APPROACHING THE ABSOLUTE ZERO OF TIME:

THEORY DEVELOPMENT IN EARLY UNIVERSE COSMOLOGY

by

Christopher Joel Smeenk

B.A., Physics and Philosophy, Yale University, 1995

M.S., Physics and Astronomy, University of Pittsburgh, 2001

M.A., Philosophy, University of Pittsburgh, 2002

Submitted to the Graduate Faculty of

Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2003

We approve the thesis of Christopher Joel Smeenk.

Date of Signature

John S. Earman
University Professor of History and Philosophy of Science
Co-Chair of Committee

John D. Norton
Professor of History and Philosophy of Science
Chair of the Department of History and Philosophy of Science
Co-Chair of Committee

Laura M. Ruetsche
Associate Professor of Philosophy

Allen I. Janis
Professor Emeritus of Physics and Astronomy

Abstract

This dissertation gives an original account of the historical development of modern cosmology along with a philosophical assessment of related methodological and foundational issues. After briefly reviewing the groundbreaking work by Einstein and others, I turn to the development of early universe cosmology following the discovery of the microwave background radiation in 1965. This discovery encouraged consolidation and refinement of the big bang model, but cosmologists also noted that cosmological models could accomodate observations only at the cost of several "unnatural" assumptions regarding the initial state. I describe various attempts to eliminate initial conditions in the late 60s and early 70s, leading up to the idea that came to dominate the field: inflationary cosmology. I discuss the pre-history of inflationary cosmology and the early development of the idea, including the account of structure formation and the introduction of the "inflaton" field. The second part of my thesis focuses on methodological issues in cosmology, opening with a discussion of three principles and their role in cosmology: the cosmological principle, indifference principle, and anthropic principle. I assess appeals to explanatory adequacy as grounds for theory choice in cosmology, and close with a discussion of confirmation theory and the issue of novelty in relation to cosmological theories.

Table of Contents

List of Figures

Acknowledgments

Without the encouragement and support from many people I would not have completed this dissertation, and I am indebted to all of them.

Several institutions supported different aspects of this research. A grant from the National Science Foundation (SES 0114760) made it possible to perform interviews with cosmologists, and I learned an enormous amount from the interviewees: John Barrow, Andrei Linde, Jeremiah Ostriker, James Peebles, Martin Rees, William Saslaw, Paul Steinhardt, and Neil Turok. I would also like to thank Spencer Weart at the American Institute of Physics (AIP) for supporting the interview project. The AIP sponsored an interview with Charles Misner, a visit to the AIP archives, and transcription of all the interviews. Patrick McCray and Spencer Weart both gave me helpful advice. The Slater Fellowship from the American Philosophical Society gave me the freedom to extend the project in various ways, and I am very grateful for the award. I am also grateful for the generous financial support in various guises that I have received from the Department of History and Philosophy of Science.

My greatest intellectual debt is to the members of my committee: John Earman, Al Janis, John Norton, and Laura Ruetsche. The work of the two Johns has set high standards for scholarship in the history and philosophy of science, and their advice and encouragement has helped me to come closer to attaining that standard. Conversations with John Earman initially inspired this project, and I owe a clear debt to his work on modern cosmology. His insightful comments at various stages have helped me to avoid a number of misconceptions. I benefitted enormously from John Norton's thoughtful responses and suggestions, and discussions with John invariably led me to a much clearer conception of the issues at hand. Al Janis read everything I handed him with much more care than I had any right to expect, and his detailed comments improved the dissertation considerably. Laura Ruetsche consistently pointed me towards fruitful intersections between technical issues and the broader philosophical literature. Finally, I learned a great deal from Rob Clifton in the early stages of this work. His tragic early death robbed the community of one of its most insightful and penetrating scholars, and I sorely missed the opportunity to benefit from his further guidance.

Many people contributed to the productive and engaging atmosphere of HPS and helped me enjoy my stint in graduate school. I thank Bernard Goldstein, Peter Machamer, and Ted McGuire for inspiring and encouraging my interest in early modern science and philosophy. Several friends have helped me to maintain my sanity through graduate school, but I am particularly indebted to Zvi Biener, Chris Martin, Greg Morgan, Wendy Parker, Gualtiero Piccinini, and John Roberts. I would like to thank my family for always encouraging me to pursue whatever career interested me, even if it meant leaving the farm and taking up something as bizarre as philosophy of science. They have been a constant source of support throughout my long career as a student.

Finally, my greatest debt is to my wife, Sarah Gallagher. She has contributed directly to this dissertation in many ways, including drawing the figures, giving me LaTeX advice, and patiently answering my questions about astrophysics. But more importantly, our relationship has been a constant source of strength and inspiration throughout the last eight years. I look forward to continuing this intellectual adventure together with Sarah.

Nature, we find, even from our limited experience, possesses an infinite number of springs and principles, which incessantly discover themselves on every change of her position and situation. And what new and unknown principles would actuate her in so new and unknown a situation as that of the formation of a universe, we cannot, without the utmost temerity, pretend to determine.

David Hume

Cosmologists are often in error, but never in doubt.

Lev Landau

Chapter 1

Introduction

Historians and philosophers of science have paid relatively little attention to modern cosmology, in contrast to both the substantial scholarly literature concerning the cosmologies of earlier eras and research focused on other developments in 20^{th} century physics. The controversy between the steady state and big bang models–with its overt focus on methodological and philosophical issues–spilled over into the philosophy of science literature in the 50s and 60s, but more recent developments have for the most part not drawn the attention of historians and philosophers. Partly this is due to the increasing size and technical sophistication of the research literature in cosmology, and other areas within physics have established traditions devoted to their history and philosophical implications. In broad terms, the aim of this dissertation is to correct this oversight, and to provide some indication of the riches awaiting historians and philosophers who turn their attention to cosmology. As with any early exploration, I follow an idiosyncratic path that colors my sense of the lay of the land, and I am keenly aware of the unexplored territory surrounding the topics I have chosen to focus on. This introductory chapter surveys the historical development of cosmology and attendant methodological debates leading up to the mid-1960s, where the account starting in Chapter 2 begins. The introduction closes with a brief overview of the dissertation.

1.1 Einstein and Cosmology

Prior to Einstein's groundbreaking paper (Einstein 1917), cosmology was hardly an area of active concern for physicists. Newton briefly discussed the application of his gravitational theory to a "cosmological" matter distribution in response to Bentley's prompting, but this line of thought—and the attendant serious conceptual difficulties with applying Newtonian gravitational theory to an inifinite distribution of matter—was explored by only a handful of scientists from Newton's time up to the 20^{th} century.[1] In this section I will briefly assess Einstein's motivations for proposing a cosmological model. I will argue that Einstein is the first in a long line of physicists pushed to cosmological speculations by purely theoretical concerns, and I will also criticize the common characterization of Einstein (1917) as a straightforward application of his newly completed theory. Einstein (1917) should be seen instead as a continuation of the road leading to general relativity.[2]

After several years of struggling to reconcile Newtonian graviational theory with special relativity, Einstein's work built to a crescendo in November of 1915. He delivered papers on the general theory of relativity to the Prussian Academy on four consecutive Thursdays. Following what was surely one of the most stenuous months of work in his life, he considered his task complete with the presentation of the final paper on November 25.[3] He had successfully formulated a field theory of gravitation that apparently satisfied the formal and physical requirements which had guided his long search for a new theory.

