SLAC – PUB – 3376
July 1984
(T)

TCP, QUANTUM GRAVITY, THE COSMOLOGICAL CONSTANT AND ALL THAT ...*

T. BANKS

Stanford Linear Accelerator Center

Stanford University, Stanford, California, 94305

and

Institute for Advanced Study

Princeton, New Jersey 08540 †

ABSTRACT

We study cosmology from the point of view of quantum gravity. Some light is thrown on the nature of time, and it is suggested that the cosmological arrow of time is generated by a spontaneous breakdown of TCP. Conventional cosmological models in which quantum fields interact with a time dependent gravitational field are shown to describe an approximation to the quantum gravitational wave function which is valid in the long wavelength limit. Two problems with initial conditions are resolved in models in which a negative bare cosmological constant is cancelled by the classical excitation of a Bose field η with a very flat potential. These models can also give a natural explanation for the observed value of the cosmological constant.

Submitted to *Nuclear Physics B*

* Work supported by the Department of Energy, contract DE–AC03–76SF00515.

† Permanent address: Physics Department, Tel-Aviv University, Ramat-Aviv, ISRAEL

1. Introduction

On the unknown continent that is quantum gravity, the shadowy outlines of three major landmarks are barely discernible. The first is the nonrenormalizability of all unitary local quantum field theories of gravity.[#1] It seems probable that we will have turn to a theory of strings, to discrete space-time, or to some other, as yet unimagined, modification of current ideas in order to understand the physics of distances less than the Planck length.

The second landmark, somewhat more shrouded in mist than the first is the complex of questions raised by the seminal work of Hawking[1] on quantum field theory in the vicinity of a black hole. This includes the question of whether the topology of space-time (and not just its metric) is a dynamical variable, as well as the problem of the apparent loss of quantum coherence in quantum gravitational phenomena. The final resolution of these questions may well depend on the nature of short distance dynamics.

The third major topographical feature of quantum gravity is the question of whether it makes sense to say that "the entire universe" is in a pure quantum state. This question does not seem to depend on short distance physics. Indeed it arises even when we apply ordinary quantum field theory to cosmology.

We will not attempt to give an answer to this question here. We need a lot more practical experience with quantum gravity before we can hope to understand such a deep conceptual issue. Our aim in the present paper will be to try to place cosmology in the context of the formalism of quantum gravity as it is presently understood.

Contemporary cosmology is usually formulated in terms of quantum field theory in a time dependent classical gravitational field. The initial state of the field theory is taken to be a thermal density matrix. Presumably this is not supposed

---

#1 Despite the recent interest in conformally invariant theories of gravity[25] there is no real evidence that the apparent violations of the fundamental principles of quantum mechanics that appear in the perturbative solution of these theories can be circumvented.

to mean that the universe is not in a pure state, but is only an approximate description of a highly excited pure state. We will see that from the point of view of quantum gravity there are two puzzling aspects to these initial conditions:

1. One can show that in the region of gravitational field configuration space that describes long wavelength fluctuations in a universe of large volume, the semiclassical (WKB) approximation to the quantum gravity wave functional is valid. There are an infinite number of WKB wave functions, one for each classical solution of Einstein's equations. For any fixed WKB wave function the usual picture of quantum field theory in a classical background space-time is valid. However, any superposition of WKB wave functions also solves the gravitational Schrödinger equation (Wheeler-DeWitt equation). The correct superposition is determined by matching the WKB wave function to a solution which is valid in the small volume region where the WKB approximation breaks down. Indeed in this region we do not even believe that we know the correct Hamiltonian for quantum gravity (because of the problems with nonrenormalizability) Apparently we must make drastic assumptions about the nature of short distance physics in order to justify the usual procedure of taking a single classical solution.

2. The Wheeler-DeWitt equation seems to have many solutions. If we are discussing the wave function of the universe we must choose one of them (see Section 2) and throw the others away. In simple examples[2] there is always a preferred "simple" solution in which matter fields are minimally excited. This does not seem to correspond to the highly excited matter state which is the initial condition for all Hot Big Bang cosmologies (even new inflationary ones). The justification of conventional cosmology must again rely on special properties of unknown short distance physics.

In the present paper we will present a mechanism which resolves these two puzzles within the confines of presently understood physics. Somewhat surprisingly, it is the same mechanism that was previously introduced to explain the

currently observed magnitude of the cosmological constant in a natural way.[3] We will describe the outlines of a cosmological model based on this mechanism.

The model explains the choice of a unique classical solution by having the universe tunnel from a non-classical region of configuration space with negative cosmological constant, to a classical region where the effective cosmological constant has been made positive by the coherent excitation of a Bose field. Since the wave function in the tunneling region is concentrated along a one dimensional path in configuration space, the subsequent classical evolution has a unique initial condition, thus resolving puzzle #1. Furthermore the penetration into the classical region is correlated with the coherent excitation of a matter field η. It is the slow relaxation of η to its minimum which drives the processes which create the part of the universe that we see.

"After" (see Section 3 for an explanation of the quotes) the tunneling event the universe can be described as an exponentially expanding classical geometry plus a classical matter field whose relaxation slowly decreases the effective cosmological constant, coupled to a quantum field theory in a minimally excited state. This system can undergo a "curvature induced" first order phase transition of the type first envisaged by A.Sakharov.[♯2] The transition generates all the matter and entropy that we see around us; it is the origin of the Hot Big Bang.

The rest of this paper is an explanation of the above (somewhat cryptic) remarks. We begin in Section 2 by recalling the main points of the canonical quantum theory of gravity.[2,4] Section 3 describes the semiclassical approximation and the thorny problem of boundary conditions for the Wheeler-DeWitt-Schrödinger equation. The two puzzling discrepancies between quantum gravity and conventional cosmology are explained in Section 4 and the resolution of the first of them via tunneling is presented.

♯2 L. Susskind has informed me that the idea of curvature induced phase transitions originates with Sakharov. I have not however been able to find a published versions of Sakharov's work.

In Section 5 we describe the classical cosmological model that is suggested by our previous considerations. We show that it undergoes a curvature induced first order phase transition which produces a myriad of bubble universes, some of which resemble our own. We discuss the difficulties of this scenario. In section six we discuss several possible incarnations of the cosmic relaxation field η. Appendix A describes the details of the WKB approximation for gravity, while Appendix B is a statement of religious dogma about spacetime topology and related matters.

2. The Canonical Formalism for Quantum Gravity

General relativity is a theory of the dynamics of Riemannian spatial geometries. Its configuration space is a set of spatial metrics $g_{ij}(x)$, modulo time independent spatial coordinate transformations:

$$g_{ij}(x) \to \frac{\partial \bar{X}^k}{\partial X^i} \frac{\partial \bar{X}^\ell}{\partial X^j} g_{k\ell}\Big(\bar{X}(x)\Big) \equiv \bar{g}_{ij}(x) \ . \tag{1}$$

The Hilbert space for quantum gravity is the space of all coordinate invariant functionals of g_{ij}:

$$\psi[g_{ij}] = \psi[\bar{g}_{ij}] \tag{2}$$

with the coordinate invariant scalar product:[♯3]

$$\langle \psi | \phi \rangle = \int \prod_x \Big[\prod_{i \le j} dg_{ij}(x) / \det{}^2 g(x) \Big] \ \psi^*[g] \, \phi[g] \ . \tag{3}$$

The set of time independent spatial coordinate transformations is thus easily incorporated as a symmetry of the theory in the ordinary quantum mechanical

♯3 This scalar product is the most general one consistent with general covariance which does not involve derivatives of the metric. The solution of $2+1$ dimensional quantum gravity seems to indicate that corrections to the measure are necessary when we go beyond the semiclassical approximation.[6,26]

sense. Not so the rest of the coordinate transformation group of space time. For example, local time translations are generated by the Hamiltonian density $\mathcal{H}(x)$. General covariance then implies that the only acceptable states are those satisfying:

$$\mathcal{H}(x)|\psi\rangle = 0 \tag{4}$$

which is known as the Wheeler-DeWitt equation.[2] It turns out that (2) and (4) are sufficient to guarantee full general covariance if $\mathcal{H}$ and $P_n(x)$ (the generator of (1)) satisfy the Dirac-Schwinger commutator algebra[4].

Although Eq. (4) was derived in the framework of ordinary quantum mechanics, it implies a profound reevaluation of the meaning of the wave function and of our notion of time. It seems to imply that the wave function is static. How does this correspond to our experience?

