13

False Vacuum: Early Universe Cosmology and the Development of Inflation

Chris Smeenk

University of California, Los Angeles, Department of Philosophy, 321 Dodd Hall, Los Angeles, CA 90095, U.S.A.; smeenk@humnet.ucla.edu

Inflationary cosmology has been widely hailed as the most important new idea in cosmology since Gamow's pioneering work on nucleosynthesis, or perhaps even since the heady early days of relativistic cosmology in the 1920s. Popular accounts typically attribute the invention of inflation to Alan Guth, whose seminal paper (Guth 1981) created a great deal of excitement and launched a research program. These accounts typically present Guth and a small band of American particle physicists as venturing into untouched territory. More careful accounts (such as Guth's memoir, Guth 1997) acknowledge that inflation's central idea, namely that the early universe passed through a brief phase of exponential expansion, did not originate with Guth. Reading this earlier research merely as an awkward anticipation of inflation seriously distorts the motivations for these earlier proposals, and also neglects the wide variety of motivations for such speculative research. Below I will describe several proposals that the early universe passed through a de Sitter phase, highlighting the different tools and methodologies used in the study of the early universe.

The early universe was the focus of active research for over a decade before Guth and other American particle physicists arrived on the scene in the late 1970s. The discovery of the background radiation in 1965 brought cosmology to the front page of the *New York Times* and to the attention of a number of physicists. In his influential popular book *The First Three Minutes*, Steven Weinberg characterized the effect of the discovery as follows:

> [Prior to discovery of the background radiation]...it was extraordinarily difficult for physicists to take seriously *any* theory of the early universe. ... The most important thing accomplished by the ultimate discovery of the 3°K radiation background in 1965 was to force us all to take seriously the idea that there *was* an early universe. (Weinberg 1977, 131–132)

Taking the early universe seriously led to efforts to extend the well understood "standard model" of cosmology developed in the 1960s, accepted by a majority of mainstream cosmologists and presented in textbooks such as Peebles (1971); Weinberg (1972), to ever earlier times. According to the standard model, the large scale structure of the universe and its evolution over time are aptly described by the simple

Friedmann–Lemaître–Robertson–Walker (FLRW) models. Extrapolating these models backwards leads to a hot, primeval "fireball," the furnace that produced both the background radiation and characteristic abundances of the light elements. Finally, the theory included the general idea that large scale structure, such as galaxies and clusters of galaxies, formed via gravitational clumping. But the standard model was not without its blemishes. In particular, it was well known that extrapolating the FLRW models led to arbitrarily high energies and a singularity as $t \to 0$.

The paper proceeds as follows. The first section below focuses on efforts by a number of Soviet cosmologists to eliminate the initial singularity. Their abhorrence of the singularity was strong enough to motivate a speculative modification of the FLRW models, namely patching on a de Sitter solution in place of the initial singularity. Gliner and Sakharov arrived at the idea by considering "vacuum-like" states of matter, whereas Starobinsky found that de Sitter space is a solution to Einstein's field equations (EFE) modified to incorporate quantum corrections. These proposals highlight two problems facing any modification of the early universe's evolution: what drives a change in the expansion rate near the singularity, and how does an early de Sitter phase lead into the standard big bang model? Section 2 turns to the influx of ideas into early universe cosmology from particle physics, focusing in particular on symmetry breaking. A group of physicists in Brussels proposed that the "creation event" could be understood as a symmetry breaking phase transition that sparked the formation of a de Sitter-like bubble, which eventually slowed to FLRW expansion. The more mainstream application of symmetry breaking to cosmology focused on the consequences of symmetry breaking phase transitions. Early results indicated a stark conflict with cosmological theory and observation. Despite this inauspicious beginning, within a few years early universe phase transitions appeared to be a panacea for the perceived ills of standard cosmology rather than a source of wildly inaccurate predictions.

13.1 Eliminating the Singularity

Cosmologists have speculated about the nature of the enigmatic "initial state" since the early days of relativistic cosmology. Research by Richard Tolman, Georges Lemaître and others in the 1930s established the existence of an initial singularity in the FLRW models, but this was typically taken to represent a limitation of the models rather than a feature of the early universe. Debates about exactly how to define a "singularity" continue to the present, but in early work singularities were usually identified by divergences in physical quantities (such as the gravitational field or curvature invariants).[1] Tolman argued that the presence of a singular state reflects a breakdown of the various idealizations of the FLRW models (Tolman 1934, 438 ff.). But by the mid-1960s cosmologists could not easily dismiss singularities as a consequence of unphysical idealizations. New mathematical techniques developed primarily by Roger Penrose, Stephen Hawking, and Robert Geroch made it possible to prove the celebrated "singularity theorems." These theorems established that singularities, signalled by the presence of incomplete geodesics,[2] are a generic feature of solutions to the field equations of general relativity that: satisfy global causality constraints (ruling out

pathologies such as closed time-like curves), contain matter fields satisfying one of the energy conditions, and possess a point or a surface such that light cones start converging towards the past. The precise characterization of these assumptions differed for various singularity theorems proved throughout the 1960s, but in general these assumptions seemed physically well motivated (see, e.g., Hawking and Ellis 1968). Thus these powerful theorems dashed the hope that a singularity could be avoided in "more realistic" cosmological models.

The prominent Princeton relativist John Wheeler described the prediction of singularities as the "greatest crisis in physics of all time" (Misner et al. 1973, 1196–1198). Confronted with this crisis many of Wheeler's contemporaries took evasive maneuvers. A number of prominent Soviet physicists (including Lev Landau, Evgeny Lifshitz, Isaak Khalatnikov, and several collaborators) analyzed the (allegedly) general form of cosmological solutions to Einstein's field equations (EFE) in the neighborhood of the singularity, with the hope of showing that the singular solutions depend upon a specialized choice of initial conditions.[3] Although this group (eventually) accepted the results of the singularity theorems, there were other ways of evading an initial singularity. Approaching the initial singularity (or singularities produced in gravitational collapse) leads to arbitrarily high energies, and theorists expected the as yet undiscovered theory of quantum gravity to come into play as energies approached the Planck scale, undercutting the applicability of the theorems.[4] But there was another obvious escape route: deny one of the assumptions. Another line of research made denial of the energy conditions more appealing: the "vacuum" in modern field theory turned out to be anything but a simple "empty" state, and in particular a vacuum state violated the energy conditions. Several Soviet cosmologists, who apparently abhorred the singularity more than the vacuum, proposed that an early vacuum-like state would lead to a de Sitter bubble rather than a singularity.

13.1.1 Λ in the USSR

Two Soviet physicists independently suggested that densities reached near the big bang would lead to an effective equation of state similar to a relativistic vacuum: Andrei Sakharov, the famed father of the Soviet H-bomb and dissident, considered the possibility briefly in a study of galaxy formation (Sakharov 1966), and a young physicist at the Ioffe Physico-Technical Institute in Leningrad, Erast Gliner, noted that a vacuum-like state would counter gravitational collapse (Gliner 1966). Four further papers over the next decade developed cosmological models on this shaky foundation (Gliner 1970; Sakharov 1970; Gliner and Dymnikova 1975; Gurevich 1975), in the process elaborating on several of the advantages and difficulties of an early de Sitter phase.

Gliner's paper took as its starting point an idea that has been rediscovered repeatedly: a non-zero cosmological constant Λ may represent the gravitational effect of vacuum energy.[5] Einstein modified the original field equations of general relativity by including a Λ term to vouchsafe cherished Machian intuitions (Einstein 1917), but later thought it marred general relativity's beauty. Even for those who didn't share Einstein's aesthetic sensibility, observational constraints provided ample evidence that Λ

must be *very close* to zero. Yet, as Gliner (1966) and others noted, Λ could be treated as a component of the stress-energy tensor, $T_{ab} = -\rho_V g_{ab}$ (where "V" denotes vacuum); a T_{ab} with this form is the only stress energy tensor compatible with the requirement that the vacuum state is locally Poincaré invariant.[6] The stress-energy tensor for a perfect fluid is given by

$$T_{ab} = (\rho + p)u_a u_b + p g_{ab}, \tag{13.1}$$

where u^a represents the normed velocity of the perfect fluid, ρ is the energy density and p is pressure. The vacuum corresponds to an ideal fluid with energy density $\rho_V \left(= \frac{\Lambda c^2}{8\pi G}\right)$ and pressure given by $p_V = -\rho_V$; this violates the strong energy condition, often characterized as a prerequisite for any "physically reasonable" classical field.[7] Yakov Zel'dovich, whom Gliner thanked for critical comments, soon published more sophisticated studies of the cosmological constant and its connection with vacuum energy density in particle physics (Zel'dovich 1967, 1968). The main thrust of Gliner's paper was to establish that a vacuum stress-energy tensor should not be immediately ruled out as "unphysical," whereas Zel'dovich (1968) proposed a direct link between Λ and the zero-point energy of quantum fields.

The novelty of Gliner's paper lies in the conjecture that high density matter somehow makes a transition into a vacuum-like state. Gliner motivated this idea with a stability argument (cf. Gliner 1970), starting from the observation that matter obeying an ordinary equation of state is unstable under gravitational collapse. For normal matter and radiation, the energy density ρ increases without bound during gravitational collapse and as one approaches the initial singularity in the FLRW models.[8] However, Gliner recognized that the energy density remains constant in a cosmological model with a vacuum as the only source. The solution of the field equations in this case is de Sitter space, characterized by exponential expansion $a(t) \propto e^{\chi t}$, where $(\chi)^2 = (8\pi/3)\rho_V$ and the scale factor $a(t)$ represents the changing distance between fundamental observers. During this rapid expansion the vacuum energy density remains constant, but the energy density of other types of matter is rapidly diluted. Thus extended expansion should eventually lead to vacuum domination as the energy density of normal matter becomes negligible in comparison to vacuum energy density.[9] It is not clear whether Gliner recognized this point. But he did argue that if matter undergoes a transition to a vacuum state during gravitational collapse, the result of the collapse would be a de Sitter "bubble" rather than a singularity. This proposal avoids the conclusion of the Hawking–Penrose theorems by violating the assumption that matter obeys the strong energy condition. In effect, Gliner prefered a hypothetical new state of matter violating the strong energy condition to a singularity, although he provides only extremely weak plausibility arguments suggesting that "vacuum matter" is compatible with contemporary particle physics.[10]

By contrast with Gliner's outright stipulation, Sakharov (1966) hoped to derive general constraints on the equation of state at high densities by calculating the initial peturbations produced at high densities and then comparing the evolution of these perturbations to astronomical observations. Sakharov argued that at very high densities (on the order of 2.4×10^{98} baryons per cm^3!) gravitational interactions would need to

be taken into account in the equation of state. Although he admitted that theory was too shaky to calculate the equation of state in such situations, he classified four different types of qualitative behavior of the energy density as a function of baryon number (Sakharov 1966, 74–76). This list of four included an equation of state with $p = -\rho$, and Sakharov noted that feeding this into FLRW dynamics yields exponential expansion. But the constraints Sakharov derived from the evolution of initial perturbations appeared to rule this out as a viable equation of state. In a 1970 preprint (Sakharov 1970), Sakharov again considered an equation of state $\rho = -p$, this time as one of the seven variants of his speculative "multi-sheet" cosmological model.[11] This stipulation was not bolstered with new arguments (Sakharov cited Gliner), but as we will see shortly Sakharov discovered an important consequence of an early vacuum state.

Three later papers developed Gliner's suggestion and hinted at fruitful connections with other problems in cosmology. Gliner and his collaborator, Irina Dymnikova, then a student at the Ioffe Institute, proposed a cosmological model based on the decay of an initial vacuum state into an FLRW model, and one of Gliner's senior colleagues at the Institute, L. E. Gurevich, pursued a similar idea. According to the Gliner and Dynmikova's model, an initial fluctuation in the vacuum leads to a closed, expanding universe. The size of the initial fluctuation is fixed by the assumption that $\dot{a} = 0$ at the start of expansion. The vacuum cannot immediately decay into radiation. This would require joining the initial fluctuation to a radiation-dominated FLRW model, but as a consequence of the assumption this model would collapse rather than expand—the closed FLRW universe satisfies $\dot{a} = 0$ only at *maximum* expansion.[12] Gliner and Dymnikova (1975) stipulated that the effective equation of state undergoes a gradual transition from a vacuum state to that of normal matter.[13] The scale factor and the mass of the universe both grow by an incredible factor during this transitional phase, as Gliner and Dymnikova (1975) noted; however, there is no discussion of whether this is a desirable feature of the model.