---

[1] See North (1965); Norton (1999) for discussions of the difficulties in formulating Newtonian cosmology.

[2] Discussions with Michel Janssen have been a great help in understanding Einstein's early work on cosmology and the Einstein-de Sitter correspondence, and I owe much of what follows to those discussions; cf. Torretti (2000).

[3] Einstein lays out the story of the twists and turns leading to the final paper in a detailed letter to Sommerfeld written just three days later (Schulmann et al. 1998, Doc. 153).

Thanks to extensive, careful historical work by a number of scholars, we now have a thorough understanding of these requirements and how Einstein arrived at the final field equations.[4] From his earliest work on gravitation (Einstein 1907), Einstein thought that any extension of the principle of relativity to accelerated systems must incorporate the striking empirical equality of inertial and gravitational mass. He initially formulated the equivalence principle, described as a generalization of the principle of relativity to accelerating reference systems, as the claim that a gravitational field is physically equivalent to uniform acceleration.[5] Eight years later, the principle amounted to the claim that there is no background inertial structure in spacetime; instead, inertial and gravitational effects are both consequences of the metric field, and they cannot be disentangled from each other. Two additional crucial requirements were that the theory incorporate an energy-momentum conservation law, and that the field equations reduce to the Newtonian form in the limit of weak, static fields.[6] The series of papers in November reveal that Einstein's breakthrough came in part by finally rejecting a number of erroneous assumptions regarding the Newtonian limit. He abandoned the field equations of an earlier theory (the so-called *Entwurf* theory) and renewed the search for generally covariant field equations. (Einstein 1916 characterized general covariance as the requirement that "the universal laws of nature must be expressed by equations which hold good for all coordinate systems, that is, are covariant with respect to arbitrary transformations.") The field equations presented on Nov. 25 appeared to satisfy all

[4]Norton (1984) (drawing on earlier work of John Stachel) is the canonical shorter account of Einstein's path to general relativity, soon to be supplemented with an authoritative book-length treatment, Renn et al. (2003). It goes without saying that there are a number of remaining questions regarding the intricate conceptual and technical issues Einstein struggled with before (and after) finding the "final" field equations (such as Einstein's understanding of "coordinate conditions"). Hopefully Renn et al. (2003) will mark out the common ground shared by its contributors, and also clarify the remaining disputes.

[5]This rough formulation follows Einstein's remarks in §17 of Einstein (1907). See, e.g., Norton (1985); Torretti (1983) for more detailed discussions of the equivalence principle.

[6]For a clear reconstruction of how the requirement of energy-momentum conservation came into play, traced through the pages of Einstein's Zurich notebook, see Norton (2000).

of Einstein's guiding principles: they upheld the equivalence principle; they had an appropriate Newtonian limit, but also gave correct predictions for departures from Newtonian theory, notably the value of Mercury's anomalous perihelion motion; they incorporated energy-momentum conservation; and, finally, they were generally covariant.

But as a consequence of discussions and extensive correspondence with the Leyden astronomer Willem de Sitter beginning in 1916, Einstein soon realized that the field equations did *not* satisfy an additional guiding principle. He added "Mach's Principle" to the list of foundational principles of general relativity:[7]

> The metric field is determined *without residue* by the masses of bodies. Since mass and energy are equivalent according to the results of the special theory of relativity, and since energy is formally described by the symmetric energy tensor T_{ij}, this means that the metric field is conditioned and determined by the energy tensor. (Einstein 1918b, p. 242, original emphasis)

Einstein explicitly added this principle to the short list after recognizing that it might not hold in his newly minted theory. In particular, boundary conditions imposed at infinity could play a role (alongside the distribution of matter and energy) in determining inertial structure. For Einstein this was contrary to the spirit, if not the letter, of his theory; he put the complaint as follows in 1917:

> In a consistent theory of relativity there can be no inertia *relative to "space"* but only an interia of masses *relative to each other*. Hence if I take a mass sufficiently far away from all the other masses in the world its inertia must fall down to zero. (Einstein 1917, p. 145, original emphasis)

[7] This is the third and final foundational principle, following the principle of relativity and the principle of equivalence. Norton (1993, §3) describes the evolution of these principles in Einstein's thought leading up to GTR and through the debates with de Sitter.

Requiring that the metric field approaches the flat Minkowski metric far from all concentrations of matter would introduce inertia relative to this flat background, which is apparently completely independent of the matter distribution.

Einstein first suggested that this remnant of "absolute space" could be disposed of by stipulating that the metric take degenerate values at infinity. The flat Minkowski values of the metric field at finite but large distances were to be explained via "distant masses." Einstein soon abandoned this proposal (see Einstein 1917 and the de Sitter correspondence, included in Volume 8 of the Collected Papers, Schulmann et al. 1998) in favor of a much more radical solution: why not get rid of infinity entirely! Einstein's bold suggestion was that space could be unbounded yet finite. In modern terms, Einstein's cosmological model is topologically $\mathbb{R} \times S^3$; it can be naturally decomposed into surfaces at constant "cosmic time," which are compact and homeomorphic to S^3. Each of these constant time slices is finite, and the need to stipulate boundary conditions is avoided since there are no boundaries. Einstein is very clear regarding his motives for introducing the model: he calls the model "nothing but a spacious castle in the air," built to accomodate Machian intuitions rather than observations (Schulmann et al. 1998, Doc. 311).

The castle in the air came at a price: Einstein was forced to modify his final field equations from November of 1915 since they did not admit this cosmological model as a solution. However, the model was a solution of the modified field equations:[8]

$$R_{ab} - \frac{1}{2} R g_{ab} + \Lambda g_{ab} = \kappa T_{ab} \tag{1.1}$$

[8] Einstein wrote down the field equations in the following form, mathematically equivalent to the equation stated in the text: $R_{\mu\nu} = \kappa(T_{\mu\nu} - \frac{1}{2} T g_{\mu\nu}) + \Lambda g_{\mu\nu}$, where T is the trace of $T_{\mu\nu}$.