In fact, (4) merely states in equations that the coordinate time that we have been using is a figment of our imagination on which no physical result can depend. Physical time measurements are correlations between two physical objects (system and clock) which necessarily interact at least gravitationally. They must both be included in the Hamiltonian density . Thus the familiar time dependent Schrödinger equation should appear as an equation for the correlations between two variables in the wave function, rather than for the dependence of the wave function on coordinate time. We should not expect the Schrödinger equation to take on its usual form unless the variable we choose as a clock is both weakly interacting and more or less classical. For example if we choose space to be infinite then the states of quantum gravity can be classified according to eigenvalues of a conserved asymptotic energy H (Arnowitt-Deser-Misner energy[4]). Time can be defined in terms of the unitary transformations $\exp(-iHt)$. Physically this corresponds to using as a clock a "test particle" which is heavy enough to be classical, but far enough away to be weakly interacting.

We will follow Einstein and Wheeler in assuming that the appropriate topology for space in discussions of "the Universe" is compact. Numerous attempts

have been made to find a quantum mechanical definition of time which would apply to compact topologies[5]. These attempts usually founder on the fact that the Wheeler-DeWitt equation is second order in all variables including the putative time. My conclusion (see also DeWitt[2]) is that time is a semiclassical concept which cannot be extended in any reasonable way into the domain in which quantum fluctuations of the gravitational field are important. In the next section we will see how the usual Schrödinger equation arises in the semiclassical approximation to the Wheeler-DeWitt equation.

The interpretation of physical time as a description of correlations in the WD wave function ψ points up a general property of ψ in quantum gravity. It is a relative probability amplitude, which describes only correlations between physical variables. This interpretation accords very well with the principle of general relativity, but it raises a technical question of some importance. What significance is there to the absolute norm of ψ, and indeed, must we require it to be normalizable? This question is bound up with the difficult problem of deciding what the correct boundary conditions for the Wheeler-DeWitt equation are. The question of boundary conditions takes on an importance here which it does not have in ordinary quantum mechanics problems. One wave function describes the entire history of the universe and everything in it. If there many solutions to the WD equation we must choose one of them and throw all the others away. We lose predictability if we do not find a compelling physical principle for deciding which solution of the equations to use.

The most attractive resolution of this problem (suggested to me by Ed Witten) is that the WD equation of the "True and Correct Theory of the World" has only one mathematically acceptable solution. The possibility of a unique solution is not totally unreasonable, although it does not seem to occur in ordinary quantum gravity. In supergravity $\mathcal{H}(x)$ is the square of a supercharge density $Q(x)$ and the WD equation becomes

$$Q_\alpha(x)|\psi\rangle = 0\ . \tag{5}$$

These are first order equations and certainly have many fewer solutions than the typical second order equation. If Witten's "one true theory" does not materialize, we will have to find a principle for choosing the right solution. Hawking and Hartle[2] (HH) have recently proposed such a principle.[7] They suggest that quantum gravity be described by an Euclidean functional integral and that the correct ψ is given by the integral over all compact Euclidean space-time geometries that have the spatial geometry described by g_{ij} as their only boundary. This integral can be shown to satisfy the WD equation by formal manipulations.[2] The HH boundary condition has a certain aesthetic appeal, but the meaning of Euclidean functional integrals in general relativity is somewhat obscure. I suspect that this issue, as well as the general problem of boundary conditions will only be resolved when we understand the correct short distance modifications of Einstein's theory. It is perfectly possible that the solution of the two problems referred to in the introduction also depends on these short distance boundary conditions. However, it would be more attractive to resolve them on the basis of physics that we understand now. To do this we have to understand the semiclassical approximation to the WD equation, which we present in the next section.

3. The Semiclassical Approximation

The semiclassical approximation to quantum gravity has been extensively discussed in the literature.[5] Semiclassical gravity in the presence of quantum matter fields (quantum field theory in curved space-time) has also received a lot of attention. But to my knowledge, the derivation of this formalism from the fully quantized WD equation was first discussed in Ref. 6. In particular, the time dependent four metrics of the curved space time formalism seem out of place in the static framework of the previous section.

In order to understand the semiclassical approximation, we will first consider the simplest non-trivial system invariant under time reparametrization. A general Lagrangian $L(q,\dot{q})$ can be made t reparametrization invariant by introducing a

1 metric $\alpha(t)$:

$$\int dt\, L(q,\dot{q}) \to \int dt\, \alpha(t)\, L\Big(q, \frac{\dot{q}}{\alpha}\Big)\ . \tag{6}$$

In the gauge $\alpha = 1$, the equations of motion take on their usual form but are supplemented by the constraint

$$p\dot{q} - L \equiv H = 0\ . \tag{7}$$

In quantum mechanics this becomes a constraint on states

$$H\,|\,\psi\rangle = 0 \tag{8}$$

whose interpretation in terms of correlations between system and clock is the same as that of Eq. (4). The simplest case is that in which both the system (q_1) and clock (q_2) have a single degree of freedom. Suppose the Hamiltonian is

$$H = \frac{P_2^2}{2M} + MV(q_2) + \frac{P_1^2}{2} + U(q_1,q_2)\ . \tag{9}$$

In general, the choice of the system to be called the clock is arbitrary. However when $M \gg 1, q_2$ behaves classically, and defines a natural time variable which conforms to our intuitive notion of time.

Let us try to solve Eq. (8) up to terms which are zeroth order in M as $M \to \infty$. If q_1 did not exist the solution for q_2 would have the familiar WKB form

$$\psi_{\rm WKB}(q_2) = V^{-1/4}(q_2)\, e^{\pm iM\int^{q_2}\sqrt{V(q)}\,dq}\ . \tag{10}$$

However, in the presence of q_1 this is modified. It is easy to verify that the correct

solution is

$$\psi(q_1, q_2) = \psi_{\rm WKB}(q_2)\,\chi\Big(q_1, t(q_2)\Big) \tag{11}$$

where

$$\pm\sqrt{V(q_2)}\,\frac{dt}{dq_2} = -1 \tag{12}$$

and

$$i\partial_t\,\chi(q_1, t) = \Big[\frac{P_1^2}{2} + U\Big(q_1, q_2(t)\Big)\Big]\,\chi(q_1, t)\ . \tag{13}$$

Note that Eq. (12) is equivalent to

$$\frac{1}{2}\left(\frac{dq_2}{dt}\right)^2 + V(q_2) = 0 \tag{14}$$

which means that $q_2(t)$ is a zero energy solution of the classical equations for the large part of the Hamiltonian. $t(q_2)$ is then the transit time for this solution: the time it takes to get from some(as yet unspecified) initial position to q_2. Equation (13) is the time dependent Schrödinger equation for q_1 in the "time dependent background field" $q_2(t)$. It is first order in time. The second order time derivatives of χ (which arise when both q_2 derivatives act on χ) contribute only at higher orders in $1/M$. Thus, within the region of configuration space where WKB is valid, and in which there is a zero energy classical solution, the problems that bedevilled previous attempts to define an intrinsic time disappear. There does not seem to be a reasonable notion of time outside of the WKB regime. Note that in regions where $V(q_2)$ is positive there are no solutions of (14) with $t(q)$ real. However there are solutions with imaginary time. These give a semiclassical description of tunneling.[7] The system variable q_1 is described by a Euclidean quantum mechanics in the tunneling region.

Physically the significance of the time dependent Schrödinger equation is that it predicts the same correlations between q_1 and q_2 that we obtain from the WD wave function . If we ask "What is the WD wave function for q_1 when q_2 takes

on the value q?", then up to an irrelevant (q_2 dependent) factor we get the same wave function predicted by the time dependent Schrödinger equation at the time that $q_2(t) = q$. Note that the time $t(q)$ plays only a cosmetic role here. It was introduced to make the Schrödinger equation look familiar. Any other function of q would have predicted the same correlations.

The treatment of general relativity is so similar to the above that I will relegate the technical details to Appendix B. The WD equation is[4] (we ignore ordering problems)

$$-\frac{\gamma_{ijk\ell}}{M^2}\,\frac{\delta^2}{\delta g_{ij}(x)\,\delta g_{k\ell}(x)} - M^2\sqrt{g(x)}\,(R^{(3)} - \Lambda) + \mathcal{H}_m\Big(\phi, g, \frac{\delta}{\delta\phi}\Big) \qquad (15)$$

where M is the Planck mass and $\mathcal{H}_m(x)$ is the Hamiltonian density describing a generic matter field ϕ and its interaction with gravity. To zeroth order as $M \to \infty$ the solution of (15) is

$$\sqrt{\det\frac{\delta^2 S}{\delta g_{ij}\,\delta\bar{g}_{k\ell}}}\;e^{iM^2 S(g,\bar{g})}\,\chi_n\Big[\phi(x), \tau(x;g)\Big] = \psi_{\bar{g},n}(g,\phi)\;. \qquad (16)$$

Here S satisfies the Einstein Hamilton Jacobi equation[8]

$$\gamma_{ijk\ell}\,\frac{\delta S}{\delta g_{ij}}\,\frac{\delta S}{\delta g_{k\ell}} = \sqrt{g}\,(R^{(3)} - \Lambda) \qquad (17)$$

and χ_n satisfies the Tomonaga-Schwinger equation

$$i\,\frac{\delta\chi_n}{\delta\tau(x)} = \mathcal{H}_m(x)\,\chi_n \qquad (18)$$

which reduces to the time dependent Schrödinger equation if we consider only global time variations. The action of any solution of Einstein's equations which contains the three geometry g_{ij} as a spacelike slice, is a solution of (17). Dimensional arguments suggest that this WKB approximation should be valid when

we consider long wavelength gravitational fields in a large volume universe. Note however that the classical field which is used in the WKB approximation is a solution of the vacuum Einstein equations. If matter fields make large contributions to the energy density then the approximation must be refined.

