass through $\tau = 0$ with considerable fine-scale variations of particle density raises interesting questions of chemical composition. Regions of high density resulting from stars crossing $\tau = 0$ from the "other side" would behave in their nuclear evolution rather like the local objects studied by Wagoner, Fowler, and Hoyle (1967), in what these authors called "the high temperature case." There is the difference of principle, however, that, whereas the local objects of Wagoner, Fowler, and Hoyle (WFH) determined their own nuclear time scale, here the time scale is that appropriate to the smoothed Einstein–de Sitter model. Fortunately, however, WFH studied a case having closely the appropriate time scale, with results shown in figure 6 of their paper. This case is directly applicable to the present discussion, and it is interesting that WFH found the resulting chemical composition to resemble that of Type II halo stars. These authors were particularly impressed at finding a fraction by mass of order 10^{-5} to be processed by neutron addition to give elements of atomic weight up to $A \sim 100$. This was the one case examined by WFH in which elements of high atomic weight, above the iron group, were produced in appreciable quantity.

The fraction of all material condensed into stars on the "other side" is of course not known. If most were indeed in the form of stars, then the chemical composition to emerge on our side would mainly be of the kind discussed briefly in the preceding paragraph, with the processed material largely locked away in a cloud of stars emerging on our side from $\tau = 0$. Some of these stars might form themselves into galactic aggregations—possibly the elliptical galaxies—but most of them may well be distributed everywhere throughout space. Yet some fraction of the material would be distributed at $\tau = 0$ with approximate spatial uniformity. Such smoothly distributed material would experience nuclear evolution, not like a uniform Einstein–de Sitter model ($q_0 = +\frac{1}{2}$), but like a model of smaller q_0. In the notation of WFH, this distributed material would behave in the manner of a model having parameter h less than the value $\sim 10^{-3}$ appropriate to the smoothly distributed Einstein–de Sitter case. It is well known that such smaller values of h can lead to $\mathrm{D/H} \simeq 10^{-5}$, the deuterium-to-hydrogen ratio now believed by many astronomers to exist everywhere through the gas clouds of our own Galaxy.

From what has just been said it is clear that the nuclear evolution of material emerging on our side from $\tau = 0$ would be likely to contain a large measure of variety, ranging from values of the h parameter of WFH less than 10^{-3} up to values of order 10^3, or even higher. Hence the nuclear species emerging from $\tau = 0$ may well have been much richer than has usually been appreciated.

In conclusion, enough has been said to show that we may owe many aspects of our present-day world to remote ancestors on the other side of the barrier which has hitherto been thought to represent the origin of the universe.

[2] Before thermalization, the number density of quanta in the external radiation field would be $\sim 10^{-2}$ cm^{-3} (defining the centimeter in the manner discussed above). The Thomson cross section, after an increase by a factor 10^{14}, was $\sim 10^{-10}$ cm^2. Hence the mean free path between successive scatterings by an electron cannot have been more than $\sim 10^{12}$ cm, very small compared with the interstellar spacing of $\sim 10^{21}$ cm. Any thermal outflow of particles from the stars evidently experienced a strong viscous drag from the external radiation field.

REFERENCES

Hoyle, F., and Narlikar, J. V. 1974, *Action-at-a-Distance in Physics and Cosmology* (San Francisco: Freeman).

Wagoner, R. V., Fowler, W. A., and Hoyle, F. 1967, *Ap. J.*, **148**, 3.

FRED HOYLE: Cockley Moor, Dockray, Penrith, Cumbria CA11 0LG, England