This equation specifies the relationship between the distribution of matter and energy (the right hand side) and spacetime geometry (the left hand side). Very briefly, in Einstein's theory spacetime is modeled by a four-dimensional manifold equipped with a metric, g_{ab}.[9] The metric defines the geometry of spacetime, in that the distance between two neighboring points x^a and $x^a + dx^a$ is given by $ds^2 = g_{ab}dx^a dx^b$. The quantities appearing on the left hand side of (1.1)—the Ricci tensor R_{ab}, and the Ricci scalar R—are both defined in terms of the metric and its first and second derivatives. The constant $\kappa = 8\pi G/c^4$ (although I will adopt geometric units below, setting $G = c = 1$), T_{ab} represents the distribution of stress and energy, and Λ is the new "cosmological constant" term which was absent from the original field equations. Although Einstein would later call the introduction of Λ his "greatest blunder," the modified field equations share the virtues of the original equations. Indeed, the left hand side of (1.1) is the most general symmetric tensor that can be constructed from the metric and its first and second derivatives, such that the covariant conservation law for T_{ab} is still satisfied. Even for those who didn't share Einstein's fine aesthetic sense (he later commented that Λ damaged the formal beauty of the field equations), observational constraints provided ample evidence that Λ must be *very close* to zero.[10] These constraints have not prevented cosmologists from re-introducing Λ again and again for various reasons, as we will see below.

Einstein was driven to cosmological speculation not by a desire to model the universe as a whole, but by the hope of showing that his new theory incorporated the Machian intuitions that had been so important in guiding his research. Observations are mentioned only once in

[9] See A.1 for definitions and a brief discussion, or for a more comprehensive introduction to GTR, see, e.g., Misner et al. (1973); Wald (1984). For consistency with later work I have written the field equations using the modern "abstract index notation" introduced in the late 60s: the indices appended to a tensor characterize the tensor, *not* its components in a particular basis, whereas in the old notation $T_{\mu\nu}$ represents the components of a tensor in a given basis. Throughout the dissertation I will use a (-+++) signature.

[10] See, e.g., Weinberg (1989) for a contemporary review of observational constraints on Λ.

Einstein (1917): Einstein supports his suggestion that the universe is static by noting the small relative velocity of the stars. In this context "static" means that the metric does not change over time; as a consequence, in Einstein's model the distance between curves traced out by freely falling particles is fixed. Any one of the "time slices" of Einstein's model is equivalent to another. This assumption is what actually forced Einstein's hand: as he explicitly noted, one might be able to construct non-static cosmological models with $\Lambda = 0$ (Einstein 1917, p. 152). Friedmann would later show that there are dynamical models with $\Lambda = 0$ that have closed spatial sections, as in Einstein's model. Thus the modification of the field equations was not necessary for preserving Mach's Principle by eliminating boundaries at infinity. The introduction of Λ was also not sufficient for preserving Mach's Principle, as Einstein's correspondent de Sitter showed with a slight modification of Einstein's model. Although the de Sitter solution is a vacuum solution, its inertial structure differs from that of flat Minkowski space; it is thus a straightforward counterexample to Mach's Principle (as Einstein formulated it).

In an extended epistolary exchange involving de Sitter, Klein, and Weyl from late 1916 to early 1919, Einstein suggested various reasons to rule out de Sitter's solution as physically inadmissible and thus preserve Mach's Principle.[11] His early criticisms focused on an apparent singularity in the de Sitter solution: Einstein argued that the divergence of the metric at the "equator" of the de Sitter solution represented a true singularity (Einstein 1918a).[12] Correspondence with Klein eventually convinced Einstein that this singularity is due to the use of static

[11] The debate is played out in a series of letters between Einstein and de Sitter from November 1916 to April 1918, with Einstein discussing related issues with Klein and Weyl until 1919; see, in particular, Michel Janssen's thorough editorial note "The Einstein-de Sitter-Weyl-Klein Debate" (Schulmann et al. 1998, pp. 351-357).

[12] In the static coordinates used by de Sitter (1917b), the "equator" is the hypersurface given by $r = \frac{\pi}{2}R$, and at that point the g_{tt} component of the metric vanishes. As subsequent work clarified, the apparent singularity corresponds to the event horizon (see, in particular, Schrödinger 1957). See the headnote referred to above for a discussion of Einstein's other objections.

coordinates, which only cover a portion of the full de Sitter space; the bad behavior of the metric on the equator reflects this poor choice of coordinates rather than any real singularity in the spacetime. When eventually forced to admit that this criticism was misplaced, Einstein retreated to the position that the de Sitter solution could not be accepted because it was not static. Einstein had earlier argued in favor of a static model on the slim observational basis that the velocities of the stars appears to be small (Einstein 1917), but rather than leading him to reconsider this assumption the exchange appears to have reinforced his conviction that a physically reasonable solution of the field equations must be static.

de Sitter did not share Einstein's concerns with preserving Machian insights, but he immediately studied possible observational consequences of the speculative cosmological models. This was fully in character: long before Einstein's successful completion of the general theory, de Sitter (1911) calculated the astronomical consequences of early work on Lorentz covariant gravitational theories (by Poincaré and Minkowski). de Sitter served as an "ambassador" for relativity theory in two different respects: he brought the new theory to the attention of the astronomical community through a series of publications on GTR (de Sitter 1916a,b, 1917a), and he published in English journals accessible to astronomers and physicists in England and America during the war. de Sitter's articles shaped the ensuing debate regarding the status of these cosmological models. In particular, he emphasized one striking observational consequence: a systematic red-shift of the spectral lines of distant stars or nebulae in his own model. By 1930 Eddington, Weyl, Jeans, de Sitter, Hubble, and a handful of others had been drawn into a debate weighing the virtues and vices of Einstein's static model and the de Sitter solution, on observational as well as theoretical grounds. One focus of the debates was the interpretation of the

redshift in de Sitter's solution. From the line-element for the de Sitter solution in static coordinates it was clear that free particles at rest would move apart, and de Sitter correctly argued that as a result spectral lines from distant stars or "nebulae" would be systematically red-shifted (de Sitter 1917a). However, determining whether this was a real or "spurious" effect, as well as the dependence of this redshift on distance from an observer, required differentiating different causes of redshift from coordinate effects. The failure to do so led to a decade-long debate regarding the nature of redshift in de Sitter's solution.[13] The Einstein vs. de Sitter debate did not end with a firm resolution in favor of either model; instead, the participants belatedly recognized that there were many other models to choose from, as we will see in the next section.

These early debates regarding cosmological models reflect two general problems: first, the difficulty in separating artifacts of particular coordinate representations from genuine physical features of the models, and second, deciding what general features should be required of "physically reasonable" models. The debates described above led to some progress on the first question: by the mid 1930s there was broad agreement regarding the genuine physical features of various cosmological models (see, e.g. Tolman 1934), although other issues such as the nature of singularities proved to be far more subtle (see §5.2). Regarding the second issue, Einstein insisted that a physically reasonable cosmological model should be both static and singularity free. The use of strong and intuitively clear principles to delimit the space of reasonable models continued to be a mainstay of arguments in cosmology. But partially due to de Sitter's influence, over the course of the 20s qualitative agreement with observational results began to play a more important role in judging the merits of cosmological models.