The above discussion resolves many of my own conceptual confusions about the application of quantum mechanics to cosmology, and I hope it will be equally helpful to others. In particular, it shows how the usual picture of field theory in a time dependent classical vacuum is compatible with the idea that the universe is in a stationary state. It also gives some sort of an answer to the question of what happened before the Big Bang. At very early times, the size of the universe is very small and quantum gravitational fluctuations become important. We do not really know the correct theory for describing these short distances but there is really no reason to expect that it is ill-defined. The above discussion shows that as we enter this regime, the intuitive concept of time loses all meaning. Thus there is no content in the question of what happened before the Big Bang, not because the universe becomes singular, but because quantum fluctuations invalidate the notion of "Absolute Time ..."[9]. It is very hard to think intuitively about the quantum regime of general relativity, but the mathematical formalism need not break down. Indeed in low dimensions, where there are no problems with renormalization, one can solve the WD equation in the small volume regime even when WKB breaks down.[6] The solution matches smoothly onto the semiclassical picture, although the words that we use to describe the semiclassical regime become inadequate.

4. Boundary Conditions for the Wheeler DeWitt Equation

We now come to one of the most confusing questions of the whole subject of quantum cosmology-boundary conditions. There are several aspects to this — first we must make a choice of configuration space, then we must choose a Hilbert space of functions on configuration space. Our configuration space is the set of

all spatial geometries (and of fields defined on these geometries), but we have not yet specified the topology of our manifold. Is it open or closed, connected or disconnected, with or without handles. Can the topology of space change with time? There are really no good answers to these questions at present, choosing between alternatives is basically a religious act. We will try to argue in Appendix B that the answers to topological questions depend on details of short distance physics which we do not know. Here we will simply choose a religion and hope for the best. We take space to be compact and connected, with the topology of a 3 sphere or torus, and we do not allow changes in topology.

The choice of Hilbert space is essentially the question of which solutions of the WD equations we should use. The ambiguity is somewhat reduced by mathematical considerations: we should only choose solutions from within the domain in which the WD operator is Hermitian. In low dimensions, where the theory is more or less well defined, this is not sufficient to single out a unique solution. It should be remembered that our wave function describes the history of the universe, changing boundary conditions changes everything. So it is very important to find a principle for choosing the right solution.

In the semiclassical approximation, the collection of solutions of the WD equation is parameterized in the following way: We first choose a solution of the vacuum Einstein equations. There is a unique solution connecting a pair of three geometries with a given time orientation. Given an initial condition $\bar{g}_{ij}$ we must then choose an initial state for the time dependent Schrödinger Eq. (18). In general then, the solution of the WD equation in the semiclassical regime has the form

$$\int d\bar{g} \sum_n C_n[\bar{g}]\, \psi_{\bar{g},n}(g,\phi)\ . \tag{19}$$

The coefficients $C_n[\bar{g}]$ are determined by matching the WKB solution to a solution of the WD equation valid in the small volume region. They thus depend on all the details of short distance physics, of which we are ignorant.

Unless the short distance dynamics is such that the coefficients $C_n[\bar{g}]$ are peaked around a particular initial condition $\bar{g}^*$, the wave function (19) will not have a simple semiclassical interpretation. In ordinary quantum mechanics we can prepare a system in a classical coherent state. Then if the WKB approximation is valid in some region of configuration space the wave function will propagate classically in this region. However there are many multidimensional quantum systems whose eigenstates obey the WKB equations in some region of configuration space, but are linear superpositions of WKB wave functions corresponding to different classical trajectories. Their behavior is not at all classical.

Clearly some solution of this problem must be found if we are to understand how to derive conventional cosmology from quantum gravity. A possible way to do this is to insist that the semiclassical dynamics lead asymptotically to states that are independent of the initial conditions. For example (and this is the only example I know) if the cosmological constant is positive then at asymptotically large times, expanding solutions of Einstein's equations probably all approach the exponentially expanding branch of DeSitter space.[10][♯4] Furthermore any state of a quantum field theory in a DeSitter background probably evolves into the unique causal, DeSitter invariant state of the field theory.[♯5]

The problem with this idea is that it is very hard to get rid of the cosmological constant once we have put it in. In standard inflationary cosmologies the cosmological constant is a transient phenomenon, a consequence of the universe's temporary sojourn in a metastable state. The state is stabilized by the thermal

♯4 There is a "no-hair" conjecture for Desitter space which states that perturbations of it are asymptotically damped at large times. If one starts with generic initial conditions, then local regions cay undergo gravitational collapse before inflation dilutes the matter density. However, even such metrics will look like DeSitter space almost everywhere (i.e. except at the points of collapse) at sufficiently late times.

♯5 The conjecture that generic states evolve into the DeSitter invariant state is due to Ed Witten. It may be verified in free field theory. However it is only true locally. That is, only operators which correspond to measurements concentrated in a region of finite proper volume will have expectation values which are indistinguishable at late times from those in the DeSitter invariant state. The point is that differences between the initial state and the DeSitter state are not washed out; they are simply inflated to sizes which grow with the scale factor.

fluctuations of a highly excited gas of massless particles which must exist prior to the "vacuum dominated" phase.

In the present context we must ask whether the above scenario describes a plausible solution to the WD equation. Certainly there are pure quantum states which accurately mimic the relevant effects of a thermal bath. However, the question of whether such states are the correct initial conditions for Eq. (18) again leads back to the matching conditions for the coefficients $C_n[g]$. So the justification for standard inflationary scenarios (or any other model in which the hot big bang extends back to the edge of the semiclassical region) depends on details of the matching to the short distance wave function. In minisuperspace models[2] and in low dimensions[6] the natural boundary conditions at short distance pick out a matter wave function which is minimally excited (the DeSitter invariant state for inflationary backgrounds), rather than the sort of highly excited state which could mimic a hot big bang. The requirement of thermal initial conditions can be regarded as a (rather stringent *a priori*) constraint on the short distance dynamics, but then we have not really solved the problem of the $C_n[g]$. We might just as well have assumed that the short distance dynamics fixed $C_n[\bar{g}]$ to be peaked around a particular initial condition. Note that if the positive cosmological constant is not a thermal effect, but a term in the Lagrangian then we get a classically evolving universe, but one that becomes virtually empty after a few e foldings.

Our arguments for rejecting the standard Hot Big Bang scenario are far from airtight. Nonetheless, there seems to be a definite difficulty in making a hot classical universe compatible with quantum gravity without making assumptions about physics in regions which we do not presently understand. We will now present a model which resolves these questions within the domain in which semiclassical physics is valid. Our basic assumption is that the cosmological constant $-\Lambda(\Lambda > 0)$ in the Einstein Lagrangian is large $(O(M^2))$[♯6] and negative. It is

♯6 Of course, in a theory with global supersymmetry, it may be natural to have a value of Λ which is much less than M^2.

worth noting that if Λ is this large then we do not have to worry about renormalization effects changing its value. In particular we do not have to discriminate between the bare Λ (renormalized at scale M) and the renormalized Λ at some low energy scale. They are of the same order of magnitude.

In pure quantum gravity, the introduction of a large negative cosmological constant is disastrous. The piece of the WD operator which depends only on the volume v of the spatial geometry is

$$\mathcal{H}_v = \frac{1}{M^2} v \frac{\partial^2}{\partial v^2} - M^2 \Lambda v \ . \tag{20}$$

While a solution of $\mathcal{H}_v = 0$ is not a solution of the full WD equation, it is the first term in the formal asymptotic solution of WD in powers of $1/v$. The only solution of $\mathcal{H} = 0$ which allows integration by parts so that the WD operator is Hermitian is

$$\psi = e^{-M^2\sqrt{\Lambda}v} \ . \tag{21}$$

It is exponentially damped in the large v region and does not have the oscillatory behavior of the WKB wave function for a real classical motion. This is not surprising. Einstein's equations with no matter and a negative cosmological constant have no solutions in which the spatial geometry is compact. The WD equation has a solution but it cannot be described in classical terms.