This proposal replaces the singularity with a carefully chosen equation of state, but Gliner and Dymnikova (1975) give no physical motivation guiding these choices. Instead, details of the transition are set by matching observational constraints. As a result of this phenomenological approach, Gliner and Dymnikova (1975) failed to recognize one of the characteristic features of a de Sitter-like phase. In particular, the following equation relates parameters of the transition (the initial and final energy densities, ρ_0 and ρ_1, and the "rate" set by the constant α) to present values of the matter and radiation density (ρ_p, ρ_{rp}):[14]

$$\sqrt{\frac{\rho_1}{\rho_{rp}}} \exp\left(\frac{2(\rho_0 - \rho_1)}{3\gamma\rho_1(1-\alpha)}\right) = \frac{\rho_0}{\rho_p}\left(1 - \frac{3H^2}{8\pi G\rho_p}\right)^{-1}. \tag{13.2}$$

This equation indicates how the length of the transitional phase effects the resulting FLRW model: for a "long" transitional phase, ρ_1 is small, and the left-hand side of the equation is exponentially large. This forces the term in parentheses on the right-hand side to be exponentially small, so that H^2 approaches $\frac{8\pi G\rho_p}{3}$, the Hubble constant for a flat FLRW model. Four years later, Guth would label his discovery of this feature a "Spectacular Realization," but Gliner and Dymnikova (1975) took no notice of it.

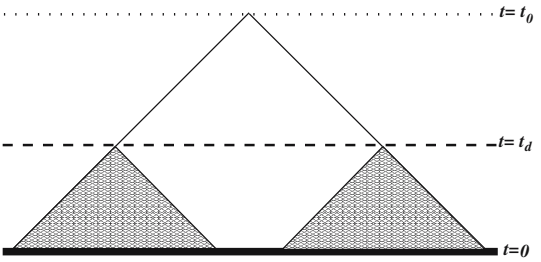

Fig. 13.1. This conformal diagram illustrates the horizon problem in the FLRW models. The singularity at $t = 0$ is stretched to a line. The lack of overlap in the past light cones at points on the surface $t = t_d$ (both within the horizon of an observer at $t = t_0$) indicates that no causal signal could reach both points from a common source.

Gurevich and Sakharov both had a clearer vision of the possible cosmological implications of Gliner's idea than Gliner himself. Gurevich (1975) noted that an initial vacuum dominated phase would provide the "cause of cosmological expansion." Gurevich clearly preferred an explanation of expansion that did not depend on the details of an initial "shock" or "explosion," echoing a concern first voiced in the 1930s by the likes of Sir Arthur Eddington and Willem de Sitter.[15] Gurevich aimed to replace various features of the initial conditions — including the initial value of the curvature, the "seed fluctuations" needed to form galaxies, and the amount of entropy per baryon — with an account of the formation and merger of vacuum-dominated bubbles in the early universe. The replacement was at this stage (as Gurevich admitted) only a "qualitative picture of phenomena" (Gurevich 1975, 69), but the goal itself was clearly articulated.

Gurevich failed to recognize, however, the implications of a vacuum-dominated phase for a problem he emphasized as a major issue in cosmology: Misner's horizon problem (Misner 1969). Horizons in relativistic cosmology mark off the region of space-time from which light signals can reach a given observer. The "particle horizon" measures the maximum distance from which light signals could be received by an observer at t_0 as the time of emission of the signal approaches the initial singularity:[16]

$$d_{ph} = \lim_{t \to 0} a(t_0) \int_t^{t_0} \frac{dt}{a(t)}. \tag{13.3}$$

This integral converges for $a(t) \propto t^n$ with $n < 1$ (satisfied for matter- or radiation-dominated FLRW models), leading to a finite horizon distance. A quick calculation shows that regions emitting the background radiation at nearly the same temperature lie *outside* each other's particle horizons. The horizon problem refers to the difficulty in accounting for this observed uniformity given the common assumption that the universe began in a "chaotic" initial state (see Figure 13.1). Misner (1969) suggested that more realistic models of the approach to the singularity would include "mixmaster oscillations," effectively altering the horizons to allow spacetime enough for causal interactions, but by 1975 a number of Gurevich's comrades (along with British cosmologists and Misner himself) had put the idea to rest (see, e.g., Criss et al. 1975,

for a *post mortem*). But mixmaster oscillations were unnecessary to solve the horizon problem; as Sakharov recognized, an odd equation of state would suffice:[17]

> If the equation of state is $\rho \approx S^{2/3}$ [where S is baryon number density; this is equivalent to $p = -\frac{\rho}{3}$], then $a \approx t$ and the Lagrangian radius of the horizon is
>
> $$\int_{t_0}^{t_1} \frac{dt}{a} \to \infty \quad \text{as} \quad t_0 \to 0, \tag{13.4}$$
>
> i.e., the horizon problem is resolved without recourse to anisotropic models.

To my knowledge this is the earliest "solution" of the horizon problem along these lines. (It is a solution only in the sense that altering the horizon structure makes causal interactions possible, but it does not specify an interaction that actually smooths out chaotic initial conditions.) Sakharov's colleagues at the Institute of Applied Mathematics in Moscow, notably including Igor Novikov and Zel'dovich, were probably aware of this result. But it appeared buried in the Appendix of a preprint that was only widely available following the publication of the *Collected Works* in 1982.

13.1.2 Starobinsky's model

During a research year in Cambridge in 1978–79, Zel'dovich's protegé Alexei Starobinsky developed an account of the early universe based on including quantum corrections to the stress-energy tensor in EFE. Starobinsky clearly shared Gliner and Dymnikova's willingness to replace the initial singularity with an early de Sitter phase. But there the similarity with Gliner and Dymnikova's work ends. Unlike their sterile phenomenological approach, Starobinsky's model drew on a rich source of ideas: recent results in semi-classical quantum gravity.

Throughout the 1970s Starobinsky was one of the main players in Zel'dovich's active team of astrophysicists at the Institute of Applied Mathematics, focusing primarily on semi-classical quantum gravity. Starobinsky brought considerable mathematical sophistication to bear on Zel'dovich's insightful ideas, including the study of particle production in strong gravitational fields and the radiation emitted by spinning black holes (a precursor of the Hawking effect). The relationship between the energy conditions and quantum effects was a recurring theme in this research. In response to an alleged "no go theorem" due to Hawking, Zel'dovich and Pitaevsky (1971) showed that during particle creation the effective T_{ab} violates the dominant energy condition.[18] Energy conditions might be violated as a consequence of effects like particle creation, but Starobinsky was unwilling to introduce new fields solely to violate the energy conditions. Shortly before developing his own model, Starobinsky criticized Parker and Fulling's (1973) proposal that a coherent scalar field would violate the strong energy condition and lead to a "bounce" rather than a singularity, pointedly concluding that "there is no reason to believe that at ultrahigh temperatures the main contribution to the energy density of matter will come from a coherent scalar field" (Starobinsky 1978, 84).[19]

Starobinsky's (1979, 1980) model accomplished the same result without introducing fundamental scalar fields. By incorporating quantum effects Starobinsky found a class of cosmological solutions that begin with a de Sitter phase, evolve through an oscillatory phase, and eventually make a transition into an FLRW expanding model. In the semi-classical approach, the classical stress-energy tensor is replaced with its quantum counterpart, the renormalized stress-energy tensor $\langle T_{ab} \rangle$, but the metric is not upgraded. Calculating $\langle T_{ab} \rangle$ for quantum fields is a tricky business due to divergences, but several different methods were developed to handle this calculation in the 1970s. Starobinsky's starting point was the one-loop correction to $\langle T_{ab} \rangle$ for massless, conformally invariant, non-interacting fields. Classically the trace for such fields vanishes, but due to regularization of divergences $\langle T_{ab} \rangle$ includes the so-called "trace anomaly."[20] Taking this anomaly into account, Starobinsky derived an analog of the Friedman equations and found a set of solutions to these equations.[21] This establishes the existence (but not uniqueness) of a solution that begins in an unstable de Sitter state before decaying into an oscillatory solution. Using earlier results regarding gravitational pair production, Starobinsky argued that the oscillatory behavior of the scale factor produces massive scalar particles ("scalarons"). Finally, the matter and energy densities needed for the onset of the standard big bang cosmology were supposedly produced via the subsequent decay of these scalarons.

In the course of describing this model, Starobinsky mentioned an observational constraint that simplifies the calculations considerably (Starobinsky 1980, 101):

> If we want our solution to match the parameters of the real Universe, then [the de Sitter stage] should be long enough: $Ht_0 >> 1$, where t_0 is the moment of transition to a Friedmann stage. This enables us to neglect spatial curvature terms ... when investigating the transition region.

The published version of a paper delivered in 1981 at the Moscow Seminar on Quantum Gravity (Starobinsky 1984) repeated a portion of this earlier paper with a page of new material added.[22] This added material explains that an extended de Sitter phase drives the universe very close to a "flat" FLRW model, with negligible spatial curvature. But Starobinsky did not present this aspect of the model as an important advantage: he commented that an extended de Sitter phase is necessary simply to insure compatibility with observations, and he did not further comment on whether an extended de Sitter phase is a *natural* or *desirable* feature of his model. Starobinsky's approach requires *choosing* the de Sitter solution, with no aim of showing that it is a "natural" state; as Starobinsky put it (Starobinsky 1980, 100), "This scenario of the beginning of the Universe is the extreme opposite of Misner's initial 'chaos'." In particular, his model takes the *maximally symmetric* solution of the semi-classical EFE as the starting point of cosmological evolution, rather than an *arbitrary* initial state as Misner had suggested.[23] In this assumption he was not alone: several other papers from the Moscow conference similarly postulate that the universe began in a de Sitter state (see, e.g., Grib et al. 1984; Lapchinksy et al. 1984).

Starobinsky's model led to two innovative ideas that held out some hope of observationally testing speculations about the early universe. The first of these was Starobinsky's prediction that an early de Sitter phase would leave an observational

signature in the form of gravitational waves. Starobinsky (1979) calculated the spectrum of long-wavelength gravitational waves, and argued that in the frequency range of $10^{-3} - 10^{-5}$ Hz an early de Sitter phase would produce gravitational waves with an amplitude not far beyond the limits of contemporary technology. Zel'dovich was thrilled at the prospect (Zel'dovich 1981, 228): "For this it would be worth living 20 or 30 years more!" Mukhanov and Chibisov (1981) introduced a second idea that would carry over to later early universe models: they argued that zero-point fluctuations in an initial vacuum state would be amplified during the expansion phase, leading to density perturbations with appropriate properties to seed galaxy formation. Both of these ideas would prove crucial in later attempts to identify a unique observational footprint of an early de Sitter-like phase.

Starobinsky's proposal created a stir in the Russian cosmological community: it was widely discussed at the Moscow Seminar on Quantum Gravity 1981, and Zel'dovich — undoubtedly the dominant figure in Soviet cosmology, both in terms of his astounding physical insight and his institutional role as the hard-driving leader of the Moscow astrophysicists — clearly regarded the idea as a major advance. Zel'dovich (1981) reviewed the situation with his typical clarity. One of the appealing features of Starobinsky's model, according to Zel'dovich, was that it provided an answer to embarassing questions for the big bang model, "What is the beginning? What was there before the expansion began [...]?" In Starobinsky's model the "initial state" was replaced by a de Sitter solution, continued to $t \to -\infty$. But Zel'dovich noted two other important advantages of Starobinsky's model. First, it would solve the horizon problem (Zel'dovich 1981, 229):[24]

> An important detail of the new conception is the circumstance that the de Sitter law of expansion solves the problem of causality in its stride. Any two points or particles (at present widely separated) were, in the distant de Sitter past, at a very small, exponentially small distance. They could be causally connected in the past, and this makes it possible, at least in principle, to explain the homogeneity of the Universe on large scales.