[13] For a blow-by-blow recounting of the debates see North (1965, pp. 92-104), but for a clear description of the different causes of redshift that dispels the confusion see Ellis (1989).

1.2 Evolutionary Cosmology

This section briefly reviews some of the developments in cosmology in the five decades after Einstein's seminal paper, focusing on homogeneous and isotropic models, the study of galaxy formation via gravitational clumping, and Gamow's theory of "big bang" nucleosynthesis. These three ideas were later incorporated into what Steven Weinberg dubbed the "standard model" of big bang cosmology (Weinberg 1972), but it is important to emphasize that in textbooks and review articles from the late 50s and early 60s they were not presented as a coherent, unified theory. Gamow's ideas, for example, entirely disappeared from view, so completely that in 1965 Peebles was unaware of work of Gamow and his collaborators until a referee for *Physical Review* rejected his paper on nucleosynthesis on the grounds that it repeated their results.[14] There is a striking contrast in approach between Gamow's application of particle physics to the early universe, various attempts to account for galaxy formation, and more mathematical studies of the features of sundry cosmological models (such as Schrödinger 1957). A richer historical account of these ideas would take up the interactions (and lack thereof) between such different approaches to cosmology, along with their relation to the controversial steady state theory, which a vocal minority advocated as an alternative to evolutionary theories.[15] Since my main goal is to set the stage for the more detailed history to follow, I will instead give only a quasi-historical introduction to these components of big bang cosmology.

[14] See Kragh (1996) §7.1 for this anecdote and more detail on the revival of research on nucleosynthesis, and Chapter 3 for a detailed treatment of Gamow's theory.

[15] There are several historical accounts covering this period: Kragh (1996); North (1965); Ellis (1989, 1990). George Gale and his frequent collaborator John Urani have also written a number of papers on the development of cosmology from the 20s to the 40s, with an explicit focus on the methodological debates that shaped the emerging science (see Gale 2002, for references).

Following Einstein's work described above, study of the large-scale structure of the universe and its dynamical evolution has been primarily based on exact solutions of Einstein's field equations (EFE). Most exact solutions bear the name of whoever possessed the mathematical insight or determination needed to find them. The daunting task of solving the field equations (in a particular coordinate system, they are generally a set of 10 coupled, non-linear partial differential equations) can be simplified considerably by making strong symmetry assumptions, and in the 20s several authors discovered a set of solutions which have been used extensively since their re-discovery in the early 30s. This set of solutions, called the Friedmann-Lemaître-Robertson-Walker (FLRW) solutions, includes all homogeneous and isotropic solutions of the field equations.[16]

Homogeneity and isotropy characterize the geometric features of spacetime as seen by stationary observers. Robertson (1929) introduced these concepts as follows. He first assumed that spacetime may be divided into three-dimensional spaces labeled by a set of coordinates x^i with $i = 1, 2, 3$ and a single time dimension. This is equivalent to the requirement that in suitable coordinates the line element has the form

$$ds^2 = -dt^2 + h_{ij} dx^i dx^j \tag{1.2}$$

(h_{ij} is the metric of the three-dimensional surfaces). Suppose that in addition we consider a specialized set of "comoving" coordinates x^i, defined such that the matter distribution is at rest

[16]These solutions are also often referred to as the FRW or RW models. Friedmann and Lemaître independently derived the dynamics of the FLRW models by solving (1.1) with particular matter sources along with assumptions regarding the overall geometry (see Torretti 1983, §6.3), whereas Robertson and Walker showed that equation (1.3) below is the general expression of the metric for a homogeneous and isotropic model. The difference in these approaches is quite striking. Gale (2002) has recently pointed out that Robertson and Walker's approach was heavily influenced by Milne's kinematic relativity.

with respect to these coordinates—i.e., the matter moves along geodesics with x^i =constant. Robertson (1929) characterized isotropy as the requirement that all spatial directions should "look the same" to any observer at rest with respect to the comoving coordinates. Such an observer cannot find a preferred spatial direction based solely on the geometry of spacetime. On the other hand, in a homgeneous spacetime different points on the same three dimensional space all "look the same" geometrically—at a given instant in time, all the different spatial points are geometrically equivalent. Resorting briefly to more familiar terminology, these concepts can be defined in terms of the isometries of the spatial metric. In an isotropic spacetime, the spatial sections are symmetric around a comoving observer; more precisely, for such an observer at the point p, any vector orthogonal to u^a, the tangent to the observer's worldline at p, can be mapped into another orthogonal vector v^a by an isometry of the metric h_{ij}. Isotropy characterizes the three-space orthogonal to u^a at a single point; in contrast, homogeneity is defined with respect to a foliation of spacetime into spacelike surfaces. A spacetime is spatially homogenous provided that such a foliation exists and any point on a given spatial surface can be mapped into any other point on the same surface by an isometry of the metric.

These two properties taken together place very tight constraints on the full metric g_{ab}. The constant time hypersurfaces are maximally symmetric—that is, they are three dimensional spaces with constant curvature. As Robertson (1929) showed, the general form of the line element for FLRW solutions is:

$$ds^2 = -dt^2 + a(t)^2 \left(\frac{dr^2}{1 - kr^2} + r^2(d\theta^2 + \sin^2\theta d\phi^2) \right) \tag{1.3}$$

The factor k classifies the three distinct possibilities for the curvature of the three-spaces: positive, $k = +1$ corresponding to spherical space; zero for flat space; and negative curvature ($k = -1$) corresponding to hyperbolic space. Although the curvature of a given spatial surface is constant, the curvature may change with time—as signaled by the presence of the scale factor $a(t)$, which measures the change over time of any selected length scale, such as the distance between neighboring galaxies at rest with respect to the comoving coordinates.