The situation is improved if we add matter fields. Let us add a scalar field with potential $U(\eta)$. We choose U to have a unique minimum at $\eta = 0$ and to satisfy $U(0) = 0$ so that $-\Lambda$ is the full cosmological constant of the model. Keeping only v and the spatially constant mode of η we get a Wheeler-DeWitt operator:

$$\frac{1}{M^2} v \frac{\partial^2}{\partial v^2} - M^2 \Lambda v + U(\eta) v - \frac{1}{2v} \frac{\partial^2}{\partial \eta^2} \simeq \mathcal{H} \ . \tag{22}$$

If $U(\eta) > \Lambda$ for some values of η then there is a region of configuration space where classical solutions exist. The wave function is concentrated near $\eta = v = 0$, but

there is some amplitude to tunnel out to the classically allowed region. Since we can only observe regions where v is large and the universe is behaving classically, and since the wave function only describes correlations between variables, we need not be concerned that the bulk of the wave function is concentrated near the origin. We care only that the tunneling amplitude to a large v, classical region is non-zero. We learn that large v classical regions are correlated with the displacement of η from its minimum into a region where $U(\eta) > \Lambda M^2$.

This correlation between large volume and the excitation of a field away from its minimum is extremely important. We will see in the next section that the classical relaxation of η might drive the processes which create our universe. Equally important is that the "starting point" for the classical evolution is a tunneling process. To see the significance of this fact let us recall some features of multidimensional tunneling processes. If the barrier to be tunneled through is high and wide enough, tunneling can always be described in the WKB approximation. The wave function in the tunneling region is concentrated along a one dimensional path in configuration space called the most probable escape path [MPEP][7]. The MPEP can be parametrized in such a way that it is a solution of the Euclidean classical equations of motion. (after which it is usually called an instanton). It pierces the barrier and penetrates into the classically allowed region at a particular point in configuration space $\vec{q}_0$. If the WKB approximation remains valid in the classically allowed region, the wave function in this region is a WKB function based on the classical solution whose initial position is $\vec{q}_0$ and initial velocity zero.[♯7] The higher order corrections to the WKB wave function are dependent on the nature of the state inside the barrier (they are determined by matching), as are the parameters of the MPEP, but the fact that a unique initial condition is picked out is not. Thus tunneling solves the problem of the unknown coefficients $C_n[g]$. The "tunnelled" wave function is concentrated around

♯7 Actually[7] the tunneling wave function becomes less concentrated as the instanton nears the point of penetration into the classical region. Consequently, there will still be a small spread in the distribution of initial values for q_0 and $\dot{q}_0$.

a particular classical initial condition. Whatever the nature of the wave function at small v, a universe which tunnels from a regime with negative cosmological constant will be described by a single classical solution. The initial conditions for the time dependent Schrödinger Eq. (18) will depend on the state at small v, as will the initial configuration of the gravitational field. This is not a source of worry. The classical equations on the large v side of the barrier have a large positive value of the effective cosmological constant. If η has a flat enough potential (see Section 5) this situation will continue for a long time, and the exponential expansion of the universe will inflate all traces of the initial conditions to a size much bigger than the horizon scale.

This section was entitled boundary conditions. We have spent most of it trying to avoid talking about the boundary conditions at all. To be complete we will discuss the one attempt that has been made to fix the WD boundary conditions: the ansatz of Hawking and Hartle.

HH propose that the correct solution of the WD equation is given by a Euclidean path integral over compact spacetime metrics and matter fields which are regular on the compact space. The boundary of the compact manifold is identified with the spatial geometry which appears in the WD wave function. In minisuperspace models in which all but a single mode of the gravitational field are suppressed, there is generally a special solution of the equations in which the matter fields are in a minimally excited state. HH show that in this case (and in the semiclassical approximation) their path integral definition gives the minimally excited state. I believe that there is something correct about this ansatz but there are several points that I find obscure. Firstly, due to the indefiniteness of the Euclidean Einstein action, it is not clear whether the functional integral is well defined outside of the semiclassical approximation. Numerous inconclusive discussions of this have appeared in the literature and I have nothing to add. Secondly, it is clear that the HH ansatz is saying something about the small volume region. Their wave function is the amplitude to find a given spatial geometry, starting from an initial state with zero volume. It seems a bit premature to make

an ansatz about the zero volume behavior of the wave function when we do not yet have the correct Hamiltonian for the system in this region. The final confusing point about the HH wave function is related to time reversal invariance. The WKB solution related to a particular classical motion with action S is $\exp(iS)$, while that for the time reversed motion is $\exp(-iS)$. The HH wave function is real and in the classical regime it looks like the sum of these two. HH interpret this as representing the two branches of a classical cosmological model in which contraction is followed by expansion (e.g. DeSitter space). This interpretation seems rather problematical when applied to the real world. Even if our present universe does recollapse, we believe that a lot of entropy was created during its expansion. The collapsing phase should not at all look like the time reverse of the expansion. For example if the universe underwent a first order phase transition during the course of expansion, we expect to experience the other branch of a hysteresis loop when it recontracts.

An alternative idea of what to do with time reversed solutions was suggested in Ref. (2). One should consider only the wave functional corresponding to the forward motion and say that TCP is spontaneously broken. Physics described by the reversed wave function is identical to that described by the forward motion but there are no correlations between the two. As is usual when a symmetry is spontaneously broken, one can use the symmetric superposition of the two time reversed wave functions if one is sufficiently careful about the operators one computes the expectation value of. Thus there is nothing wrong with the HH prescription, but it may lead to confusion if we try to study the wrong quantities. For example, taken literally it would imply that the wave function vanishes at certain values of the volume in the classical regime. Of course, it is possible that wave functions that have a pure $\exp(iS)$ behavior in the classical regime are not valid solutions of the WD equation. Examples of this behavior are found both in $1+1$ dimensions when gravity is coupled to matter[2] and in $2+1$ dimensions for spaces with spherical topology and a positive cosmological constant.[6] They seem to occur when the classical equations predict a time symmetric bounce

cosmology with contraction followed by expansion. One of the quantum wave functions decreases as it penetrates the classically forbidden region with small volumes while the other increases. The WD operator is not Hermitian when applied to the increasing wave function because one is not able to integrate by parts at $v = 0$. The good wave function is real in the classically allowed region. It is not clear whether one can make a sensible interpretation of this superposition of states describing two different parts of the classical motion.

One should note that in the real world one may not be able to throw out the "bad" wave function. Its pathology occurs at small volumes where we expect the theory to be cut-off, and the true boundary conditions at small volume may allow a pure $\exp(iS)$ behavior in the classical region. It is clear however that it is much easier to interpret the quantum wave function when the classical motion is time asymmetric.

To summarize this section we can say that although short distance boundary conditions can profoundly affect the question of whether quantum gravity is compatible with conventional cosmology, it is possible to find a set of models in which one can derive semiclassical cosmology without making special assumptions about the short distance behavior of the theory. These models all have a negative "bare" cosmological constant and a Bose matter field with an extremely flat potential. The universe tunnels out of the nonclassical region where most of its wave function is concentrated, into a classically allowed region where the Bose field is coherently excited so as to produce a positive value for the effective cosmological constant. The semiclassical history of the universe is driven by the slow classical relaxation of the Bose field to its minimum. In the next section we will see whether models of this sort can produce cosmological scenarios which agree with what we know of cosmic history.

5. The Classical Regime

The scalar field that we introduced in Section 4 will be called the isichon

$(\eta\sigma\chi_0\nu)$ because it relaxes the cosmological constant. We will assume that its Lagrangian is such that it can be treated classically after it crosses the barrier. Thus we must solve the classical equations for η coupled to the gravitational field.

The initial conditions for these equations are obtained by solving the tunneling problem described in the last section. As usual[7] $\dot{\eta} = 0$ and most of the $\dot{g}_{ij}$ are determined from g_{ij} and η by the constraint equations

$$\mathcal{H} = P_m = 0 \,. \tag{23}$$

Since the potential for η is very flat and $U(n) > \Lambda$, we know that the homogeneous mode of η is very large. The non-homogeneous components are probably much smaller as may be seen by the following argument. The tunneling process is controlled by the minimum action solution of the Euclidean field equations. At any Euclidean time t we can expand η in eigen modes of the Laplacian of the instantaneous spatial geometry. Since all these geometries have volumes of order M^{-3} the nonhomogeneous modes contribute quadratic terms to the action which will be large if $\eta_0 > M$. Thus for nonzero n η_n is bounded by $\cdot M$. Similar arguments indicate that there is no reason for the initial values for g_{ij} and $\dot{g}_{ij}$ to be larger than simple dimensional estimates. On the other hand η_0 must be much larger than M in order to have $U(\eta) > \Lambda M^2$. Finally, the initial value for $U(\eta) - \Lambda M^2$ should also be given by dimensional considerations. It is a positive number of order ΛM^2.