Second, perturbations produced in the transition to an FLRW model might produce gravitational waves as well as the density perturbations needed to seed galaxy formation. But Zel'dovich also emphasized the speculative nature of this proposal, concluding optimistically that "there is no danger of unemployment for theoreticians occupied with astronomical problems" (Zel'dovich 1981, 229).

13.1.3 Common Problems

These proposals illustrate common problems faced by speculative theories of the early universe's evolution. First, what is the physical source of an early vacuum-like state? Second, how could an early de Sitter-like phase make a transition into FLRW expansion, during which the vacuum is converted to the incredibly high matter and radiation densities required by the hot big bang model? Gliner's outright stipulations leave little room to refine or enrich the proposal by incorporating believable physics. The contrast with Starobinsky's model is stark: in 1980, Starobinsky's model appeared to be on the

verge of being developed systematically into a detailed model of the early universe based on speculative but actively studied aspects of semi-classical quantum gravity. As we will see in the next section, cosmologists would instead develop a detailed model of an early de Sitter phase based on a rich new idea from particle physics: symmetry breaking phase transitions.

13.2 Symmetries and Phase Transitions

This section focuses primarily on the study of early universe phase transitions, but this line of research was just one of many threads tying together cosmology and particle physics. In the 1970s the particle physics community began to study several different aspects of the "poor man's accelerator," as Zel'dovich called the early universe. Following the consolidation of the Standard Model of particle physics in the mid 1970s, nearly every bit of data from accelerator experiments had fallen in line. The drive to understand physics beyond the Standard Model led to exorbitantly high energies: the relevant energy scales for Georgi and Glashow's $SU(5)$ GUT proposed in 1974 was $10^{15}\ GeV$, far beyond what would ever be accessible to earth-bound accelerators. Any sense that cosmology was too data-starved to compete with the precise science of accelerator physics was dispelled by a trio of young researchers well-versed in cosmology and particle physics. In 1977 Gary Steigman, David Schramm, and Jim Gunn argued that the number of lepton types had to be less than 5 for particle physics to be consistent with accounts of nucleosynthesis (Steigman et al. 1977). Unlike earlier cases of interaction between particle physics and cosmology, the three answered a fundamental problem in particle physics on the basis of cosmological constraints. In a time of decreasing support for ever-larger accelerators, the price tag of the poor man's accelerator must have been appealing; and Steigman, Schramm, and Gunn showed that even this bargain accelerator could be used to address fundamental issues.

The first intensive study of GUTs applied to the early universe focused on "baryogenesis." For a given GUT, one can directly calculate an observable feature of the early universe — the baryon-to-photon ratio usually denoted η — and in 1978 Motohiko Yoshimura argued that an SU(5) GUT predicted a value of η compatible with observations. Yoshimura (1978) kicked off a cottage industry focused on developing an account of baryogenesis similar in its quantitative detail to the account of light element nucleosynthesis. The account of baryogenesis has been widely hailed as one of the "greatest triumphs" of particle cosmology (Kolb and Turner 1990, 158).[25] Below I will focus on another aspect of GUTs in cosmology, the study of symmetry breaking and restoration in the early universe.

13.2.1 Symmetries: broken and restored

The understanding of symmetries in quantum field theory (QFT) changed dramatically in the 1960s due to the realization that field theories may exhibit spontaneous symmetry breaking (SSB). A typical one-line characterization of SSB is that "the laws of nature may possess symmetries which are not manifest to us because the vacuum

state is not invariant under them" (Coleman 1985, 116). Symmetry breaking in this loose sense is all too familiar in physics: solutions to a set of differential equations almost never share the full symmetries of the equations. The novel features of symmetry breaking in QFT arise as a result of a mismatch between symmetries of the Lagrangian and symmetries which can be implemented as unitary transformations on the Hilbert space of states $\mathcal{H}$. Roughly, systems for which a particular symmetry of the Lagrangian *cannot* be unitarily implemented on $\mathcal{H}$ exhibit SSB. This failure has several consequences: observables acquire non-invariant vacuum expectation values, and there is no longer a unique vacuum state. Physicists first studied symmetry breaking in detail in condensed matter systems displaying these features, but Yoichiro Nambu and others applied these ideas to problems in field theory starting in the early 1960s (see Brown and Cao 1991, Pickering 1984 for historical studies).

The introduction of SSB led to a revival of interest in gauge theories of the weak and strong interactions. Yang–Mills style gauge theories seemed to require massless gauge bosons (like the photon), in stark conflict with the short range of the weak and strong interactions. Adding mass terms for the gauge bosons directly to the Lagrangian would break its gauge invariance and, according to the conventional wisdom, render the theory unrenormalizable.[26] SSB garnered a great deal of attention in the early 1960s, but a general theorem due to Jeffrey Goldstone seemed to doom symmetry breaking in particle physics barely after its inception: SSB implies the existence of spin-zero massless bosons (Goldstone 1961; Goldstone et al. 1962).[27] Experiments ruled out such "Goldstone bosons," and there seemed to be no way to modify the particle interpretation of the theory to "hide" the Goldstone bosons along the lines of the Gupta–Bleuler formalism in QED.[28] Goldstone et al. (1962) concluded by reviewing the dim prospects for SSB; Weinberg added an epigraph from *King Lear* — "Nothing will come of nothing: speak again" — to indicate his dismay, which was (fortunately?) removed by the editors of *The Physical Review* (Weinberg 1980, 516). But there was a loophole: Goldstone's theorem does not apply to either discrete or local gauge symmetries.[29]

Philip W. Anderson was the first to suggest that breaking a gauge symmetry might cure the difficulties with Yang–Mills theory (by giving the gauge bosons mass) without producing Goldstone bosons. Anderson noted that this case may resemble condensed matter systems exhibiting SSB, in that the Goldstone bosons "become tangled up with Yang–Mills gauge bosons, and, thus, do not in any true sense really have zero mass" (Anderson 1963, 422; cf. Anderson 1958). He speculated that this "tangling" between Goldstone and gauge bosons could be exploited to introduce a massive gauge boson, but he supported these provocative remarks with neither a field theoretic model nor an explicit discussion of the gauge theory loophole in Goldstone's theorem. Within a year of Anderson's suggestive paper, Brout, Englert, Guralnik, Kibble and Higgs all presented field theoretic models in which gauge bosons acquire mass by "tangling" with Goldstone bosons (Englert and Brout 1964; Guralnik et al. 1964; Higgs 1964).

In the clear model presented by Peter Higgs, the massless Goldstone modes disappear from the physical particle spectrum, but in their ghostly gauge-dependent presence the vector bosons acquire mass.[30] Higgs began by coupling the simple scalar field of the Goldstone model with the electromagnetic interaction. Take a model in-

corporating a two component complex scalar field, such that $\phi = \frac{1}{\sqrt{2}}(\phi_1 - i\phi_2)$ with an effective potential

$$V(\phi) = \frac{1}{2}\lambda^2|\phi|^4 - \frac{1}{2}\mu^2|\phi|^2. \tag{13.5}$$

The effective potential includes all the terms in the Lagrangian other than the kinetic terms, and it represents the potential energy density of the quantum fields.[31] At first glance the second term appears to have the wrong sign; with the usual + sign, $V(\phi)$ has a unique global minimum at $\phi = 0$. The "incorrect" sign leads to degeneracy of the vacuum state; with a − sign, $V(\phi)$ has minima at $\phi_0 = \frac{\mu}{\lambda}$. Including the electromagnetic interaction leads to the following Lagrangian:

$$\mathcal{L} = (D_\mu\phi)^\dagger(D^\mu\phi) - V(\phi) - \frac{1}{4}F_{\mu\nu}F^{\mu\nu}, \tag{13.6}$$

where $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$, and D is the covariant derivative operator defined as $D_\mu = \partial_\mu + ieA_\mu$. Rewriting the effective potential $V(\phi)$ by expanding the field ϕ around the "true vacuum" ϕ_0 shows that the ϕ_1 field acquires a mass term whereas ϕ_2 is the massless "Goldstone boson." Higgs realized that a clever choice of gauge can be used to "kill" the latter component, which then appears *not* as a massless boson but instead as the longitudinal polarization state of a massive vector boson. The Lagrangian is invariant under the following gauge transformations:

$$\phi(x) \to e^{-i\theta(x)}\phi(x), \tag{13.7}$$

$$A_\mu \to A_\mu + \frac{1}{m}\partial_\mu\theta(x), \tag{13.8}$$

where m is a constant. The "Higgs mechanism" involves choosing a value of $\theta(x)$ to cancel the imaginary part of ϕ. This choice of $\theta(x)$ also effects the vector potential, leading to the following Lagrangian:

$$\mathcal{L} = (\partial_\mu\phi)(\partial^\mu\phi) + m^2\phi^2 A_\mu A^\mu - V(\phi) - \frac{1}{4}F_{\mu\nu}F^{\mu\nu}. \tag{13.9}$$

The vector field A_μ has acquired a mass term (the second term), as has the "Higgs boson" (although it is buried in the expression for $V(\phi)$), and the dreaded "Goldstone boson" has disappeared from the Lagrangian.

The Higgs mechanism could be used to fix and combine two appealing ideas, ridding both Yang–Mills style gauge theories and SSB of unwanted massless particles. Several theorists hoped that the trail blazed by Higgs et al. would lead to a gauge theory of the strong and weak interactions.[32] Three years after Higgs' paper, Weinberg incorporated the Higgs mechanism in a unified theory of the electromagnetic and weak interactions (Weinberg 1967), and a similar theory was introduced independently by Abdus Salam. These theories faced a roadblock, however: although several theorists suspected that such theories are renormalizable, they were not able to produce convincing arguments to that effect (Weinberg 1980, 518). Without a proof of renormalizability or direct experimental support the Salam–Weinberg idea drew little attention.[33]

Although theories with *unbroken* gauge symmetries were known to be renormalizable term-by-term in perturbation theory, it was not clear whether SSB would spoil renormalizability. Progress in the understanding of renormalization (due in large part to the Nobel Prize winning efforts of the Dutch physicists Gerard 't Hooft and Martinus Veltman) revealed that the renormalizability of a theory is actually *unaffected* by the occurrence of SSB. In his 1973 Erice lectures, Sidney Coleman advertised this as the main selling point of SSB (Coleman 1985, 139).

Testing the Higgs mechanism required a venture into uncharted territory. Although accelerator experiments carried out throughout the 1970s probed various aspects of the electroweak theory (see, e.g., Pickering 1984), they did little to constrain or elucidate the Higgs mechanism itself. Physicists continue to complain three decades later that the Higgs mechanism remains "essentially untested" (Veltman 2000, 348). Although the Higgs mechanism was the simplest way to reconcile a fundamentally symmetric Lagrangian with phenomenology, physicists actively explored alternatives such as "dynamical" symmetry breaking.[34] Indeed, treating the fundamentally symmetric Lagrangian as a formal artifact rather than imbuing it with physical significance was a live option. However, several physicists independently recognized that treating the Higgs mechanism as a description of a physical transition that occurred in the early universe, rather than as a bit of formal legerdemain, has profound consequences for cosmology. Weinberg emphasized at the outset that this line of research "may provide some sort of answer to the question" of "whether a spontaneously broken gauge symmetry should be regarded as a true symmetry" (Weinberg 1974b, 274).

In the condensed matter systems that originally inspired the concept of symmetry breaking, a variety of conditions (such as high temperature or large currents) lead to restoration of the broken symmetry. Based on a heuristic analogy with superconductivity and superfluidity, David Kirzhnits and his student Andrei Linde, both at the Lebedev Physical Institute in Moscow, argued that the vacuum expectation value ϕ_0 in a field theory with SSB varies with temperature according to $\phi_0^2(T) = \phi_0^2(T = 0) - c\lambda T^2$, where c and λ are non-zero constants (Kirzhnits 1972; Kirzhnits and Linde 1972). Symmetry restoration occurs above the critical temperature T_c, defined by $\phi_0^2(T_c) = 0$ (for $T > T_c$, $\phi_0(T)$ becomes imaginary). In the Weinberg model $\phi_0(0) \approx G^{1/2}$ (G is the weak interaction coupling constant), and (assuming that $c\lambda \approx 1$) Kirzhnits and Linde estimated that symmetry restoration occurs above $T_c \approx G^{-1/2} \approx 10^3\ GeV$. They concluded that the early universe underwent a transition from an initially symmetric state to the current broken symmetry state at the critical temperature, which corresponds to approximately 10^{-12} seconds after the big bang in the standard hot big bang model.