For such geometrically simple spacetimes solving EFE is a relatively straightforward calculation rather than a mathematical nightmare. The only matter source compatible with the symmetries of the FLRW solution is that of a perfect fluid, i.e. $T_{ab} = (\rho+p)v_a v_b + p g_{ab}$. Solving (1.1) yields two independent equations governing the time evolution of the scale factor (where $\dot{a} =: \frac{da}{dt}$) (cf. Appendix A.2):

$$\frac{\ddot{a}}{a} = -\frac{4\pi}{3}(\rho + 3p) + \frac{\Lambda}{3} \tag{1.4}$$

$$\left(\frac{\dot{a}}{a}\right)^2 = -\frac{k}{a^2} + \left(\frac{8\pi}{3}\right)\rho + \frac{\Lambda}{3}. \tag{1.5}$$

With the cosmological constant set to zero, the evolution of the scale factor depends upon the choices of ρ, p, k; the equation of state for different states of matter fixes the values of ρ and p, but the value of k depends upon the overall matter density of the universe. The stress-energy tensor for "normal" matter and energy obeys the following inequalities: $\rho \geq 0, \rho + 3p \geq 0$.[17] With Λ set to zero, these inequalities imply that the first term on the RHS in equation (1.4) is strictly

[17] I am assuming that "normal" matter satisfies both the strong and weak energy conditions. The weak energy condition requires that any observer moving with a velocity v^a measures a non-negative energy density: $T_{ab}v^a v^b \geq 0$, whereas the strong energy condition holds that $T_{ab}v^a v^b \geq \frac{1}{2}T$ for any timelike V^a (for a perfect fluid, this implies that $\rho + 3p \geq 0$). The standard terminology here is misleading, since the strong energy condition does not imply the weak energy condition. Attitudes towards energy condition-violating fields have changed considerably with the realization that various classical and quantum fields violate both of these conditions.

negative, and it follows that $\ddot{a} < 0$; the expansion of the universe decelerates (for any choice of k). In addition, from (1.5) the velocity of the expansion can change from positive to negative if and only if $k = 1$; for $k \leq 0$ the right hand side of (1.5) is strictly positive, and the expansion does not halt. This remarkably direct argument shows that the simplest cosmological solutions of EFE are inherently dynamical: as long as the stress-energy tensor obeys the constraint above, the scale factor evolves over time with $\ddot{a} < 0$. Predilections for a static cosmology can be satisfied by allowing a non-zero Λ, following the example of Einstein (1917). In a vacuum solution with a non-zero cosmological constant, the right hand side of (1.4) is positive if $\Lambda > 0$, and consequently $\ddot{a} > 0$. Thus in contrast to "normal" types of matter and energy, the cosmological constant produces a repulsive force that counteracts the attraction of gravitation. Setting $\ddot{a} = 0$ in equation (1.4) to obtain a precise balance between repulsion and attraction implies that $\Lambda = 4\pi(\rho + 3p)$.

Friedmann and Lemaître's work was virtually ignored throughout the 20s, while debates in observational cosmology focused exclusively on the Einstein and de Sitter solutions.[18] There were theoretical and observational reasons for abandoning the static models. Hubble's famous observations of a linear relationship between red-shift and distance provided strong evidence in favor of expanding models. Furthermore, Eddington realized that Einstein's solution is unstable to slight variations in the value of Λ (Eddington 1930).[19] de Sitter's solution could apparently accomodate the red-shift observations, but it is a vacuum solution. (Although the de Sitter solution is *not* a static solution, it was often written in a static form by choosing coordinates that only

[18] Einstein (1922)'s claim that Friedmann's solutions were incompatible with EFE must have damped interest in them, and his retraction a year later probably did little to revive it.

[19] Eddington did not immediately discard Einstein's solution due to instability; instead, he suggested that the universe may have evolved from the unstable Einstein solution to the de Sitter solution.

cover a portion of the space.) In Eddington's memorable question to the Royal Astronomical Society in 1931: "Shall we put a little motion into Einstein's world of inert matter, or shall we put a little matter into de Sitter's Primum Mobile?" Following the meeting, Lemaître notified Eddington of his earlier work, ending the neglect of the FLRW models.

The utility of the simple FLRW models depends upon whether the universe is even approximately homogeneous and isotropic. The status of the "cosmological principle" (Milne's term for the assumption that the universe is homogeneous and isotropic) differed widely among the cosmologists working during the time period from the late 30s to the 60s. Bondi, Gold, and Hoyle followed Milne's deductive approach and amplified the cosmological principle to include uniformity over time, calling it the "perfect cosmological principle". The enhanced principle was then used to derive the steady state theory.[20] Both Milne and the steady state theorists held that these cosmological principles could be given strong *a priori* justifications of a purely philosophical nature.

The cosmological principle was one of the focal points of lively, often downright vicious, debates among cosmologists in the 30s and 40s.[21] Despite agreement regarding the *content* of the cosmological principle, its status varied widely. In Milne's hands it was the consequence of an *a priori* requirement that every observer in the universe should see the same cosmos. In contrast to this deductive style, Robertson and others sought to justify uniformity as a useful extrapolation of observed regularities. But the nature of this extrapolation was unclear; for example, Robertson's colleague at Caltech Richard C. Tolman was wary of the unrelenting uniformity built into the

[20] I will discuss the methodological stance of the steady state theorists in slightly more detail below, but see Kragh (1996), Chapter 2 *et passim* and North (1965) for historical accounts of the development of the steady state theory and the the controversy between this theory and rival evolutionary models.

[21] See Gale (2002) for a wonderful account of these debates (and references to related work). Gale corrects a serious oversight of previous accounts by clarifying Milne's influence on Robertson and Walker, as well as on Bondi and the steady state theorists.

FLRW solutions. On Tolman's view, the cosmological principle was an assumption valuable for mathematical reasons, justified only indirectly by its apparent compatibility with observations. Introducing a discussion of the FLRW solutions in his influential textbook (Tolman 1934), he commented that:

> ...although we shall make great use of homogeneous models in our studies, we shall have to realize that we do this primarily in order to secure a definite and relatively simple mathematical problem, rather than to secure a correspondence with known reality. (Tolman 1934, p. 332)

Tolman re-emphasized this point eloquently in his concluding remarks:

> [W]e must be careful not to substitute the comfortable certainties of some simple mathematical model in place of the great complexities of the actual universe. (Tolman 1934, p. 487)

This sentiment was echoed by several cosmologists working in America, notably George McVittie.[22] Significant observational results such as the linear relationship between the red shift of the spectral lines of galaxies and their distance, the similarity of galaxy distributions in different directions, and the number counts of galaxies and radio sources could all be accomodated (at least qualitatively) in the FLRW models. This set the FLRW models apart from the other known exact solutions of EFE, even as more exact solutions were discovered.