The classical evolution of the universe thus begins with a large positive value for the effective cosmological constant. Since U is very flat and $\dot{\eta}$ is zero initially, the cosmological constant will remain large for a long time and most aspects of the initial values of η and g will be inflated away. We can approximate the metric by a closed Robertson-Walker geometry, and η by its homogeneous mode. (This is not really a crucial assumption at present. Inhomogeneities would not significantly alter the qualitative conclusions that we will come to.) In units in

which the Planck mass is one the classical equations are:

$$\left(\frac{\dot{R}}{R}\right)^2 + \frac{k}{R^2} = \frac{\dot{\eta}^2}{2} + U(\eta) - \Lambda \tag{24}$$

$$\ddot{\eta} + 2H\dot{\eta} = -\frac{\partial U}{\partial \eta} \qquad \left(H \equiv \frac{\dot{R}}{R}\right) . \tag{25}$$

Initially H is positive and $0(\sqrt{\Lambda})$.

As the universe expands the energy in the η field red shifts to zero and H decreases. However if U is very flat the decrease is very slow. To see this compute:

$$\dot{H} = -\frac{3}{2}\dot{\eta}^2 , \tag{26}$$

(we have taken $k = 0$ for simplicity). $\dot{\eta}$ starts at zero; even if it did not, the large frictional term in (25) would soon bring it down to a value for which the frictional force $3H\dot{\eta}$ is of the order of magnitude of the restoring force $\partial U/\partial \eta$. To see what sort of a potential is needed we note that if $U(\eta) = \frac{1}{2}\mu^2\eta^2$. then we must have $\mu^2\eta^2 \sim \Lambda$ to cancel the negative cosmological constant The restoring force is $-\mu^2\eta$ so we find $\dot{\eta} \sim \mu\sqrt{\Lambda}/3H$. Thus $\dot{H}$ will be small for sufficiently small μ.

A flat potential will thus lead to a long period of inflation with a slowly decreasing cosmological constant. Eventually H will reach zero but unfortunately it does not stop there. $\dot{\eta}$ will be nonvanishing when $H = 0$ unless the initial conditions are fine tuned so that $\dot{\eta}$ is zero there. Even if this is done the third derivative of H will be $|\partial U/\partial \eta|^2$. Since the point $U = \Lambda$ is not a minimum of the potential (to insist that it be so is the usual fine tuning), H will pass through zero to negative values and the universe will collapse.

Clearly the potential must be flat enough for at least the entire "known" history of the universe to fit into the time interval in which the cosmological constant is small but H has not yet reached zero. To see what this entails let

us again consider the case $U = \frac{1}{2}\mu^2\eta^2$, $k = 0$ and add a nonrelativistic matter density ρ to Einstein's equations.

$$H^2 \equiv \left(\frac{\dot{R}}{R}\right)^2 = \frac{\dot{\eta}^2}{2} + \frac{\mu^2\eta^2}{2} + \rho = \Lambda \tag{27}$$

$$\ddot{\eta} + 3H\eta = -\mu^2\eta \tag{28}$$

$$\dot{\rho} = -3H\rho \ . \tag{29}$$

We are interested now in the period of universal history described by classical cosmology.[11] Thus $\rho \ll \Lambda$ as is H. Define new variables by

$$\mu\eta = \sqrt{2(H^2 + \Lambda - \rho)}\,\cos\theta/2 \tag{30}$$

$$\dot{\eta} - \sqrt{2(H^2 + \Lambda - \rho)}\,\sin\theta/2 \ . \tag{31}$$

Then

$$\dot{H} = -3\sin^2\theta/2\ (H^2 + \Lambda - \rho) - \frac{3}{2}\rho \tag{32}$$

$$\dot{\theta} = \mu - 3H\sin\theta \ . \tag{33}$$

In order for these equations to resemble the Friedmann cosmologies, θ must be small throughout the period under discussion (10-20 billion years). We can therefore approximate Eq. (33) by

$$\dot{\theta} = \mu - 3H\theta \tag{34}$$

whose solution is

$$[\theta(t) - \theta(t_0)] = \mu \int_{t_0}^{t} dx\, e^{-3\int_s^t H(u)du} \ . \tag{35}$$

A bound on θ can now be obtained by replacing $H(t)$ by its value at the beginning of the evolution (e.g. at a time when the temperature of the universe was 0(1

GeV)). This gives

$$\Delta\theta \geq \frac{\mu}{3H}\left[1 - e^{-3H(t_0)(t-t_0)}\right]. \tag{36}$$

Since $H(t_0)(t - t_0)$ is 0(1), we find that we need

$$\mu^2 < H^2(t_0)\rho \sim 10^{-240} \tag{37}$$

in order to retain the Friedmann form of the Eq. (32). This would appear to be more fine tuning than is necessary to simply cancel the cosmological constant by hand. However, we will describe several scenarios in the next section in which the smallness of μ is at least technically natural.

Now that we know what values of the parameters are necessary in order to be at least compatible with the last ten billion years or so of history, we must go back to much earlier times and try to come up with a scenario which can naturally explain the initial conditions at the time that the temperature of the universe was 100 MeV. This of course means that we must resolve the usual puzzles about flatness, horizons, etc., but first we must do something much more basic than that. We must explain where all the matter in the universe came from. In our model the universe goes through a long period of exponential expansion with a Planck scale cosmological constant. Whatever the initial state of the matter, it will quickly settle down into a state which locally resembles the DeSitter invariant state of whatever quantum field theory describes it.[45] The effective cosmological constant changes much more slowly than the interaction time scales of the quantum field theory, so once settling into its DeSitter invariant state, the theory will move adiabatically from the DeSitter vacuum with one value of Λ to the vacuum with another. Unless some violent paroxysm intervenes, it will end up in the Minkowski vacuum when the cosmological constant gets to zero. It is clear that we need some sort of a first order phase transition to generate the required energy. At first sight however, this does not seem to be enough. We must also guarantee that the cosmological constant at the time the transition

occurs be small. Otherwise inflation will dilute the energy density to zero long before the cosmological constant reaches zero.

Surprisingly, a model can be constructed in which this apparent coincidence of the time at which the transition occurs and the time when $\Lambda_{\rm eff}$ reaches 0 does not require finely tuned initial conditions. Our model will consist of the η field and another scalar field ϕ which represents all the rest of the matter in the universe. ϕ could be a Higgs field in a grand unified theory. We will assume that ϕ has a potential of the form shown in Figure 1. The long flat stretch between the two local minima of the potential is as unnatural as it is in new inflationary universe scenarios. Indeed we will assume that all of the usual fine tunings which are necessary to make new inflation work have been done here. We will resolve the problem of the cosmological constant but none of the other problems of cosmology. This is clearly a defect of the present model, but perhaps we may be excused for doing one thing at a time.

The crucial ingredient in our model will be the quantum corrections to the effective potential of the ϕ field in curved spacetime. Many papers have been written on this subject[12] The largest single correction is the logarithmically divergent renormalization of the $R\phi^2$ term in the potential. (R is the scalar curvature.) We will take only this large effect into account, and assume that it comes in with the sign necessary to stabilize the minimum at the origin (if the flat space potential has a very flat maximum near the origin then the $R\phi^2$ term will produce a minimum for sufficiently large R). At early times, when the cosmological constant is large, the $R\phi^2$ term dominates the potential and the origin is the global minimum. As η slowly relaxes, the effective cosmological constant will decrease and the origin becomes metastable. We will witness a curvature induced first order phase transition.

Curvature induced phase transitions in cosmology were apparently first discussed by A. Sakharov.[42] They allow one to start from a cold universe and generate entropy. However, in conventional models where all fields are in their ground

states at zero temperature, there is no way to change the space time curvature. The novelty of our approach is the η field whose classical relaxation allows us to vary the curvature even at zero temperature.

We assume that the parameters of the potential are arranged so that the kinetics of the phase transition are described by bubble formation. Basically this means that the barrier between the metastable and true vacua is high and wide enough that the minimum action of paths crossing the barrier is large. Bubbles will begin to be nucleated when $\sqrt{\Lambda}$ is smaller than the scale which characterizes the structure in the flat space potential for ϕ. Presumably this is the GUT scale ($\sqrt{MM_G}$) in a realistic model. This is a rather large value of the cosmological constant. Thus the first bubbles to be formed will inflate, then "fall over the edge", and then inflate more slowly. Any matter and radiation that is produced in falling over the edge, will quickly redshift to zero. These bubbles will be large enough to be our universe, but they will be empty. However, all is not lost. As first shown by Guth and Weinberg[13], bubbles of true vacuum in an inflating background, generally do not percolate. The background, with larger cosmological constant, expands too fast for the bubble growth to catch up. This means that the phase transition does not end with the formation of the first bubbles. The false vacuum keeps on nucleating new bubbles of true vacuum which float off into the blue. While this is going on the cosmological constant (the η field energy) continues to decrease. Inevitably, there will come a time when bubbles are nucleated with an effective cosmological constant which is close to zero. If the authors of Ref. (14) have built their models carefully, these bubbles will resemble our universe.