Within two years Kirzhnits and Linde and a group of Cambridge (Massachusetts) theorists had developed more rigorous methods based on finite-temperature field theory to replace this heuristic argument.[35] Finite-temperature field theory includes interactions between quantum fields and a background thermal heat bath at a temperature T.[36] These more detailed calculations showed that, roughly speaking, symmetry restoration occurs as a consequence of the temperature dependence of quantum corrections to the effective potential. The full effective potential includes a zero-temperature term along with a temperature-dependent term, $\bar{V}(\phi, T)$. Symmetry breaking occurs

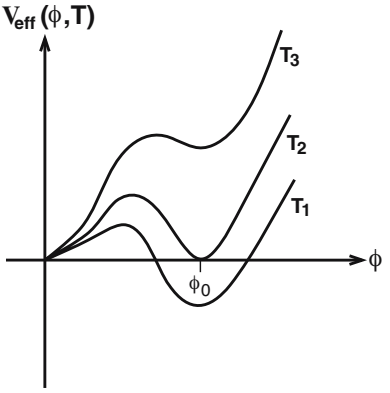

Fig. 13.2. This figure illustrates the temperature dependence of the effective potential of the Higgs field $V_{eff}(\phi, T)$ in the Weinberg–Salam model. T_2 is the critical temperature (approximately 10^{14} GeV), and $T_3 > T_2 > T_1$.

in a theory with $V(\phi) = \frac{1}{2}\lambda^2|\phi|^4 + \frac{1}{2}\mu^2|\phi|^2$, for example, if $\bar{V}(\phi, T)$ includes a mass correction that changes the sign of the second term above a critical temperature. Whether symmetry restoration occurs depends upon the nature of $\bar{V}(\phi, T)$ and the zero temperature effective potential in a particular model.[37] In the Weinberg–Salam model (with suitable choices for coupling constants), the global minimum at $\phi = 0$ for temperatures above the critical temperatures develops into a local minimum with the true global minimum displaced to ϕ_0 (see Figure 13.2 for an example). Determining the nature and consequences of such phase transitions drew an increasing number of particle physicists into the study of early universe cosmology throughout the 1970s, as we will see in Section 2.3. But before continuing with the discussion of this line of research, I will briefly turn to more speculative uses of SSB in cosmology.

13.2.2 Conformal Symmetry Breaking

By the late 1970s symmetry breaking was an essential piece in the field theorists' technical repertoire, and its successful use in electroweak unification and the development of the Standard Model encouraged more speculative variations on the theme. The "Brussels Consortium" (as I will call Robert Brout, François Englert, and their various collaborators) described the origin of the universe as SSB of conformal symmetry, but this imaginative line of research led to an increasingly rococo mess rather than a well constrained model. At roughly the same time, Anthony Zee developed an account of gravitational symmetry breaking motivated by the desire to formulate a "unified" gravitational theory with no dimensional constants other than the mass term of a fundamental scalar field.

Like their countryman Lemaître decades earlier, the Brussels Consortium focused on a quantum description of the "creation" event itself. Brout et al. (1978) aimed to replace "the 'big bang' hypothesis of creation—more a confession of desperation

and bewilderment than the outcome of logical argumentation" with an account of the "spontaneous creation of all matter and radiation in the universe. [...] The big bang is replaced by the fireball, a rational object subject to theoretical analysis" (Brout et al. 1978, 78). As with Tyron's (1973) earlier proposal, this account of spontaneous creation did not violate conservation of energy. Their theoretical analysis builds on an alleged "deep analogy" between relativistic cosmology and conformally invariant QFT, which in practice involves two fundamental assumptions.[38] First, the Consortium assumes that the universe must be described by a conformally flat cosmological model, which implies that the metric for any cosmological model is related to Minkowksi space-time by $g_{ab} = \phi^2(x^i)\eta_{ab}$, where η_{ab} is the Minkowski metric.[39] The conformal factor $\phi(x^i)$ is treated as a massless scalar field conformally coupled to gravitation. Second, a fluctuation of $\phi(x^i)$, which breaks the conformal symmetry of the pristine initial state (constant $\phi(x^i)$ in a background Minkowski space-time), bears the blame for the creation of the universe.

The devil is in providing the details regarding the outcome of the "rational fireball" triggered by such a modest spark. The Consortium's original script runs as follows: the fluctuation initially produces a de Sitter-like bubble, with the expansion driven by an effective equation of state with negative pressure. This equation of state is due to particle creation via a "cooperative process": the initial fluctuation in $\phi(x^i)$ perturbs the gravitational field; variations in the gravitational field produce massive scalar particles; the particles create fluctuations in the gravitational field; and so on. Eventually the cooperation ends, and the primeval particles decay into matter and radiation as the universe slows from its de Sitter phase into FLRW expansion. Although the details of these processes are meant to follow from the fundamental assumptions, a number of auxiliary conditions are needed to insure that the story culminates with something like our observed universe. The evolution of the Consortium's program belies the malleability of the underlying physics: Brout et al. replace the earlier idea regarding "cooperative processes" with the suggestion that particle production is a result of a "phase transition in which the 'edge of the universe' is the boundary wall between two phases" (Brout et al. 1980, 110).

Despite these difficulties, the Consortium often attributed a great deal of importance to their "solution" of the "causality problem." The basis for this solution was buried in an Appendix of Brout et al. (1978), but mentioned more prominently in later papers, including the title of Brout et al. (1979) — "The Causal Universe." Brout et al. (1978) note that in their model the integral in equation (13.3) diverges. There are no horizons. But there is also no pressing horizon *problem* in Misner's sense: conformal symmetry is stipulated at the outset, so there is simply no need to *explain* the early universe's uniformity via causal interactions. However, the absence of horizons is still taken to solve the "causality problem," in the sense that the universe and all its contents can ultimately be traced back to a simple single cause, the initial fluctuation of $\phi(x^i)$. Whatever the appeal of this solution, the Consortium ultimately failed to develop a believable model that realized their programmatic aims. However, the Princeton theorist J. Richard Gott III developed a variation of the Consortium's idea that would eventually lead to the development of "open inflation" models (Gott 1982).

Anthony Zee also solved the horizon problem with a variation on the theme of SSB. Zee (1979, 1980) proposed that incorporating symmetry breaking into gravitational theory (by coupling gravitation to a scalar field) leads to replacing the gravitational constant G with $(\epsilon\phi_v^2)^{-1}$, where ϵ is a coupling constant and ϕ_v is the vacuum expectation value of the scalar field.[40] If the potential (and the minima) of this field varies with temperature, then the gravitational "constant" varies as well. Zee (1980) argues that $\phi^2 \approx T^2$ at high temperatures, so that $G \propto 1/T^2$. This alters the FLRW dynamics so that $a(t) \propto t$; and it will come as no suprise that the integral in equation (13.3) diverges as a result. According to Guth's recollections (Guth 1997, 180–81), a lunchtime discussion of Zee's paper in the SLAC cafeteria led him to consider the implications of his own ideas for horizons.

13.2.3 Phase Transitions

The study of early universe phase transitions held out the promise of deriving stringent observational constraints from the cosmological setting for aspects of particle physics far beyond the reach of accelerators. Throughout the 1970s physicists studied three different types of consequences of symmetry breaking phase transitions: (1) effects due to the different nature of the fundamental forces prior to the phase transition, (2) defect formation during the phase transition, (3) effects of the phase transition on cosmological evolution. As we will see below, initial results ran the gamut from disastrous conflict with observational constraints to a failure to find any detectable imprint.

The first type of effect drew relatively little attention. Kirzhnits (1972); Kirzhnits and Linde (1972) briefly mentioned the possible consequences of long-range repulsive forces in the early universe. Prior to the electroweak phase transition any "weak charge" imbalance would result in long-range repulsive forces, and according to Kirzhnits and Linde such forces would render both a closed, positive curvature model and an isotropic, homogeneous model "impossible" (Kirzhnits and Linde 1972, 474).[41] By way of contrast, a group of CERN theorists suggested that interactions at the GUT scale would help to *smooth* the early universe. Ellis et al. (1980) consider the possibility that a "grand unified viscosity" would effectively insure isotropization prior to a symmetry breaking phase transition; they conclude that although these interactions damp some modes of an initial perturbation spectrum, they will not smooth a general anisotropic cosmological model.

The study of defect formation in the early universe was a much more fruitful line of research. An early study of CP-symmetry breaking (Zel'dovich et al. 1975) showed that the resulting inhomogeneity (with energy density concentrated in domain walls) would be far too large to fit observational constraints.[42] But Zel'dovich et al. (1975) also calculated the equation of state for this "cellular medium" (averaged over a volume containing both domain walls and the empty cells), and remarked that evolution dominated by matter in this state might solve the horizon problem.[43] The authors did not highlight this point (it was not mentioned in the introduction, abstract, or conclusion); their main interest was to establish that cosmology rules out discrete symmetry breaking, in itself a remarkable constraint on particle physics.

Later work on the formation of defects in theories with SSB of local gauge symmetries also ran afoul of observational constraints. Tom Kibble, an Indian-born British physicist at Imperial College, established a particularly important result (Kibble 1976): defect formation depends on the topological structure of the vacuum solutions to a particular field theory, and is thus relatively independent of the details of the phase transition. Roughly, defects result from the initial domain structure of the Higgs field, which Kibble argued should be uncorrelated at distances larger than the particle horizon at the time of the phase transition. This complicated domain structure disappears if the Higgs field in different regions becomes "aligned," but in some cases no continuous evolution of the field can eliminate all nonuniformities; topological defects are the resulting persistent structures. Kibble (1976) noted that point-like defects (called monopoles and previously studied by 't Hooft 1974; Polyakov 1974) might form, but thought that they would "not be significant on a cosmic scale." However, given the absence of any natural annihilation mechanism, Zel'dovich and Khlopov (1978); Preskill (1979); Einhorn et al. (1980) established a dramatic conflict between predicted monopole abundance and observations: in Preskill's calculation, monopoles alone would contribute a mass density 10^{14} times greater than the *total* estimated mass density![44]

The resolution of this dramatic conflict would ultimately come from considerations of the third type of effect. Linde, Veltman and Joseph Dreitlein at the University of Colorado independently realized that a non-zero $V(\phi)$ would couple to gravity as an effective Λ term.[45] Linde (1974) argued that although earlier particle physics theories "yielded no information" on the value of Λ (following Zel'dovich, he held that Λ is fixed only up to an arbitrary constant), theories incorporating SSB predicted a tremendous shift – 49 orders of magnitude – in $V(\phi)$ at the critical temperature T_c.[46] However, this dramatic change in the cosmological "constant" would apparently have little impact on the evolution of the universe (Linde 1974, 183):[47]

> To be sure, almost the entire change [of Λ] occurs near $T_c = 10^{15} - 10^{16}$ deg. In this region, the vacuum energy density is lower than the energy density of matter and radiation, and therefore the temperature dependence of Λ does not exert a decisive influence on the initial stage of the evolution of the universe.

Linde implicitly assumed that the phase transition was second-order, characterized by a transition directly from one state to another with no intermediate stage of "mixed" phases.[48] Unlike Linde, Veltman (1974) regarded the idea that an arbitrary constant could be added to the vacuum energy density to yield a current value of $\Lambda \approx 0$ as "ad hoc" and "not very satisfactory." Veltman took the "violent" disagreement with observational constraints on Λ and the value calculated using the electroweak theory as one more indicator that the Higgs mechanism is "a cumbersome and not very appealing burden" (Veltman 1974, 1).[49] Dreitlein (1974) explored one escape route: an *incredibly* small Higgs mass, on the order of $2.4 \times 10^{-27} MeV$, would lead to an effective Λ close enough to 0. Veltman (1975) countered that such a light Higgs particle would mediate long-range interactions that should have already been detected. In sum, these results were thoroughly discouraging: Veltman had highlighted a discrepancy between calculations of the vacuum energy in field theory and cosmological constraints that would come to be called the "cosmological constant problem" (see

Rugh and Zinkernagel 2002). Even for those willing to set aside this issue and focus only on the shift in vacuum energy, there appeared to be "no way cosmologically to discriminate among theories in which the symmetry is spontaneously broken, dynamically broken, or formally identical and unbroken" (to quote Bludman and Ruderman 1977, 255).