Although the unrelenting uniformity of the FLRW models conflicts with the non-uniformities of the local universe, from the mid 30s onward cosmologists generally assumed that at large enough scales these non-uniformities would disappear. In 1934 observational evidence for homogeneity on large scales was limited to a distance of about 10^8 light years. Hubble initiated

[22] McVittie left Britain for the Illinois observatory at least partially to escape the intellectual climate of British cosmology, and in his textbook McVittie (1965) he characterizes the arguments in support of a deductive approach as "largely illusory" (p. 187), and comments that: "If therefore the uniform models... are employed in the interpretation of astronomical data, the reason lies not in necessity, but because these models appear to provide the simplest first step that can be taken."

an observational program (working with Humason) to measure the uniformity of the large scale distribution of nebulae (galaxies) in 1926, and throughout the 30s several theorists including Tolman invoked Hubble's observational data to bolster the assumption of homogeneity (see Tolman 1934, §177-185). However, Hubble's conclusions were challenged by other observers, most prominently Shapley, who disagreed about the significance of small scale clumping in the distribution of galaxies.[23] Hubble acknowledged the presence of this clumping, but he argued that at sufficiently large scales the distribution was indeed uniform. Observational evidence gathered throughout the 50s and 60s did not settle the question.[24] Although the observational evidence did not fully vindicate the focus on the FLRW models, it also did not force cosmologists to abandon the comfortable certainties of these simple models. In an extensive review of the observational tests of cosmology within the reach of the 200 inch Hale telescope (then under construction), Sandage (1961) focused almost exclusively on deciding which of the three FLRW solutions or the steady state theory most accurately models the universe. By this point, the focus of observational cosmology (at least as Sandage practiced it) had shifted to measuring important parameters characterizing the FLRW models (as the title of Sandage 1970 has it: "Cosmology: A search for two numbers") rather than testing the applicability of these models.[25]

However, two significant limitations of the FLRW models were widely acknowledged.[26] The first was mentioned by Tolman: the great complexity of the universe could be manifested in a breakdown of symmetry in the early universe. Tolman studied the approach to an infinite

[23] See the first chapter of Peebles (1980) for a brief review of this debate.

[24] Indeed, the most recent data on galaxy counts indicates the presence of structure on the largest scales. For a recent review, see Peacock (1999), Chapter 13.

[25] The two numbers are the values of the Hubble "constant," $H = \frac{\dot{a}(t)}{a(t)}$, and $\Omega_0 = \frac{\rho_o}{\rho_{crit}}$ where $\rho_{crit} = \frac{3H_0^2}{8\pi}$ in geometric units. The subscripts indicate that the quantities in question are evaluated at the present time t_0.

[26] Anderson (1967) §14-8 gives a very clear discussion of both problems.

density "singular state" in a closed FLRW solution in some detail, and he concluded that the idealizations of the model fail to hold as the singular state is approached.[27] Placing the blame for the singular behavior on the "unphysical" symmetries of the FLRW models remained common practice until the mid-60s. The second limitation of the FLRW models was that of explaining the formation of galaxies and clusters of galaxies: the formation of "clumps" and other irregularities is precisely what the idealization of a perfect fluid neglects. Both of these limitations arise from the presumed breakdown of the idealizations of the FLRW models.

The second limitation is connected with a problem which still vexes cosmologists: how does matter cluster into structures such as galaxies or clusters of galaxies?[28] The general problem is to account for how a uniform mass distribution develops "clumps" of higher density of the appropriate size. This requires balancing the attractive force of gravitation with dissipation due to thermal motion at high temperatures and the expansion of the universe. Lemaître showed that small initial wrinkles could grow into prominent irregularities, and proposed a model with a preferred epoch of galaxy formation arranged by temporarily halting expansion with $\Lambda \neq 0$ (Lemaître 1934). Lifshitz (1946) and Bonnor (1956, 1957) developed methods to study the behavior of linear perturbations in an expanding universe, and they both noted that graviational instability could produce galaxies in the available time only if the initial perturbations were remarkably large (as compared to expected statistical fluctuations). This problem spurred the research of Gamow and collaborators into a speculative "primeval turbulence" theory of galaxy formation (Gamow 1952, 1954). Gamow hoped to show that primeval turbulence produced

[27] See Tolman (1934), pp. 438-439, 484-486. I will return to this point in Chapter 2. For a brief discussion of Tolman's research regarding singularities, see Earman (1999).

[28] See also the much more detailed discussions of structure formation in Peebles (1980), Chapter 1, and Kragh (1996), p. 288 ff., both of which I draw on here.

inhomogeneities of a characteristic size, and that structures of this size develop into galaxies.[29] Few cosmologists shared Gamow's enthusiasm for the idea. By way of contrast, Sciama (1955) proposed a promising mechanism for galaxy formation in a steady state theory: the "tidal wake" of a galaxy would lead to an increase in density of the intergalactic gas, which would then collapse to form a galaxy with a characteristic mass.[30]

Hoyle made good use of this lacuna in criticizing the evolutionary models:

> Undoubtedly, the greatest shortcoming of all cosmological theories lies in their failure to provide a working model of the formation of galaxies. Evolutionary cosmology provides no model at all. Galaxies are supposed to arise from initial fluctuations, every necessary property being inserted into the theory as an initial condition. (Burbidge et al. 1963, p. 874)[31]

Although Hoyle admits that the steady state theory also lacks an account of galaxy formation, simply assuming that the initial fluctuations have all the right properties to seed galaxies has all the advantages of theft over honest toil. Dissatisfaction with simply extrapolating current observations back to an initial spectrum of fluctuations also motivated Gamow's approach and a handful of other theories. But in the mid-60s there was no generally accepted theory of galaxy formation in evolutionary cosmology which could produce this initial spectrum of perturbations, or which relied on a mechanism other than gravitational instability to produce clustering.

Gamow's interest in galaxy formation was a byproduct of his detailed study of nuclear reactions in the early universe, which he undertook in collaboration with Ralph Alpher and Robert

[29] Gamow had originally claimed to have a correct theory of galaxy formation in 1948 (Gamow 1948a): during the transition from radiation- to matter- dominated expansion, Gamow argued that gaseous condensations of a characteristic size (determined by Jeans' criterion of gravitational instability) would form. Criticisms raised by Gamow's collaborators Alpher and Herman (Alpher and Herman 1949) (involving both calculational mistakes and fundamental problems with applying the Jeans criteria) led Gamow to abandon this idea.

[30] This idea is also described qualitatively in Sciama (1959).

[31] The body of the paper is concerned with difficulties with a version of Sciama's theory with a "hot" intergalactic gas—so the shortcoming is not limited to evolutionary theories.

Herman with the aim of explaining the observed relative abundances of the elements.[32] This research in turn drew on considerable work done in the late 30s (in which Gamow played a part) regarding nuclear reactions occurring in stars. Studies of stellar processes indicated that the nuclear reactions in stars probably could not produce heavier elements, and in addition the "equilibrium theory" appeared to require a variety of physical conditions at very high temperature in order to reproduce observed element abundances.[33] Although several cosmologists had speculated about the possibility of element formation in the early universe, Gamow recognized the importance of the connection between reaction rates for various nuclear processes and the expansion rate of an FLRW model. On the assumption that the universe originally consisted of a dense neutron gas and photons, Gamow argued that neutron decay and capture would produce heavier elemets as the universe cooled to a temperature such that the average kinetic energy was lower than the binding energy for stable nuclei.[34] Reactions such as the capture of one neutron by a proton ($^1H \rightarrow ^2 H$) occur with high probability when the following condition holds: $v\Delta t n\sigma \approx 1$ (where v is the thermal velocity, Δt is the expansion timescale, n is the baryon number, and σ is the scattering cross section). With knowledge of the scattering cross section and the expansion rate for a radiation-dominated FLRW model, Gamow was able to calculate element abundances. One of the major problems for the Gamow theory was that accounting for

[32] Kragh (1996), Chapter 3, provides a detailed and reliable account of Gamow's theory.