We have therefore found a class of models which will produce a myriad of bubble universes, some of which resemble our own. In particular we easily find bubbles with small or zero cosmological constant. However, it may be argued that this is insufficient, and that we must find a model in which the typical bubble resembles our own universe. To even ask this question we must first decide what we mean by typical. Do we count all bubbles, or only those in which intelligent

life could develop? It is this point that the infamous anthropic principle enters our considerations. I would contend that in the present context, the use of the anthropic principle is not quite as ridiculous as (for example) it is in attempts to explain the value of the fine structure constant. Our model incorporates a physical mechanism which really generates a large number of "sample universes". It makes sense to ignore those in which intelligent life could not develop. This principle definitely restricts the values of the effective cosmological constant which are allowed. A universe with a large positive value of the cosmological constant quickly becomes empty. Non-relativistic matter and radiation energy densities are diluted to zero before any interesting structures can form. On the other hand a bubble that is nucleated with a large negative cosmological constant will collapse on itself in a short period of time. Note that the Coleman-DeLuccia mechanism[15] which can prevent the nucleation of bubbles with large enough negative values of $\Lambda_{\rm eff}$ is not relevant here. The bubble is nucleated on the plateau region of Figure 1, so that the cosmological constant which goes into the bubble nucleation calculation has nothing to do with the true cosmological constant at the bottom of the well. However, bubbles nucleated with negative Λ will live only as long as the radiation energy created in reheating is larger in absolute magnitude than Λ.

The question of the bounds on Λ that come from requiring the existence of intelligent life is a difficult problem in the interaction of physics with biology. It touches on all sorts of questions which evolutionary biology simply has no answers to. I believe however that it is conservative to state that a universe with intelligent life is inconceivable unless the absolute value of $\Lambda M_{p\ell}^2$ is less than ten times the critical density:

$$\rho_c = H_{\rm NOW}^2 \qquad \text{(in units where } M_{pl} = 1) \ .$$

If the probability density for Λ (number of bubbles with cosmological constant Λ/total number of bubbles) is fairly flat in the region between + and $-10\ \rho_c$,

then the typical livable bubble is reasonably similar to our own. Traditionally, one said that the observed cosmological constant was zero, but recently values of order 0.7 or 0.8 , have been invoked to explain the discrepancy between estimates of clustered dark matter and the matter density necessary for consistency with inflationary cosmologies.[16] This would be a reasonably typical value for a flat distribution concentrated between $\pm 10\ \rho_c$.

Unfortunately the probability distribution for Λ in our model is far from flat. The tunneling amplitude for nucleating bubbles is essentially independent of Λ in the range of interest. However this is a tunneling amplitude per unit proper volume, and the volume of false vacuum is exponentially expanding as the cosmological constant changes. Since the cosmological constant does not change on time scales as long as the age of our universe, and the e folding time for the false vacuum is determined by the GUT scale (say $10^{15}\ GeV$). the probability distribution for Λ grows extremely rapidly as Λ decreases. The number of bubbles with $\Lambda = 0.8\ \rho_c$ is smaller than the number with $\Lambda = -10\,\rho_c$ by a factor greater than $\exp(10^{100})$.

This is a deathblow for the present model. Even the anthropic principle cannot save it. It predicts that the typical bubble with intelligent life in it is one that has a negative cosmological constant which is on the verge of causing catastrophic gravitational collapse. This is clearly not a description of our universe, which if anything has a positive Λ. Bubbles of our type exist but only in miniscule numbers (the probability of catching one in a random sampling of bubbles is much smaller than the probability that all the air in this room will gather itself in a corner of the ceiling).

The reason that I have presented this model despite its glaring fault is that I believe that most of its structure is sound. Models with a varying cosmological constant can explain the presently observed value in a natural way and they can explain the generation of heat in a universe which is initially in its ground state. It is now necessary to find a model which will give a reasonably flat probability

distribution for bubbles as a function of their cosmological constant.[#8]

6. Models for the η Field

The flat potential required for the η field must seem extremely unnatural to many readers. Numerically, the fine tuning required to obtain such a potential in a generic field theory model is much worse than that required to simply cancel the cosmological constant by hand. However, the cosmological constant problem is not the worst of the hierarchy problems of physics because of the magnitude of the fine tuning it requires. It is a disturbing problem because it requires fine tuning in a very low energy effective Lagrangian, which describes a regime in which we think we know the physics quite well. To see this consider a generic field theory and integrate out (in the sense of Wilson and Kadanoff) degrees of freedom with momenta greater than 1 MeV. According to Wilson's renormalization group philosophy, if the underlying theory was local, then physics below 1 MeV is described to a very good approximation by a local effective Lagrangian containing only light degrees of freedom (for renormalizable local field theories this claim is the content of the Symanzik-Appelquist-Carrazone theorem.)[17] The effective Lagrangian will contain photons, electrons, gravitons, neutrini, axions (?) and any other light particles we may invent. Now suppose that some magical symmetry in the underlying theory has enabled us to prove that the cosmological constant that appears in the effective Lagrangian is exactly zero. We now proceed to integrate out degrees of freedom with momenta between 1 keV and 1 MeV. At this point we believe that we know the true Lagrangian and we can actually perform the calculation in perturbation theory. The answer is $0(1\ \mathrm{MeV}^4)$ which is too large by 36 orders of magnitude.

There are only two ways out of this dilemma. Either we can assume that the 1 MeV effective Lagrangian had a non-zero cosmological constant which exactly

[#8] Susskind has suggested a variant of the present model which has a flat probability distribution for Λ. However a proper analysis of this idea requires extensions of the approximations discussed in this paper.

cancels the results of integrating out scales between 1 MeV and 0.001 eV, or we must assume that the low energy Lagrangians for any scale down to 0.001 eV have a mechanism in them which can cancel any given value of Λ. The first proposal is contrary to all of our experience with local field theory. Cancellations due to symmetries occur momentum scale by momentum scale. The second proposal implies that we should be able to construct a mechanism for cancelling which does not utilize any fancy symmetries. The spectrum of very low energy particles simply does not contain anything even approximately degenerate with the electron which could cancel the virtual electron contribution to Λ. The η field is of course such a mechanism. Its virtue is that it shifts the burden of explanation from the gravitational field (which interacts with everything) to a scalar field which, at least at low energies, need interact only with gravity. We know how to produce scalar fields which have no potentials whatsoever. These are Goldstone bosons, the explanation for whose masslessness can be a symmetry that is spontaneously broken at very high energies. We also know how to give such Goldstone bosons extremely small symmetry breaking potentials.[♯9]

Unfortunately, Goldstone bosons for ordinary compact symmetries will not do the job for us. These fields live on compact manifolds and the only way we can make the restoring force small is by making the potential very small. This means however that it cannot cancel the cosmological constant. Noncompact global symmetries have recently become popular in discussions of extended supergravity.[18] They can certainly be spontaneously broken. What is not clear is whether such symmetries can also be explicitly broken by nonperturbative effects in weakly coupled field theories. This is a question which merits investigation, for the pseudo-Goldstone boson for such a symmetry would be a perfect candidate

♯9 Extend the standard model by adding 3 colorless Weyl fermions in the $N+\bar{N}$ representation of an $SU(N)$ hypercolor group. Let the hypercolor scale be much larger than 100 GeV, and let the Glashow Salam Weinberg gauge bosons couple to the $SU(2)\times U(1)$ subgroup of the diagonal $SU(3)$ subgroup of the $SU(3)\times SU(3)$ flavor group of hypercolor. Spontaneous breakdown of $SU(3)\times SU(3)$ leads to Goldstone bosons. However, some of the relevant currents have $SU_2\times U_1$ anomalies, and the corresponding Goldstone bosons will get mass from weak instanton effects. Obviously, many variations of this scenario are possible.

for the η field.

The superpartners of ordinary Goldstone bosons in supersymmetric field theories are often noncompact fields with flat potentials. In typical models however, the flatness does not survive supersymmetry breaking to the extent that would be necessary to construct an η field. It is probably possible to construct models with sufficiently many decoupled sectors to construct a very small flat potential, but nothing really compelling has emerged from preliminary glances at this idea.

The final candidate for the η field is the third rank antisymmetric tensor gauge field that I discussed in (3) (see also Ref. 19). If a dynamical Higgs mechanism[20] occurs for this field, its dynamics is identical to those of a massive scalar field. The dynamical Higgs phenomenon for third rank tensors in four dimensional space is too poorly understood to know whether a mass of the requisite (tiny) magnitude is plausible.

To summarize, there are several possible candidates which might naturally give rise to a field with the properties required by the cosmology of Section 5. They are all technically natural, but it is not clear whether a sufficiently flat potential can be generated.