By the end of the 1970s several physicists had discovered that this conclusion does not hold if the Higgs field became trapped in a "false vacuum" state (with $V(\phi) \neq 0$). Demosthenes Kazanas, an astrophysicist working at Goddard Space Flight Center, clearly presented the effect of persistent vacuum energy (Kazanas 1980): the usual FLRW dynamics is replaced with a phase of exponential expansion. He also clearly stated an advantage of incoroporating such a phase (L62):

> Such an exponential expansion law occurring in the very early universe can actually allow the size of the causally connected regions to be many orders of magnitude larger than the presently observed part of the universe, thus potentially accounting for its observed isotropy.

But it was not clear how to avoid an undesirable consequence of a first-order phase transition, namely the production of large inhomogeneities due to the formation of "bubbles" of the new "true" vacuum phase immersed in the old phase. Linde and Chibisov (Linde 1979, 433–34) explored the possibility of combining Zel'dovich's "cold universe" idea with a first-order phase transition, but they did not see a way to avoid excessive inhomogeneity.[50] During a stay at NORDITA in Copenhagen, the Japanese astrophysicist Katsuhiko Sato studied first-order phase transitions in considerable detail, focusing on the consequences of a stage of exponential expansion driven by a false vacuum state. Sato (1981) derived constraints on various parameters, such as the rate of bubble formation and coupling constants.[51] Although Sato appears to have been optimistic that these constraints could be met, a slightly later collaborative paper with the University of Michigan theorist Martin Einhorn (Einhorn and Sato 1981) ended on a skeptical note (401):[52]

> We have seen that most of the difficulties with the long, drawn-out phase transition discussed in Section V stems [sic] from the exponential expansion of the universe. This was due to the large cosmological constant. If a theory could be developed in which the vacuum did not gravitate, i.e., a theory of gravity which accounts for the vanishing cosmological constant term in a natural way, then the discussion would be drastically changed. Although scenarios have been developed in which the effect of the cosmological constant term remains small for all times, we would speculate that the problem here is less the choice of GUT but rather reconciling gravity with quantum field theory.

To avoid the unpalatable consequences of a first order phase transition Einhorn and Sato were willing to abandon the starting point of this entire line of thought.[53]

By the time these papers appeared in print, the young American physicist Alan Guth had presented an argument that an "inflationary" stage is a desirable consequence of an early universe phase transition, rather than a source of difficulties. After

persistent lobbying from his friend and collaborator Henry Tye, Guth undertook serious study of GUTs in the summer of 1979, focused on production of monopoles in the early universe (Guth 1997, chapter 9). Tye and Guth discovered that a first-order transition could alleviate the monopole problem: within each bubble produced in a first-order transition, the Higgs field is uniform. Monopoles would only be produced at the boundaries between the bubbles as a consequence of bubble wall collisions. Thus the abundance of monopoles ultimately depends upon the nucleation rate of the bubbles. Guth and Tye (1980) argued that reasonable models of the phase transition have a low nucleation rate, leading to a tolerably low production of monopoles. Einhorn and Sato (1981) highlighted various difficulties with this proposal, commenting that "although it is *possible* to meet the necessary requirements, it is unclear whether this scenario is *natural* in the sense that it may require fortuitious relationships between the magnitude of the gauge coupling and the parameters of the Higgs potential" (Einhorn and Sato 1981, 385) and noting the difficulties associated with a phase of exponential expansion. Shortly after Guth and Tye (1980) was submitted, Guth independently discovered that the equation of state for the Higgs field trapped in a "false vacuum" state drives exponential expansion. In short order, he discovered several appealing features of what he called, alluding to economic worries at the end of Carter's presidency, an "inflationary universe."

13.2.4 Guth's "Spectacular Realization"

Guth modestly concluded as follows (Guth 1981, 354):

> In conclusion, the inflationary scenario seems like a natural and simple way to eliminate both the horizon and flatness problems. I am publishing this paper in the hope that it will highlight the existence of these problems and encourage others to find some way to avoid the undesirable features of the inflationary scenario.

To say that Guth's paper (and the series of lectures he gave before and after it appeared) achieved these goals would be a dramatic understatement. This success stemmed not from fundamentally new physics, but from the clear presentation of a rationale for pursuing the idea of inflation. Even those who had been aware of the work discussed above, such as Martin Rees, have commented that they only understood it in light of Guth's paper.[54] Guth's paper significantly upped the explanatory ante for early universe cosmology: he showed that several apparently independent features of the universe could be traced to a common source, an early stage of inflationary expansion. This effectively set a new standard for theory choice in early universe cosmology. The situation resembles several other historical episodes in which a significant success set new standards. Einstein's accurate prediction of the anomaly in Mercury's perihelion motion raised the bar for gravitational theories: although the perihelion motion was not regarded as a decisive check prior to his prediction, it subsequently served as a litmus test for competing theories of gravitation. Similarly, following Guth's paper the ability to solve these problems served as an entrance requirement.

To my knowledge Guth was the first to explicitly recognize the connection between an inflationary stage and a puzzling balance between the initial expansion rate and energy density. Guth's work notebook dated Dec. 7, 1979 begins with the following statement highlighted in a double box: "SPECTACULAR REALIZATION: This kind of supercooling can explain why the universe today is so incredibly flat—and therefore resolve the fine-tuning paradox pointed out by Bob Dicke." [55] Dicke's paradox highlights an odd feature of the density paramter Ω. Using the Friedmann equation, we can write Ω as follows:[56]

$$\Omega := \frac{8\pi G}{3H^2}\rho = \left(1 - \frac{3k}{8\pi G\rho}\right)^{-1}. \tag{13.10}$$

During expansion ρ scales as $\propto a^{-3}$ for normal matter and $\propto a^{-4}$ for radiation. Thus, if the value of Ω initially differs from 1, it evolves rapidly away from 1; the value $\Omega = 1$ is an unstable fixed point under dynamical evolution. For the observed universe to be anywhere close to $\Omega = 1$ (as it appears to be), the early universe must have been *incredibly* close to the "flat" FLRW model ($\Omega = 1, k = 0$). Guth discovered that during exponential expansion Ω is driven rapidly *towards* 1; ρ is a constant for a false vacuum state, so Ω approaches 1 as a^{-2} during inflation. If the universe expands by a factor $Z \geq 10^{29}$, where $Z =: e^{\chi \Delta t}$ and Δt is the duration of the inflationary stage, then $\Omega_0 = 1$ to extremely high precision, for nearly any pre-inflationary "initial value" of Ω.

Unlike the horizon problem, the flatness problem was not widely acknowledged as a legitimate problem prior to Guth's paper. In an appendix added to "convince some skeptics," Guth comments that (Guth 1981, 355):

> In the end, I must admit that questions of plausibility are not logically determinable and depend somewhat on intuition. Thus I am sure that some physicists will remain convinced that there really is no flatness problem. However, I am also sure that many physicists agree with me that the flatness of the universe is a peculiar situation which at some point will admit a physical explanation.

Whether or not this argument swayed many physicists, several of the interviewees in Lightman and Brawer (1990) made remarks similar to Misner's (Lightman and Brawer 1990, 240):

> I didn't come on board thinking that paradox [Dicke's flatness paradox] was serious until the inflationary models came out. [...] The key point for me was that inflation offers an explanation. Even if it's not the right explanation, it shows that finding an explanation is a proper challenge to physics.

The existence of a proposed solution to the flatness problem lent it an air of legitimacy; the universe's flatness had been previously regarded as puzzling (Dicke and Peebles 1979), but following Guth's paper it was widely interpreted as a telling sign of an early inflationary stage.

Several proposals discussed above implied that horizons would disappear, as the horizon distance in equation (13.3) diverges. A transient inflationary phase increases

the horizon distance by a factor of Z; for $Z > 5 \times 10^{27}$ the "horizon problem disappears" in the sense that the horizon length at the time of the emission of the background radiation approaches the current visual horizon. Particle horizons don't disappear, but they are stretched enough to encompass the visible universe. Guth stressed the striking difference between initial conditions needed in the inflationary universe and the standard cosmology (Guth 1981, 347): for the standard cosmology, "the initial universe is assumed to be homogeneous, yet it consists of at least $\approx 10^{83}$ separate regions which are causally disconnected." For an inflationary period with sufficiently large Z, a single homogeneous pre-inflationary patch of sub-horizon scale expands to encompass the observed universe.

Despite these successes, Guth's original proposal did not solve the transition problem. As Einhorn and Sato (1981) had argued, bubbles of new phase formed during the phase transition do not percolate, i.e., they do not join together to form large regions of the same phase. The energy released in the course of the phase transition is concentrated in the bubble walls, leading to an energy density far too high near the bubble walls and far too low in the interior. Frequent bubble collisions would be needed to smooth out the distribution of energy so that it is compatible with the smooth beginning of an FLRW model.[57] The phase transition never ends, in the sense that large volumes of space remain "stuck" in the old phase, with vast differences in the energy density between these regions and the bubble walls. In summary, a first-order phase transition appropriate for inflation also produces a universe marred by the massive inhomogeneities due to the formation of bubbles, rather than the smooth early universe required by observations.

The solution to the transition problem led to difficulties with Guth's original identification of the Higgs field of an $SU(5)$ GUT as the source of an inflationary stage. Briefly, Albrecht and Steinhardt (1982); Linde (1982) both developed models of the phase transition based on a Coleman–Weinberg effective potential for the Higgs field. In these new models the inflationary expansion persists long enough that the initial bubble is much, much larger than the observed universe; within this single bubble the matter and radiation density needed for the big bang model is generated via decay of the Higgs field. Within a year theorists had turned to implementing Chibisov's (1981) idea that small fluctuations stretched during inflation would serve as the seeds for galaxy formation. The intense work on structure formation during the Nuffield workshop, a conference held in Cambridge from June 21–July 9, 1982, led to the "death and transfiguration" of inflation (from the title of the conference review in *Nature*, Barrow and Turner 1982). Inflation "died" since detailed calculations of the density perturbations produced during an inflationary era indicated that an $SU(5)$ Higgs field could not drive inflation, as originally thought. The "transfiguration" of the field involved a significant shift in methodology: the focus shifted to implementing inflation successfully rather than treating it as a consequence of independently motivated particle physics. In his recollections of the Nuffield conference, Guth wrote:

> [A] key conclusion of the Nuffield calculations is that the field which drives inflation cannot be the same field that is responsible for symmetry breaking. For the density perturbations to be small, the underlying particle theory must

> contain a new field, now often called the *inflaton* field [...], which resembles the Higgs field except that its energy density diagram is much flatter. (Guth 1997, 233–34)

The "inflaton" may resemble the Higgs, but the rules of the game have changed: it is a new fundamental field distinct from any scalar field appearing in particle physics.

The explosion of research interest in inflationary cosmology in the early 1980s attests to its appeal. Inflation allowed theorists to replace several independent features of the initial conditions — overall uniformity, flatness, lack of monopoles and other relics, and the presence of small scale fluctuations — with a theoretical entity they knew how to handle: the effective potential of a fundamental scalar field.[58] The discussion of earlier proposals highlights an important advantage of inflation: the Higgs mechanism is a central component of the Weinberg–Salam model and of GUTs, which provided a rich source of ideas for further refinements of inflation. Starobinsky drew on the more esoteric subject of quantum corrections to the stress-energy tensor in semi-classical quantum gravity, and the other proposals discussed above required a number of bald stipulations. Inflation still has not solved the source problem, in the sense that there is still no canonical identification of the "inflaton" field with a particular scalar field. The fertile link with particle physics has instead produced an embarassment of riches: inflation has been implemented in a wide variety of models, to such an extent that cosmologists have sometimes complained of the difficulty in coining a name for a new model.