[33] As the name suggests, this approach was based on the assumption that the elements were initially in a state of dynamical equilibrium. Using Boltzmann's equation and observed relative abundances, one can estimate the temperature and density of a "neutron gas" at which the elements would have been in equilibrium. Roughly, the problem with equilibrium theory was that observed abundances of the elements could not have been in equilibrium under similar conditions. Not everyone agreed with Gamow, Alpher, and Herman that this was a serious defect of the theory: see North (1965), p. 256-258 for a brief review of the debate and references.

[34] The original Alpher et al. (1948) paper considers only a neutron gas, but Gamow (1948b,a) include photons.

heavier elements by the process of neutron capture appeared impossible due to the “mass gaps,” the lack of stable nuclei with atomic weights of 5 or 8.

Although it generated a great deal of interest in the late 40s and early 50s (include coverage in the popular press and Gamow’s own popularizations), the Gamow-Alpher-Herman program was nearly forgotten following their last publication on the topic (Alpher et al. 1953). This reception may have been partially due to the mass gap problem, which subsequent elaborations of the idea showed no signs of resolving. But other factors were also undoubtedly important (cf. Kragh 1996). None of the main proponents of the idea continued pursuing research in cosmology: Alpher and Herman took industrial jobs, while Gamow attempted (among other things) to decode the genetic cipher. Research interest shifted to the possibility of nucleosynthesis in other “hot places,” namely the centers of massive stars and supernovae (Burbidge et al. 1957). Perhaps most importantly, the big bang program fell through the cracks between well-established disciplines, in a sense. Very few physicists trained in America in the 50s possessed expertise in both relativistic cosmology and nuclear physics. The early stages of research utilized results and expertise developed in the study of various nuclear reactions as part of the Manhattan Project. Further refinements of the program resulted from carrying out the nuclear physics calculations with much greater care, but there were certainly more promising research projects for those with an expertise in nuclear physics. The big bang program was also not immediately relevant to mainstream lines of research being pursued in astronomy.

In closing, I should emphasize that the three ideas discussed above were not originally introduced as interlocking components of an integrated cosmological theory. A number of further insights were incorporated in what Weinberg (1972) dubbed the “Standard Model” of cosmology, as I will describe in Chapter 2.

1.3 Overview of the Dissertation

This dissertation focuses on methodological issues relevant to the development of early universe cosmology, divided somewhat artificially into three historical and three philosophical chapters. Here I will briefly outline the main themes explored below.

The three historical chapters cover roughly twenty years (1965-1985) in the history of cosmology. As with the early stages of research in many fields, this period is characterized by the gradual development of a consensus regarding the central problems in the field and the methodology and theoretical tools to be used to solve them. Historians are familiar with many cases of theory development driven by empirical anomalies or logical inconsistencies within existing theory, but by the late 60s cosmologists had developed a standard model compatible with observations and free from inconsistencies. What drove early universe cosmology was an ambitious extension of the scope of explanatory questions theorists hoped to answer. In particular, cosmologists hoped to show that many observed regularities of the early universe result from fundamental physics, rather than treating these regularities as features of the universe's initial state.

Chapter 2 describes the development of the standard model and the recognition of several "unnatural" assumptions it requires regarding the initial state; the initial state appeared to be "finely tuned." The account of nucleosynthesis was one of the standard model's great successes: the abundances of light elements produced in the early universe could be calculated using nuclear physics, rather than simply ascribing these abundances to the initial state. This account involved no physics beyond what had been tested in Los Alamos and other nuclear laboraties. By way of contrast, attempts to solve other fine-tuning problems drew on speculative new ideas

in physics. Charles Misner focused on a particularly striking fine-tuning problem: observations of the background radiation indicate that the simple FLRW models are actually incredibly good approximations for the early universe. But why should the early universe be as uniform as the highly symmetric FLRW models? Rather than accepting uniformity as an unexplained feature of the initial singularity, Misner proposed a new dynamical theory of the early universe that produced uniformity as an "output" regardless of the initial state. Several other research groups applied a similar methodology to this and other fine-tuning problems, by introducing various effects, including particle creation in strong gravitational fields and new ideas from particle physics. Although Misner's attempt to explain uniformity dynamically ultimately failed, by the mid-70s a number of cosmologists had adopted Misner's methodology in attempting to solve fine-tuning problems. Misner also clearly identified the presence of particle horizons as an obstacle to solving fine-tuning problems.

The third chapter traces the provenance of an idea that would take center stage in the 80s, namely that the early universe passed through a transient de Sitter-like phase. The differences among the various proposals incorporating this idea reflect the variety of theoretical tools and methodological assumptions prevalent in cosmology at the time. Soviet cosmologists emphasized that a de Sitter-like phase allowed one to avoid an initial singularity, whereas a group of theorists in Brussels argued that it was a consequence of a treatment of the "creation event" itself. These ideas faced two common problems: what was the source of the early de Sitter phase, and how did a transition into the observed FLRW expansion occur? However, these imaginative and quite speculative suggestions did not have links to more well established physical theories that might have fostered their further development. By way of contrast, two other proposals did have such links. Drawing on the active research field of semi-classical quantum gravity, Starobinsky

showed that de Sitter space is an unstable solution to the classical field equations modified by taking quantum effects into account. The second proposal was a by-product of developments in particle physics: successful unification of the electromagnetic and weak forces incorporated a novel treatment of symmetry and the vacuum, a feature carried over to Grand Unified Theories (GUTs). The early universe proved to be one of the few testing grounds for these innovative ideas, and several theorists interested in these developments in the conceptual foundations of field theory turned to cosmology.

Research in the 70s explored a number of possible consequences of the application of GUTs to the early universe, such as the formation of topological defects and the effects of baryon number non-conserving interactions. Chapter 4 describes the introduction and early development of an idea that came to be the almost exclusive focus of early universe cosmology. Alan Guth gave the idea that the early universe passed through a de Sitter-like phase an apt name, "inflationary cosmology," and more importantly provided a compelling argument in its favor. Guth was the first to present a stage of de Sitter-like or "inflationary" expansion as an appealing package deal: three apparently independent features of the early universe—overall uniformity, flatness, and lack of magnetic monopoles—could all be traced to this inflationary stage rather than initial conditions. Guth's presentation (in Guth 1981, and in several talks) did not gloss over his failure to account for a transition to FLRW expansion, and several of his contemporaries took on the task of successfully implementing inflation. Following Guth's work the development of a successful model of inflation was the central problem in early universe cosmology, and within six months Linde (1982); Albrecht and Steinhardt (1982) had both proposed solutions of the "graceful exit" problem. At the Nuffield workshop, inflationary cosmology underwent "death and transfiguration": several theorists independently developed accounts of the production of

density perturbations during inflation, but observational constraints on the amplitude of these perturbations indicated that the field responsible for inflation probably was not the Higgs field of a GUT, as it was in current models. Shortly after the workshop the "inflaton" field was introduced, as the fundamental scalar field with the appropriate properties to drive an inflationary stage. Opinions still differ regarding the importance of a tight connection between the inflaton and particle physics, but several cosmologists have argued that the observational case for inflation is strong enough to warrant introduction of the inflaton regardless of whether it can be identified with an independently motivated fundamental scalar field.