7. Conclusions

We have shown that the formalism of quantum gravity can tell us interesting things about cosmology even if we restrict our attention to the semiclassical regime in which the short distance pathologies of Einstein's Lagrangian are not important. We have understood the relationship between stationary quantum states and time dependent classical solutions of Einstein's equations (a result which goes back to Wheeler and DeWitt) and shown how to rederive the "quantum field theory in curved spacetime" formalism as a description of correlations between gravitational and matter variables in the time independent Wheeler-DeWitt wave function. This description, and the classical notion of time make sense only within the domain of validity of the semiclassical approximation. We have also shown how quantum gravity may lead to spontaneous breakdown of TCP invariance, thus making the cosmic arrow of time compatible with TCP.

These results indicate the compatibility of our conventional picture of cosmology with the (presumably) more fundamental formalism of quantum gravity. However, there are two possible points of incompatibility. First, the general WD wave function in the semiclassical approximation is a superposition of wave functions for different classical trajectories. To understand why we see a classical universe, we must argue that all these different wave functions become equivalent or orthogonal (in which case we must resort to a reduction of the wave packet of the universe) after some time. Alternatively we can search for a reason why the correct WKB wave function contains only one classical solution. We showed that this occurred in models with a negative cosmological constant which can be compensated by the classical excitation of a scalar field η. These models also resolve the potential conflict between the natural initial conditions for quantum matter fields in our formalism, and the initial conditions for the Hot Big Bang. They envisage that the classical history of the universe "began" with a tunneling event from the region of small volume and unexcited η, (where the wave function is concentrated) to the classically allowed region where η is excited, the cosmologi-

cal constant is positive and the universe can expand. If the classical relaxation of η to its minimum is slow enough, it acts like a variable cosmological constant. We presented a model in which this variation of the cosmological constant induces a "curvature driven" first order phase transition which generates all of the matter in the universe, setting up the initial conditions for the Hot Big Bang. Models of this type can generate many universes and provide a basis for the application of the anthropic principle to the observed value of the cosmological constant. Unfortunately in the present version of the model the probability distribution for the cosmological constant is rapidly growing as Λ decreases thru 0. Consequently, even if we invoke the anthropic principle, we are led to expect that the universe is on the edge of oblivion. The universe should have the largest possible negative value of the cosmological constant compatible with the existence of intelligent life. This is not a good description of our universe, which probably has a positive cosmological constant if it has one at all. Hopefully a model of this type can be found which has a flatter probability distribution in the region of interest. In addition, it would be encouraging if we could find a model which resolved some of the more mundane issues of cosmology, like the flatness problem and galaxy formation. At present we have merely mimicked the new inflationary universe models with all of their faults. "Of course", work on these matters is in progress.

Our aim in the present paper has been to try to find models in which the classical history of the universe, including initial conditions, could be described without reference to the ill-understood short distance behavior of quantum gravity. From a certain point of view it would be a pity if such a model exists. It would mean that there will be no cosmological clues for the construction of the correct short distance theory.

Appendix A — Semiclassical Approximation

Our objective is to solve

$$\left[\frac{\gamma_{ijk\ell}}{M^2}\,\frac{\delta^2}{\delta g_{ij}\,\delta g_{k\ell}} - M^2\sqrt{g}\,(R^{(3)}-\Lambda) + \mathcal{H}_m\right]\psi = 0 \qquad (A.1)$$

to zeroth order accuracy in M as $M \to \infty$. To this end write

$$\psi = e^{iM^2 S}\psi_1 \qquad (A.2)$$

where S satisfies the Einstein-Hamilton-Jacobi equation

$$\gamma_{ijk\ell}\,\frac{\delta S}{\delta g_{ij}}\,\frac{\delta S}{\delta g_{k\ell}} - \sqrt{g}\,(R^{(3)}-\Lambda) = 0\,. \qquad (A.3)$$

Terms of zeroth order in M arise when we differentiate S twice, and when we let one derivative act on $\exp[iM^2S]$ and the other on ψ. Thus

$$i\,\gamma_{ijk\ell}\,\frac{\delta^2 S}{\delta g_{ij}\,\delta g_{k\ell}}\,\psi_1 + i\,\gamma_{ijk\ell}\,\frac{\delta S}{\delta g_{ij}}\,\frac{\delta \psi_1}{\delta g_{k\ell}} + \mathcal{H}_m\psi_1 = 0\,. \qquad (A.4)$$

Let ψ_{vv} be the solution of Eq. (4) when $\mathcal{H}_m = 0$. It is the generalization of the Van-Vleck determinant to quantum gravity, and depends on the operator ordering that we have chosen for the kinetic term. The full solution to order M^0 is then

$$\psi - e^{iM^2 S}\psi_{vv}\chi \qquad (A.5)$$

where

$$i\,\gamma_{ijk\ell}\,\frac{\delta S}{\delta g_{ij}}\,\frac{\delta \chi}{\delta g_{k\ell}} + \mathcal{H}_m\chi = 0\,. \qquad (A.6)$$

Now let $\tau(x;y)$ be the functional of $g_{ij}(x)$ defined by

$$\gamma_{ijk\ell}\,\frac{\delta S}{\delta g_{ij}(x)}\,\frac{\delta \tau(y)}{\delta g_{k\ell}(x)} = \delta(x-y) \qquad (A.7)$$

and assume χ depends on g_{ij} only through τ. Then

$$i\frac{\delta\chi}{\delta\tau(x)} = \mathcal{H}_m(x)\chi \ . \tag{A.8}$$

The solutions of Eq. (3) are parametrized by a spatial metric $\bar{g}$. $S[g;\bar{g}]$ is the action of the solution of Einstein's equations which interpolates between the spatial metrics $\bar{g}$ and g. Given this classical spacetime, g is simply the induced metric on a certain spacelike surface. Under shifts and reparametrizations of this spacelike surface, g_{ij} transforms as

$$\delta g_{ij}(x) = K_{ij}(x)\delta N(x) + \nabla_{(i}\delta N_{j)} \tag{A.9}$$

K_{ij} is the extrinsic curvature of the surface. This means that we can change d of the degrees of freedom in g by changing surfaces and coordinates within the fixed classical space time. The other $d(d-1)/2$ can only be varied by changing the classical solution. Thus, for a fixed spacetime geometry Eq. (9) defines N and N_i as functionals of g.

Now note that

$$\gamma_{ijk\ell}\frac{\delta S}{\delta g_{k\ell}} = +K_{ij} \tag{A.10}$$

so that Eq. (7) reads:

$$+K_{ij}(x)\frac{\delta\tau(y)}{\delta g_{ij}(x)} = \int dz\,\frac{\delta g_{ij}(z)}{\delta N(x)}\,\frac{\delta\tau(y)}{\delta g_{ij}(z)} = \frac{\delta\tau(x)}{\delta N(y)} = \delta(x-y) \tag{A.11}$$

which has the solution

$$\tau(x) = N(y) \ . \tag{A.12}$$

Equation (8) is thus the Tomonage-Schwinger equation. It describes the change in the wave function of the matter fields we make an infinitesimal local change in the spacelike surface on which it is defined. The space-time geometry is a classical solution of Einstein's equations.

Appendix B

Questions of topology and change of topology in quantum gravity are a source of great controversy, and have been ever since Wheeler introduced the idea of space-time foam. The simplest topological issue is the question of whether the universe is spatially open or closed. Spatial closure was considered desirable by Einstein and Wheeler because it is supposed to incorporate Mach's principle into general relativity. Certainly it seems odd to consider an observer who sits outside the universe measuring asymptotic time and defining an absolute reference frame. On the other hand, in current cosmologies the question of closure seems to be one that must be settled by observation. It seems silly to prejudice an empirically resolvable question by what is essentially a philosophical rather than a physical argument. In new inflationary cosmologies and models of the type proposed in the present paper this question takes on a new aspect. The observable universe in these models is the interior of a small bubble in a larger (unobservable?) meta-universe. Thus we can indulge our philosophical prejudices without affecting the experimentally testable question of whether the universe will re-collapse. It really doesn't matter whether we take the meta-universe to be open or closed. As indicated in the text, we will take it to be closed because the formalism is simpler.

The question of the local topology of space, and of changes in it is more difficult to resolve. Hawking has argued that topologically non-trivial instantons (which change the topology of space) have to be included in the Euclidean functional integral for quantum gravity. He also argues[21] that these instantons will violate the unitary time evolution property of quantum mechanics and force us to introduce a theory of density matrices in which pure states can evolve into mixed states. This proposal has been criticized by several authors.[22] It is certain that the conventional canonical formalism for quantum gravity does not extend simply to describe changes in spatial topology. However it is neither clear that this is impossible nor that one must really include instantons in the path integral. Indeed, the reason one must include instanton contributions in ordinary quantum

mechanics is to guarantee unitarity. If gravitational instantons violate unitarity, then what have we gained by including them? More technically, Gross and Witten argue that the rationale for introducing instantons in Yang-Mills theory is that topologically trivial configurations with widely separated instanton-anti-instanton pairs must certainly be included in the path integral. If the dynamics allows the instantons to separate, then cluster decomposition requires that we include topologically non-trivial sectors as well. On the other hand, gravitational instantons which change the space topology carry a CPT invariant quantum number, the Euler number. We cannot construct topologically trivial configurations which consist of two widely separated pieces, each of which has nonzero Euler number. Hawking however argues that topologically non-trivial metrics can be approximated arbitrarily closely by trivial ones (though not in the same clustering manner as Yang-Mills instantons).