In closing, I should emphasize an important difference between inflation and other cases of "upping the explanatory ante." Prior to Einstein's work, astronomers agreed that there was a discrepancy between the observed perihelion motion of Mercury and Newtonian calculations, although this was not seen as a telling failure of Newtonian theory. By way of contrast, several critics of inflation have not been convinced that inflation has cured *genuine* explanatory deficiencies of the standard big bang model.[59] Intellectual descendants of Ludwig Boltzmann such as Roger Penrose (see, in particular Penrose 1979, 1989) *expect* the universe to be in an initially "improbable" state, which is ultimately responsible for the second law of thermodynamics and the arrow of time. Special initial conditions play the crucial role of insuring that the observed universe has an arrow of time; they are not something to be avoided by introducing new dynamics that "washes away" the dependence on an initial state. Two of the proposals above also did not take this approach to "erasing" the singularity: Starobinsky accepted that his proposal would require stipulating that the early universe began in an early de Sitter state, and the Brussels Consortium aimed to develop an account of the creation event itself. In developing theories of the early universe, the methodological strategy exemplified by inflation was by no means mandatory.

13.3 Conclusions

In the epilogue of their recent textbook, Kolb and Turner (1990) contrast the adventurous attitude of their contemporaries with those of earlier cosmologists, commenting that (Kolb and Turner 1990, 498):

> Whatever future cosmologists write about cosmology in the decades following the discovery of the CMBR, we can be certain they will not criticize contemporary cosmologists for failure to take their theoretical ideas — and sometimes wild speculations — seriously enough.

Following a story of speculative theories regarding the universe at $t \approx 10^{-35}$ *s* after the big bang, it is easy to agree with their assessment. As I have described above, various problems and opportunities led cosmologists to develop theories of the early universe. The incredible extrapolations to the early universe allowed theorists to grapple with issues that have no bearing on more directly accessible phenomena, including the creation of particles in strong gravitational fields and the predictions of symmetry restoration at incredibly high temperatures. Many theoretical roads led to the consideration of an early de Sitter phase, and all faced the difficulties of identifying a believable physical source driving the de Sitter expansion and accounting for the transition to customary big bang expansion. Guth's seminal work on inflation did not introduce new physics, and did not solve these problems, but it did provide a rationale that has done much to underwrite the adventurous optimism characterizing the field.

Acknowledgments

I would like to thank John Earman, Al Janis, Michel Janssen, David Kaiser, John Norton, and Laura Ruetsche, for helping in various ways to make this a better paper. My research was supported in part by the NSF under grant SES 0114760.

References

Aitchison, Ian J. R. (1982). *An informal Introduction to Gauge Field Theories*. Cambridge University Press.

Albrecht, Andreas and Steinhardt, Paul (1982). Cosmology for grand unified theories with induced symmetry breaking. *Physical Review Letters* **48**, 1220–1223.

Anderson, Philip W. (1963). Plasmons, gauge invariance, and mass. *Physical Review* **130**, 439–442.

— (1958). Coherent excited states in the theory of superconductivity: Gauge invariance and the Meissner effect. *Physical Review* **110**, 827–835.

Barrow, John D. and Turner, Michael S. (1982). The inflationary universe – birth, death, and transfiguration. *Nature* **298**, 801–805.

Bekenstein, Jacob D. (1975). Nonsingular general-relativistic cosmologies. *Physical Review D* **11** , 2072–2075.

Belinskii, V. A., Khalatnikov, I. M. and Lifshitz, E. M. (1974). General solutions of the equations of general relativity near singularities. In *Confrontation of Cosmological Theories with Observational Data*, M. Longair, ed. No. 63 in IAU Symposium, D. Reidel, Dodrecht, 261–275.

Bernard, Claude W. (1974). Feynman rules for gauge theories at finite temperature. *Physical Review D* **9**, 3312–3320.

Birrell, Neil C. and Davies, Paul C. W. (1982). *Quantum Fields in Curved Space*. Cambridge University Press, Cambridge.

Bludman, Sidney A. and Ruderman, Malvin A. (1977). Induced cosmological constant expected above the phase transition restoring the broken symmetry. *Physical Review Letters* **38**, 255–257.

Brout, R., Englert, F. and Gunzig, E. (1980). Spontaneous symmetry breaking and the origin of the universe. In *Gravitation, Quanta, and the Universe*, A. R. Prasanna, Jayant V. Narlikar and C. V. Vishveshwara, eds., 110–118.

— (1979). The causal universe. *General Relativity and Gravitation* **10**, 1–6.

Brout, Robert, Englert, François and Gunzig, Edgard (1978). The creation of the universe as a quantum phenomenon. *Annals of Physics* **115**, 78–106.

Brown, Laurie M. and Cao, Tian Yu (1991). Spontaneous breakdown of symmetry: its rediscovery and integration into quantum field theory. *Historical Studies in the Physical and Biological Sciences* **21**, 211–235.

Cao, Tian Yu (1997). *Conceptual Developments of 20th Century Field Theories*. Cambridge University Press, Cambridge.

Coleman, Sidney (1985). *Aspects of Symmetry*. Cambridge University Press. Selected Erice lectures.

Criss, Thomas, Matzner, Richard, Ryan, Michael and Shepley, Louis (1975). Modern theoretical and observational cosmology. In *General Relativity and Gravitation*, Shaviv and Rosen, eds. 7, New York: John Wiley & Sons, 33–108.

de Sitter, Willem (1931). The expanding universe. *Scientia* **49**, 1–10.

Dicke, Robert and Peebles, P. J. E. (1979). The big bang cosmology–enigmas and nostrums. In *General relativity: an Einstein centenary survey*, S. W. Hawking and W. Israel, eds., Cambridge University Press, Cambridge, 504–517.

Dolan, Louise and Jackiw, Roman (1974). Symmetry behavior at finite temperature. *Physical Review D* **9**, 3320–3341.

Dreitlein, Joseph (1974). Broken symmetry and the cosmological constant. *Physical Review Letters* **20**, 1243–1244.

Earman, John (1995). *Bangs, Crunches, Whimpers, and Shrieks*. Oxford University Press, Oxford.

— (1999). The Penrose-Hawking singularity theorems: History and implications. In *The Expanding Worlds of General Relativity*, Jürgen Renn, Tilman Sauer and Hubert Gönner, eds. No. 7 in Einstein Studies, Birkhäuser Boston, 235–270.

— (2001). Lambda: The constant that refuses to die. *Archive for the History of the Exact Sciences* **55**, 189–220.

Earman, John and Mosterin, Jesus (1999). A critical analysis of inflationary cosmology. *Philosophy of Science* **66**, 1–49.

Eddington, Arthur S. (1933). *The Expanding Universe*. MacMillan, New York.

Einhorn, Martin B. and Sato, Katsuhiko (1981). Monopole production in the very early universe in a first order phase transition. *Nuclear Physics* **B180**, 385–404.

Einhorn, Martin B., Stein, Daniel L. and Toussaint, Doug (1980). Are grand unified theories compatible with standard cosmology? *Physical Review D* **21**, 3295–3298.

Einstein, Albert (1917). Kosmologische betrachtungen zur allgemeinen relativitätstheorie. *Preussische Akademie der Wissenschaften (Berlin). Sitzungs-*

berichte, 142–152. Reprinted in translation in: *Principle of Relativity*, Lorentz et al. eds., Dover, New York, 1923.

Eisenstaedt, Jean (1989). The early interpretation of the Schwarzschild solution. In *Einstein and the History of General Relativity*, John Stachel and Don Howard, eds., Vol. 1 of *Einstein Studies*. Birkhäuser Boston, 213–33.

Ellis, George F. R. and Rothman, Tony (1993). Lost horizons. *American Journal of Physics* **61**, 883–893.

Ellis, John, Gaillard, Mary K. and Nanopoulos, Dimitri V. (1980). The smoothness of the universe. *Physics Letters B* **90**, 253–257.

Englert, François and Brout, Robert (1964). Broken symmetry and the mass of gauge vector mesons. *Physical Review Letters* **13**, 321–23.

Farhi, Edward and Jackiw, Roman, eds. (1982). *Dynamical Gauge Symmetry Breaking: A collection of reprints*. World Scientific, Singapore.

Gibbons, Gary W. and Hawking, Stephen W. (1977). Cosmological event horizons, thermodynamics, and particle creation. *Physical Review D* **15**, 2738–51.

Gliner, Erast B. (1966). Algebraic properties of the energy-momentum tensor and vacuum-like states of matter. (Translated by W. H. Furry.) *Soviet Physics JETP* **22**, 378–382.

— (1970). The vacuum-like state of a medium and Friedman cosmology. *Soviet Physics Doklady* **15**, 559–561.

Gliner, Erast B. and Dymnikova, Irina G. (1975). A nonsigular Friedmann cosmology. *Soviet Astronomy Letters* **1**, 93–94.

Goldstone, Jeffrey (1961). Field theories with 'superconductor' solutions. *Nuovo Cimento* **19**, 154–164.

Goldstone, Jeffrey, Salam, Abdus and Weinberg, Steven (1962). Broken symmetries. *Physical Review* **127**.

Gott, J. Richard (1982). Creation of open universes from De Sitter space. *Nature* **295**, 304–307.

Grib, Andrej A., Mamayev, S. G. and Mostepanenko, V. M. (1984). Self-consistent treatment of vacuum quantum effects in isotropic cosmology. In Markov and West (1984), 197–212, 197–212. Proceedings of the second Seminar on Quantum Gravity; Moscow, October 13–15, 1981.

Guralnik, G. S., Hagen, C. Richard and Kibble, T. W. B. (1964). Global conservation laws and massless particles. *Physical Review Letters* **13**, 585–587.

Guralnik, Gerald S., Hagen, C. Richard and Kibble, Thomas W. B. (1968). Broken symmetries and the Goldstone theorem. In *Advances in Particle Physics*, R. L. Cool and R. E. Marshak, eds., vol. 2. 567–708.

Gurevich, L. E. (1975). On the origin of the metagalaxy. *Astrophysics and Space Science* **38**, 67–78.

Guth, Alan (1981). Inflationary universe: A possible solution for the horizon and flatness problems. *Physical Review D* **23**, 347–56.

— (1997). *The Inflationary Universe*. Addison-Wesley, Reading, MA.

Guth, Alan and Tye, S.-H. Henry (1980). Phase transitions and magnetic monopole production in the very early universe. *Physical Review Letters* **44**, 631–34.

Guth, Alan H. and Weinberg, Erick J. (1983). Could the universe have recovered from a slow first order phase transition? *Nuclear Physics* **B212**, 321.

Hagedorn, Rolf (1970). Thermodynamics of strong interactions at high energy and its consequences for astrophysics. *Astronomy and Astrophysics* **5**, 184–205.

Hawking, Stephen W. (1970). Conservation of matter in general relativity. *Communications in Mathematical Physics* **18**, 301–306.

Hawking, Stephen W. and Ellis, George F. R. (1968). The cosmic black-body radiation and the existence of singularities in our universe. *Astrophysical Journal* **152**, 25–36.

Higgs, Peter W. (1964). Broken symmetries, massless particles, and gauge fields. *Physical Review Letters* **12**, 132–133.

Janssen, Michel (2002). Explanation and evidence: COI stories from Copernicus to Hockney. *Perspectives in Science* **10**, 457–552.

Kazanas, Demosthenes (1980). Dynamics of the universe and spontaneous symmetry breaking. *Astrophysical Journal Letters* **241**, L59–L63.

Kibble, Thomas W. B. (1976). Topology of cosmic domains and strings. *Journal of Physics* **A9**, 1387–97.

Kirzhnits, David A. (1972). Weinberg model in the hot universe. *JETP Letters* **15**, 529–531.

Kirzhnits, David A. and Linde, Andrei (1972). Macroscopic consequences of the Weinberg model. *Physics Letters B* **42**, 471–474.

Kolb, Edward W. and Turner, Michael S. (1990). *The early universe*, vol. 69 of *Frontiers in Physics*. Addison-Wesley, New York.

Kolb, Edward W. and Wolfram, Stephen (1980). Spontaneous symmetry breaking and the expansion rate of the early universe. *Astrophysical Journal* **239**, 428.

Lapchinksy, V. G., Nekrasov, V. I., Rubakov, V. A. and Veryaskin, A. V. (1984). Quantum field theories with spontaneous symmetry breaking in external gravitational fields of cosmological type. In Markov and West (1984), 213–230, 213–230. Proceedings of the second Seminar on Quantum Gravity; Moscow, October 13–15, 1981.

Lemaître, Georges (1934). Evolution of the expanding universe. *Proceedings of the National Academy of Science* **20**, 12–17.

Lightman, Alan and Brawer, Roberta (1990). *Origins: The Lives and Worlds of Modern Cosmologists*. Harvard University Press, Cambridge.

Linde, Andrei (1974). Is the Lee constant a cosmological constant? *Soviet Physics JETP* **19**, 183–184.

— (1979). Phase transitions in gauge theories and cosmology. *Reports on Progress in Physics* **42**, 389–437.

— (1982). A new inflationary universe scenario: a possible solution of the horizon, flatness, homogeneity, isotropy, and primordial monopole problems. *Physics Letters B* **108**, 389–393.

Lindley, David (1985). The inflationary universe: A brief history. Unpublished manuscript.

Markov, M. A. and West, P. C., eds. (1984). Proceedings of the second Seminar on Quantum Gravity; Moscow, October 13–15, 1981. *Quantum Gravity*. Plenum Press, New York.

Misner, Charles W. (1968). The isotropy of the universe. *Astrophysical Journal* **151**, 431–457.
— (1969). Mixmaster universe. *Physical Review Letters* **22**, 1071–1074.
Misner, Charles W., Thorne, Kip and Wheeler, John Archibald (1973). *Gravitation*. W. H. Freeman & Co., New York.
Mukhanov, Viatcheslav F. and Chibisov, G. V. (1981). Quantum fluctuations and a nonsingular universe. *JETP Letters* **33**, 532–535.
Parker, Leonard and Fulling, Stephen A. (1973). Quantized matter fields and the avoidance of singularities in general relativity. *Physical Review D* **7**, 2357–2374.
Peebles, Phillip James Edward (1971). *Physical Cosmology*. Princeton University Press, Princeton.
Penrose, Roger (1979). Singularities and time-asymmetry. In *General Relativity: An Einstein centenary survey*, Stephen Hawking and Werner Israel, eds. Cambridge University Press, Cambridge, 581–638.
— (1989). Difficulties with inflationary cosmology. *Annals of the New York Academy of Sciences* **271**, 249–264.
Pickering, Andrew (1984). *Constructing Quarks: A sociological history of particle physics*. University of Chicago Press, Chicago.
Polyakov, Alexander M. (1974). Particle spectrum in quantum field theory. *JETP Letters* **20**, 194–195.
Preskill, John P. (1979). Cosmological production of superheavy magnetic monopoles. *Physical Review Letters* **43**, 1365–8. Reprinted in Bernstein and Feinberg, pp. 292–298.
Press, William H. (1980). Spontaneous production of the Zel'dovich spectrum of cosmological fluctuations. *Physica Scripta* **21**, 702–702.
Rindler, Wolfgang (1956). Visual horizons in world models. *Monthly Notices of the Royal Astronomical Society* **116**, 662–677.
Rugh, Svend E. and Zinkernagel, Henrik (2002). The quantum vacuum and the cosmological constant problem. *Studies in the History and Philosophy of Modern Physics* **33**, 663–705.
Ryder, Lewis H. (1996). *Quantum field theory*. 2nd ed. Cambridge University Press, Cambridge.
Sakharov, Andrei D. (1966). The initial state of an expanding universe and the appearance of a nonuniform distribution of matter. *Soviet Physics JETP* **22**, 241–249. Reprinted in *Collected Scientific Works*.
— (1970). A multisheet cosmological model. Preprint, Moscow Institute of Applied Mathematics. Translated in *Collected Scientific Works*.
— (1982). *Collected Scientific Works*. Marcel Dekker, New York.
Sato, Katsuhiko (1981). First-order phase transition of a vacuum and the expansion of the universe. *Monthly Notices of the Royal Astronomical Society* **195**, 467–479.
Starobinsky, Alexei (1978). On a nonsingular isotropic cosmological model. *Soviet Astronomy Letters* **4**, 82–84.
— (1979). Spectrum of relict gravitational radiation and the early state of the universe. *JETP Letters* **30**, 682–685.

— (1980). A new type of isotropic cosmological models without singularity. *Physics Letters B* **91**, 99–102.
— (1984). Nonsingular model of the universe with the quantum-gravitational de Sitter stage and its observational consequences. In Markov and West (1984). Proceedings of the second Seminar on Quantum Gravity; Moscow, October 13–15, 1981.
Steigman, Gary, Schramm, David N. and Gunn, James E. (1977). Cosmological limits to the number of massive leptons. *Physics Letters B* **66**, 202–204.
't Hooft, Gerard (1974). Magnetic monopoles in unified gauge theories. *Nuclear Physics* **B79**, 276–284.
Tolman, Richard C. (1934). *Relativity, Thermodynamics, and Cosmology*. Oxford University Press, Oxford. Reprinted by Dover.
Tryon, Edward (1973). Is the universe a vacuum fluctuation? *Nature* **246**, 396–397.
Veltman, Martinus J. G. (1974). Cosmology and the Higgs mechanism. Rockefeller University Preprint.
— (1975). Cosmology and the Higgs mass. *Physical Review Letters* **34**, 777.
— (2000). Nobel lecture: from weak interactions to gravitation. *Reviews of Modern Physics* **72**, 341–349.
Wald, Robert (1984). *General Relativity*. University of Chicago Press, Chicago.
Weinberg, Steven (1967). A model of leptons. *Physical Review Letters* **19**, 1264–66.
— (1972). *Gravitation and Cosmology*. John Wiley & Sons, New York.
— (1974a). Gauge and global symmetries at high temperature. *Physical Review D* **9**, 3357–3378.
— (1974b). Recent progress in gauge theories of the weak, electromagnetic, and strong interactions. *Reviews of Modern Physics* **46**, 255–277.
— (1977). *The First Three Minutes*. Basic Books, Inc., New York.
— (1980). Conceptual foundations of the unified theory of weak and electromagnetic interactions. *Reviews of Modern Physics* **52**, 515–523.
Yoshimura, Motohiko (1978). Unified gauge theories and the baryon number of the universe. *Physical Review Letters* **41**, 281–284.
Zee, Anthony (1979). Broken-symmetric theory of gravity. *Physical Review Letters* **42**, 417–421.
— (1980). Horizon problem and broken-symmetric theory of gravity. *Physical Review Letters* **44**, 703–706.
— (1982). Calculating Newton's gravitational constant in infrared stable Yang–Mills theories. *Physical Review Letters* **48**, 295–298.
Zel'dovich, Yakov B. (1967). Cosmological constant and elementary particles. *JETP Letters* **6**, 316–317.
— (1968). The cosmological constant and the theory of elementary particles. (Translated by J. G. Adashko.) *Soviet Physics Uspekhi* **11**, 381–393.
— (1981). Vacuum theory - A possible solution to the singularity problem of cosmology. *Soviet Physics Uspekhi* **133**, 479–503.
Zel'dovich, Yakov B. and Khlopov, Maxim Yu. (1978). On the concentration of relic magnetic monopoles in the universe. *Physics Letters B* **79**, 239–41.

Zel'dovich, Yakov B., Kobzarev, Igor Yu. and 'Okun, Lev B. (1975). Cosmological consequences of a spontaneous breakdown of a discrete symmetry. *Soviet Physics JETP* **40**, 1–5.

Zel'dovich, Yakov B. and Pitaevsky, Lev P. (1971). On the possibility of the creation of particles by a classical gravitational field. *Communications in Mathematical Physics* **23**, 185–188.

Notes

[1]A singularity cannot be straightforwardly defined as "the points at which some physical quantities diverge," since the metric field itself diverges; given the usual assumption that this field is defined and differentiable everywhere on the space-time manifold, these points are *ex hypothesi* not in space-time. The subtleties involved in giving a precise definition were more important for disentangling horizons and coordinate effects from genuine singularities in the Schwarzschild and de Sitter solutions; to my knowledge there were no published debates about whether there is a genuine initial singularity in the FLRW models. See Eisenstaedt (1989); Earman (1999) for historical discussions of the Schwarzschild singularity and the singularity theorems (respectively), and Wald (1984); Earman (1995) for more recent treatments of the intricate conceptual and mathematical issues involved.

[2]An incomplete geodesic is inextendible in at least one direction, but does not reach all values of its affine parameter; even though it does not have an endpoint it "runs out" within finite affine length. Loosely speaking, one can think of an incomplete geodesic as corresponding to "missing points" in a manifold; unfortunately, this idea can be made precise for a Riemannian metric but not for a pseudo-Riemannian metric like that used in general relativity.

[3]More precisely, this research program aimed to show that the general solution describes a "bounce"—the matter reaches a maximum density, but then expands rather than continuing to collapse—and that the bounce fails to occur only for specialized initial conditions. This program resulted in detailed studies of the evolution of anisotropic, homogeneous vacuum solutions in the neighborhood of the initial singularity (see Belinskii et al. 1974, and references therein).

[4]Cosmological models that reached a finite limiting temperature at early times were explored during this time (see, e.g. Hagedorn 1970), but were never widely accepted.

[5]Lemaître (1934) appears to have been the first to clearly state this idea in print. See Earman (2001) for an account of Λ's checkered history, and Rugh and Zinkernagel (2002) for a detailed discussion of the relation between Λ and vacuum energy density in QFT.

[6]Gliner noted that he is only concerned with local Poincaré invariance, but does not recognize the difficulties in extending Poincaré invariance to general relativity. As

a result, in general the "vacuum" cannot be uniquely specified by requiring that it is a Poincaré invariant state. I thank John Earman for emphasizing this point to me (cf. Earman 2001, 208–209).

[7]The strong energy condition requires that there are not tensions larger than or equal to the (positive) energy density; more formally, for any time-like vector v, $T_{ab}v^a v^b \geq \frac{1}{2}T^a_a$. In particular, for a diagonalized T_{ab} with principal pressures p_i, this condition requires that $\rho + \sum_{i=1}^{3} p_i \geq 0$ and $\rho + p_i \geq 0 (i = 1, 2, 3)$, clearly violated by the vacuum state.

[8]Turning this rough claim into a general theorem requires the machinery used by Penrose and Hawking. Gliner refers to Hawking's work in Gliner (1970), but his argument does not take such finer points into account.

[9]This was formulated more clearly as a "cosmic no hair theorem" by Gibbons and Hawking (1977) and in subsequent work. "No hair" alludes to corresponding results in black hole physics, which show that regardless of all the "hairy" complexities of a collapsing star, the end state can be described as simply as a bald head.

[10]Gliner was not alone in this preference; several other papers in the early 1970s discussed violations of the strong energy condition as a way of avoiding the singularity, as we will see in the next section.

[11]Briefly, Sakharov's multi-sheet model is a cyclic model based on Novikov's suggestion that a true singularity could be avoided in gravitational collapse, allowing continuation of the metric through a stage of contraction to re-expansion. I have been unable to find any discussions of the impact of Sakharov's imaginative work in cosmology or its relation to other lines of research he pursued, especially the attempt to derive gravitational theory as an induced effect of quantum fluctuations, but this is surely a topic worthy of further research.

[12]This point is clearly emphasized by Lindley (1985); although it appears plausible that this line of reasoning motivated Gliner and Dymnikova (1975), they introduce the "gradual transition" without explanation or elaboration.

[13]An alert reader may have noticed the tension between this assumption and vacuum dominance mentioned in the last paragraph: the proposed equation of state rather unnaturally guarantees the opposite of vacuum dominance, namely that the *vacuum* is diluted and the density of normal matter and radiation increases in the course of the transition.

[14]Gliner and Dymnikova (1975) derive this equation by solving for the evolution of the scale factor from the transitional phase to the FLRW phase, with matching conditions at the boundary; see Lindley (1985) for a clearer discussion. The constant $0 < \alpha < 1$ fixes the rate at which the initial vacuum energy decays into energy density of normal matter and radiation. H is the (poorly named) Hubble "constant," defined by $H := \frac{1}{a}\frac{da}{dt}$.

[15]Eddington (1933, 37) and de Sitter (1931, 9-10) both argued that a non-zero Λ was needed for a satisfactory explanation of expansion, despite the fact that the FLRW

models with $\Lambda = 0$ describe expanding models; I thank John Earman for bringing these passages to my attention.