I believe that these questions cannot be resolved by arguments based on continuum field theories. If one goes beyond the semiclassical approximation to the path integral, one finds new divergences in topologically non-trivial sectors (in marked contrast to Yang-Mills theories). The relative weights of sectors with different Euler numbers is infinite in perturbation theory. This seems to indicate that the question of whether to include non-trivial topologies is bound up with short distance physics. Indeed,topology changes seem to be more easily incorporated in discrete theories[23] which make definite predictions about short distance behavior.

A final argument about the necessity of including topological changes comes from the study of gravity in low dimensions[24] where the short distance behavior is controllable. Perfectly consistent theories with Hermitian Hamiltonians can be constructed without including topology changes. This again indicates that if it is necessary to consider topology changes in the real world this must have something to do with the short distance modifications of Einstein's theory which are necessary in four or more dimensions. These low dimensional theories also shed some light on the question of what topology to choose for space if the

topology cannot change. One can construct sensible quantum theories of $2+1$ dimensional pure gravity if space has the topology of a torus or a sphere but not if it has a more complicated topology. The Hamiltonians for higher topologies have short distance problems.[24] These results motivated our restriction of three dimensional spatial topology in the text.

To summarize: there are many indications that topology changes do not have to be allowed in quantum gravity in order to construct a consistent theory. The consistency of theories which allow topology changes probably cannot be determined without a better understanding of what goes on at short distances, and it is conceivable that short distance physics will force us to include topology changes in four dimensional quantum gravity.

ACKNOWLEDGEMENTS

This work was begun when I was a member of the Institute for Advanced Study. I would like to thank Professor Harry Woolf for his hospitality and to acknowledge enlightening conversations with Mike Dine and Ed Witten. I enjoyed the hospitality of the SLAC theory group while completing this manuscript and I would like to thank E. Martinec, M. Peskin, M. Weinstein and especially Lenny Susskind for many discussions.

REFERENCES

1. S. Hawking, Comm. Math. Phys. 87, 395 (1982); Comm Math. Phys. 80, 421 (1981); Proceedings of the Second Oxford Conference on Quantum Gravity, pp 393, 415 (Oxford 1980); Phys. Lett. 86B, 175 (1979) (with D. Page and C. Pope); Nucl. Phys. B170, 283 (1980) (with D. Page and C. Pope); Phys. Rev. D14, 2460 (1976); Phys. Rev. D13, 191 (1976).

2. S. Hawking, "Quantum Cosmology", DAMTP preprint 84 0114, presented at Les Houches Summer School, June 27-August 4, 1983; "The Quantum State of the Universe", DAMTP preprint 84 0117, November 1983; Phys. Rev. D28, 2960 (1983) (with J. Hartle); T. Banks and L. Susskind, Int. J. Theor. Phys. 23, 475 (1984); "Two Lectures on 2D Gravity", IAS preprint, November 1983; T. Banks, W. Fischler, L. Susskind, Stanford preprint, August 1984.

3. T. Banks, IAS preprint, January 1984, published in Phys. Rev. Lett., April 1984.

4. PAM Dirac Lectures on Quantum Mechanics, Belfer Graduate School Monograph 2, Yeshiva University, New York 1964; R. Arnowitt, S. Deser, C. Misner in Witten "Gravitation", Wiley NY (1962) p. 227; B. DeWitt, Phys. Rev. 160, 1113 (1967); C. Misner, K. Thore, J. Wheeler, "Gravitation", W. H. Freeman, San Francisco (1971); A. Hanson, T. Regge, C. Teitelboim, "Constrained Hamiltonian Systems", Lincei Academy Report No. 2 (1976); M. Pilati, Phys. Rev. D26, 2645 (1982); C. J. Isham, Lectures given at 1983 Les Houches School on Relativity Groups and Topology, June 27-August 4, 1983; Proc. R. Soc. A368, 33 (1979).

5. C. Misner in Relativity (eds. M. Carmeli, S. Fickler, L. Witten), Plenum NY (1970), p. 55; in Magic Without Magic (ed. Klauder), W. H. Freeman, San Francisco (1982) and references cited there.

6. T. Banks, W. Fischler, L. Susskind, op. cit.

7. J. L. Gervais and B. Sakita, H. deVega, Nucl. Phys. B139, 20 (1978); Phys. Rev. D16, 3507 (1977); S. Coleman and C. Callan, Phys. REv. D16, 1762 (1977); S. Coleman, Phys. Rev. D15, 2929 (1977); T. Banks, C. M. Bender, T. T. Wu, Phys. Rev. D8, 3346 (1973); T. Banks, C. M. Bender, Phys. Rev. D8, 3366 (1973).

8. A. Peres, Nuovo Cimento 26, 53 (1962); U. Gerlach, Phys. Rev. 177, 1929 (1969).

9. I. Newton, Philosophiae Naturalis Principia Mathematica (1687), University of California Press, Berkeley and Los Angeles (1934).

10. S. Hawking and I. G. Moss, Phys. Lett. 110B, 35 (1982).

11. S. Weinberg, Gravitation and Cosmology, John Wiley, NY (1972); P.J.E. Peebles, Physical Cosmology, Princeton University Press, Princeton, NJ (1971).

12. A. Sakharov, unpublished; L. Abbott, Nucl. Phys. B185, 233 (1981); A. Vilenkin, Nucl. Phys. B226, 504 (1983); Phys. Rev. D26, 1231 (1982); Phys. Lett. 115B, 91 (1982).

13. A. Guth and E. Weinberg, Phys. Rev. D23, 876 (1981).

14. A. D. Linde, Lebedev preprint 83-30, December 1982; Phys. Lett. 116B, 335 (1982); Phys. Lett. 116B, 340 (1982); Phys. Lett. 114B, 431 (1982); A. Albrecht and P. J. Steinhardt, Phys. Lett. 131B, 45 (1983); Phys. Rev. Lett. 48, 1220 (1982); B. Ovrut, P. Steinhardt, Phys. Lett. 133B, 161 (1983); S. Gupta, H. Quinn, SLAC-PUB-3269.

15. S. Coleman, F. DeLuccia, Phys. Rev. D21, 3305 (1980).

16. M. Turner, G. Steigman, L. Krauss, Bartol Institute Preprint, BA 84 12, April 1984; P. J. E. Peebles, Princeton preprint (1984).

17. K. Symanzik, Comm. Math. Phys. 34, 7 (1973); T. Appelquist and J. Carrazone, Phys. Rev. D11, 2856 (1975).

18. E. Cremmer, B. Julia, Nucl. Phys. B159, 141 (1979); J. Ellis, M. Gaillard, B. Zumino, Nucl. Phys. B224, 427 (1983); J. Ellis, C. Kounnas, D. Nanopoulos, CERN TH 3768, November 1983.

19. S. Hawking, Phys. Lett. 134B, 403 (1984).

20. A. Aurilia, Y. Takahashi, Prog. Theor. Phys. 60, 693 (1981).

21. S. Hawking, op. cit., Ref. 1 and DAMTP preprint 1984.

22. T. Banks, M. Peskin, L. Susskind, SLAC-PUB-3258, December 1983; D. Gross, Nucl. Phys. B236, 349 (1984).

23. M. Weinstein and V. Kaplunovsky, SLAC-PUB-3156, July 1983.

24. E. Martinec, SLAC-PUB-3306, May 1984.

25. A. Zee, Ann. Phys. 151, 431 (1983).

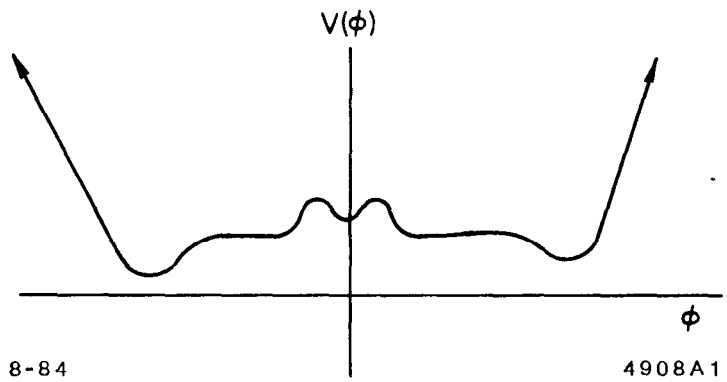

Fig. 1. The flat space effective potential for the scalar field which generates the energy in the universe in the model of Section 5.