[16]Rindler's classic paper introduced and defined various horizons (Rindler 1956); for a recent discussion see Ellis and Rothman (1993). Here I am following the conventional choice to define horizon distance in terms of the time when the signal is received rather than the time of emission (as signalled by the $a(t_0)$ term).

[17]Sakahrov's equation of state is *not* that for a vacuum dominated state, although it is easy to see that the integral diverges for $p = -\rho$ as well.

[18]Hawking's (1970) theorem showed that a vacuum spacetime would remain empty provided that the dominant energy condition holds. The dominant energy condition requires that the energy density is positive and that the pressure is always less than the energy density; formally, for any timelike vector v, $T_{ab}v^a v^b \geq 0$ and $T_{ab}v^a$ is a spacelike vector.

[19]Bekenstein (1975) also discussed the possibility that scalar fields would allow one to avoid the singularity. Starobinsky's (1978) main criticism is that Parker and Fulling dramatically overestimate the probability that their model will reach a "bounce" stage, even granted that the appropriate scalar field exists: they estimate a probability of .5, whereas Starobinsky finds 10^{-43}!

[20]The expression for the trace anomaly was derived before Starobinsky's work; in addition, it was realized that de Sitter space is a solution of the semi-classical EFE incorporating this anomaly (see, e.g. Birrell and Davies 1982). Starobinsky was the first to consider the implications of these results for early universe cosmology.

[21]In the course of this calculation Starobinsky assumed that initially the quantum fields are all in a vacuum state. In addition, the expression for the one-loop correction includes constants determined by the spins of the quantum fields included in $\langle T_{ab} \rangle$, and these constants must satisfy a number of constraints for the solutions to hold. Finally, Starobinsky argued that if the model includes a large number of gravitationally coupled quantum fields, the quantum corrections of the gravitational field itself will be negligible in comparison.

[22]This extended discussion was clearly motivated by Guth's (1981) discussion of the "flatness problem" (which Starobinsky duly cited), but Starobinsky notably did not endorse Guth's emphasis on the methodological importance of the flatness problem.

[23]Misner (1968) advocated an approach to cosmology that focused on "predicting" various features of the observed universe, in the sense of finding features insensitive to the choice of initial conditions.

[24]Zel'dovich's review does not include any references. He had already discussed the horizon problem in a different context (Zel'dovich et al. 1975), see section 2.3 below.

[25]However, this is more a triumph of approach than actual implementation; a decade after this assessment an account of baryogenesis consistent with all the constraints has yet to be developed.

[26]*Very* roughly, in a renormalizable theory such as QED divergent quantities can be "absorbed" by rescaling a finite number of parameters occuring in the Lagrangian (such as particle masses and coupling constants); these techniques did not carry over to massive Yang–Mills theories (see, e.g., §10.3 of Cao 1997 for an overview).

[27]The three "proofs" of Goldstone's theorem given in Goldstone et al. (1962) hold rigorously for classical but not quantum fields; see, e.g., Guralnik et al. (1968) for a detailed discussion of the subtleties involved.

[28]Quantizing the electromagnetic field in Lortenz gauge leads to photons with four different polarization states: two transverse, one longitudinal, and one "time-like" (or "scalar"). In the Gupta–Bleuler formalism, the contributions of the longitudinal and time-like polarizations states cancel as a result of the Lorentz condition $\partial_\mu A^\mu = 0$, leaving only the two transverse states as true "physical" states. See, e.g., Ryder (1996), section 4.4 for a brief description of the Gupta-Bleuler formalism.

[29]Goldstone's theorem held for Lagrangians invariant under the action of a continuous, "global" gauge transformation of the fields, but not for "local" symmetries or discrete symmetries (such as parity). As Chris Martin has pointed out to me, the terms "local" and "global" suggest a misleading connection with space-time: global gauge groups are finite dimensional Lie groups (such that a specific element of the group can be specified by a finite number of *parameters*), whereas local gauge groups are infinite dimensional Lie groups whose elements are specified via a finite number of *functions*.

[30]This discussion of the Higgs mechanism is by necessity brief; for a clear textbook treatment see, for example, Aitchison (1982).

[31]See, e.g., Coleman (1985, chapter 5) for a concise introduction to the effective potential and arguments that it represents the expectation value of the energy density for a given state.

[32]Englert and Brout (1964) explicitly mentioned the possibility: "The importance of this problem [whether gauge mesons can acquire mass] resides in the possibility that strong-interaction physics originates from massive gauge fields related to a system of conserved currents." The other papers introducing the Higgs mechanism are more directly concerned with exploiting the loophole in Goldstone's theorem.

[33]The number of citations of Weinberg (1967) jumped from 1 in 1970 to 64 in 1972, following 't Hooft and Veltman's proof of renormalizability (Pickering 1984, 172).

[34]In dynamical symmetry breaking, bound states of fermionic fields play the role of Higgs field; see the various papers collected in Farhi and Jackiw (1982) for an overview of this research, which was pursued actively throughout the 1970s and early 1980s.

[35]The Cambridge theorists, including Claude Bernard, Sidney Coleman, Barry Harrington, and Steven Weinberg at Harvard, and Louise Dolan and Roman Jackiw at MIT, seem to have worked fairly closely on this research, based on the acknowledgements and references to personal communication in their papers (Weinberg 1974b;

Dolan and Jackiw 1974; Bernard 1974). See Linde (1979) for a review of this literature.

[36]Conventional QFT treats interactions between fields in otherwise empty space, neglecting possible effects of interactions with a background heat bath. Finite temperature field theory was developed in the 1950s in the study of many-body systems in condensed matter physics.

[37]Weinberg (1974a) gives examples of models with no symmetry restoration and even low-temperature symmetry restoration; symmetry restoration can also be induced by large external fields or high current densities. See Linde (1979) for further discussion and references.

[38]In general relativity a conformal transformation is a map: $g_{ab} \rightarrow \Omega^2 g_{ab}$ where Ω is a smooth, non-zero real function. A field theory is conformally invariant if $\phi' = \Omega^s \phi$ is a solution to the field equations with the metric $\Omega^2 g_{ab}$ whenever ϕ is a solution with the original metric, for a given number s (called the conformal weight) (see, e.g., Wald 1984, Appendix D). A field theory is said to be "conformally coupled" if additional terms are introduced to insure conformal invariance; the conformally coupled Klein–Gordon equation, for example, includes a term, $\frac{1}{6}R$, absent from the "minimally coupled" equation obtained by replacing normal derivatives with covariant derivatives.

[39]I call this an assumption since I cannot understand the argument in favor of it, which invokes Birkhoff's theorem along with the conformal flatness of the FLRW models (see Brout et al. 1978, 78–79).

[40]Zee (1982) described the rationale for this approach in greater detail. The program (partially based on Sakharov's conception of "induced gravity") aimed to formulate a renormalizable, conformally invariant theory in which the gravitational constant is fixed by vacuum fluctuations of the quantum fields.

[41]Kirzhnits and Linde defer the detailed argument for this conclusion to a later paper, which apparently did not appear; in any case it is not clear to me that long range repulsive forces are necessarily incompatible with either a closed or uniform model.

[42]"C" denotes charge conjugation, a transformation implemented by replacing field operators for a given particle with those for its anti-particle; "P" stands for the parity transformation, which (roughly speaking) maps fields into their mirror image.

[43]They comment that "Owing to the peculiar expansion law during the initial (domain) stage it is quite possible that $X_c >> X_p$ [X_c is the causal horizon, X_p is the particle horizon]." The averaged equation of state for the domain stage is $p = -\frac{2}{3}\rho$, leading to $a(t) \propto t^2$ during the "cellular medium"-dominated stage of evolution.

[44]Zel'dovich and Khlopov (1978) calculated the abundance of the lighter monopoles produced in electroweak symmetry breaking, with mass on the order of $10^4\ GeV$, whereas Preskill (1979) calculated the abundance of monopoles (with mass on the order of $10^{16}\ GeV$) produced during GUT-scale symmetry breaking.

[45]The stress energy tensor for a scalar field is given by $T_{ab} = \nabla_a\phi\nabla_b\phi - \frac{1}{2}g_{ab}g^{cd}\nabla_c\nabla_d\phi - g_{ab}V(\phi)$; if the derivative terms are negligible, $T_{ab} \approx -V(\phi)g_{ab}$.

[46]Linde estimated that before SSB the vacuum energy density should be $10^{21}g/cm^3$, compared to a cosmological upper bound on the total mass density of $10^{-28}g/cm^3$. In an interview with the author, Linde noted that the title of this paper was mistranslated in the English edition (see the bibliography); the correct title is "Is the Cosmological Constant a Constant?"

[47]The radiation density $\rho_{rad} \propto T^4$, which dominates over the vacuum energy density for $T > T_c$; Bludman and Ruderman (1977); Kolb and Wolfram (1980) bolstered Linde's conclusion with more detailed arguments.

[48]This assumption was not unwarranted: Weinberg (1974a) concluded that the electroweak phase transition appeared to be second order since the free energy and other thermodynamic variables were continuous (a defining characteristic of a second-order transition).

[49]Veltman described the idea of "cancellation" of a large vacuum energy density as follows: "If we assume that, before symmetry breaking, space-time is approximately euclidean, then after symmetry breaking ... a curvature of finite but outrageous proportions result [sic]. The reason that no logical difficulty arises is that one can assume that space-time was outrageously "counter curved" before symmetry breaking occurred. And by accident both effects compensate so precisely as to give the very euclidean universe as observed in nature."

[50]In a 1987 interview he commented that "we understood that the universe could exponentially expand, and bubbles would collide, and we saw that it would lead to great inhomogeneities in the universe. As a result, we thought these ideas were bad so there was no reason to publish such garbage" (Lightman and Brawer 1990, 485–86).

[51]Sato apparently hoped that an early phase transition would effectively separate regions of matter and anti-matter, so that observations establishing baryon asymmetry could be reconciled with a baryon-symmetric initial state; he also mentions the possibility that small inhomogeneities could seed galaxy formation.

[52]The original draft of this paper was completed in July 1980, revised in November of 1980 partially in response to comments from Guth and his collaborator, Erick Weinberg. Einhorn and Guth met and discussed phase transitions in November of 1979, but judging from Guth's comments in Guth (1997, 180), Einhorn and Sato hit upon the idea of false-vacuum driven exponential expansion independently.

[53]Einhorn and Sato were not alone in making this suggestion; a year earlier, the Harvard astrophysicist Bill Press had proposed an account of structure formation in which vacuum energy does not couple to gravity. In Press's (1980) scenario, inhomogeneities in the vacuum are converted into fluctuations in the energy density of matter and radiation. This "conversion" only works if the vacuum does not itself gravitate; Press noted the speculative nature of this suggestion, but argued that the other possibility — an incredibly precise cancellation of vacuum energy density — is equally unappealing.

[54]Rees attended talks about the early universe by both Starobinsky and Englert before 1981, but by his own account he did not see the appeal of these ideas until he had read Guth's paper (Lightman and Brawer 1990, 161).

[55]See Guth (1997), chapter 10 for a detailed account (quotation on 179). Guth attended a lecture by Princeton's Bob Dicke, in which he mentioned the flatness problem, on Nov. 13, 1978.

[56]The density parameter is defined as the ratio of the observed density to the critical density, namely the value such that $k = 0$ in the Friedmann equation. The Friedmann equation is given by: $H^2 = \frac{8\pi G}{3}\rho - \frac{k}{a^2(t)}$, where $k = 0$ for a flat model, $k > 0$ for a closed model, and $k < 0$ for an open model.

[57]Guth and Weinberg (1983) later showed that for a wide range of parameters the bubbles do not percolate, and they also do not collide quickly enough to thermalize.

[58]Michel Janssen has recently argued that "common origin inferences" (COIs) play a central role in scientific methodology (Janssen 2002). These inferences license a preference for a theory that traces several apparent coincidences to a common origin. Guth's case for inflation is a particularly clear example of this style of reasoning. I have benefitted from extensive discussions with Janssen regarding whether the case for inflation should be treated as another "COI" story, but I do not have space to explore the issue further here.

[59]For a detailed discussion of the demand for explanatory adequacy see Earman (1995), and for a critical overview of inflationary cosmology see Earman and Mosterin (1999).