Inflationary cosmology has decisively solved several of the fine-tuning problems of standard big bang cosmology, and in many quarters it is regarded as firmly established. But the status of the fine-tuning problems themselves undercuts the significance of these achievements. One of the appealing features of empirical anomalies is that they are not defeasible, assuming that the theory and the observational or experimental procedures involved are both well understood. Given certain assumptions regarding the distribution of masses in the solar system, accepting Newtonian gravitation forces one to also accept a particular value of Mercury's perihelion advance at odds with observational results. None of this reasoning requires guesswork regarding future theories. By contrast, this is exactly what the fine-tuning problems driving early universe *do* require: the "problems" are defined by a contrast between the presumed "natural initial state" and the observed state. The "initial state" is acknowledged to lie beyond the reach of current theory, and thus postulating its nature involves considerable epistemic risk. This educated guess may turn out to be validated by future theory, but it also may not. The recent ekpyrotic scenario, for example, incorporates a different conception of the initial state that renders the problems solved by inflation superfluous. Whether or not the ekpyrotic scenario stands the test of time, it

serves to illustrate the risk involved in granting great confidence to theories that solve fine-tuning problems.

The philosophical part of the thesis considers various issues connected to the development of early universe cosmology. Chapter 5 tackles the implications of the uniqueness of the universe somewhat obliquely by focusing on three principles: the cosmological, indifference, and anthropic principles. The importance of fine-tuning is often established by appealing to the "indifference principle," namely that a theory which does not require special initial conditions is to be preferred. I argue that this principle requires stronger metaphysical commitments than cosmologists typically admit. I treat the cosmological principle, on the other hand, as a principle of general scope that licenses local to global inferences. Finally, I give a brief deflationary account of anthropic reasoning, emphasizing that weak anthropic principles simply acknowledge selection effects whereas strong anthropic principles make a strong and unwarranted explanatory demand.

Chapter 6 focuses on the role of explanatory adequacy in theory choice. The historical account illustrates that various explanatory virtues of inflation clearly played some role in its widespread acceptance. However, I argue against taking either unification or causation as epistemic virtues that constitute a component of rational theory choice. In the case of unification, even Kitcher's well-developed account fails to handle the tradeoffs that are typically involved in developing a new theory that is unified in two distinct senses: namely, it draws together various phenomena, and it successfully combines distinct theories that both apply to the same domain. I analyze causal arguments relevant to inflation in terms of the "law of conditional independence," then go on to formulate a notion of robustness intended to capture the alleged advantages of

inflation. Finally, I argue that assessments of either explanatory advantage depend upon assumptions about the initial state, and in that sense resemble fine-tuning problems in that they depend upon projections regarding the future course of research.

In chapter 7 I consider the prospects for a straightforward empirical case in favor of inflation. First I assess two problems with the original trio of inflationary predictions: they are not robust, in the sense of holding for all "natural" inflationary models, and they are furthermore not "distinctive," in the sense of differing from what one would expect based on alternative theories of the early universe. By contrast, the comparison of inflationary cosmology with the imprint of density perturbations on the CMBR provides the best possibility for further refining the theory. I will assess the idea of use novelty as applied to this case, and argue that the apparent importance of novelty can be traced instead to the importance of independent constraints on theoretical entities, in this case the inflaton potential.

Part I

History of Cosmology

Chapter 2

The Origins of Early Universe Cosmology

The discovery of the cosmic microwave background radiation (CMBR) "by telephone" in 1965 made the front page of the *New York Times*, above the fold. Readers were treated to a large picture of the horn antenna used to make the discovery, under the headline "Signals Imply a 'Big Bang' Universe" (Sullivan 1965). This discovery (arguably) occured when Arno Penzias telephoned Robert Dicke to inform him of the excess radio noise he and Robert Wilson had observed with the horn antenna, and Dicke recognized the theoretical implications of this radio noise.[1] Dicke's interpretation of the CMBR as the left-over black body radiation from a hot big bang has been almost universally accepted since 1965. According to the big bang theory, the CMBR provides a "snapshot" of the surface of last scattering, when the universe rather suddenly became transparent as free electrons combined with nuclei to form atoms.[2] Prior to this decoupling, the photons and matter were in thermal equilibrium due to frequent scattering reactions between photons and free electrons; after decoupling from the matter, the photons cooled adiabatically with the expansion of the universe.[3] These photons carry a tremendous amount of

[1] See Kragh (1996), Chapter 7, for a detailed account of this episode. Penzias and Wilson were awarded the 1978 Nobel Prize in Physics for their observational work, despite their initial doubts regarding its theoretical implications. Several observers detected the background radiation by a variety of methods prior to Penzias and Wilson (including Le Roux, Shmaonov, Ohm, and McKellar), but the signifi cance of these observations was not recognized prior to Penzias's phone call.

[2] Decoupling occurs at about 10^{13} seconds after the big bang by current estimates, when the scattering cross section for the photons drops from that due to Compton scattering to that due to scattering by neutral hydrogen and helium, which is much lower.

[3] The effect of this adiabatic expansion is to lower the temperature of the black-body spectrum of the photons: $T \propto 1/a$ (where a is the scale factor). This effect of expansion was known long before the discovery of the CMBR (see Tolman 1934, , §171).

information about the very early universe to astronomer's radio telescopes. Researchers working in cosmology and related fields at the time still describe their initial reaction to these observations with remarkable clarity.[4] Papers published soon after the discovery betray the excitement of the field: the data-starved field of cosmology had discovered a "Cosmic Rosetta Stone," which several research groups planned to use to decipher the early universe's history. Indeed, mainstream cosmologists often give May of 1965 as the birthdate of "scientific" cosmology. The idea that scientific cosmology began in 1965 has become a firmly entrenched component of the popular mythology of the field, reiterated in review articles, popularizations, and even (more surprisingly) in some history of science articles (such as Brush 1993). However, reserving the honorific of "scientific" for post-1965 cosmology betrays a serious misunderstanding of the field; precise observational work in cosmology, usually taken to be the demarcation criteria in these discussions, certainly existed before 1965.[5]

The sparse scholarly literature on the history of cosmology typically emphasizes the importance of this discovery in ending the debate between the big bang and steady state theories. This emphasis is misleading in two respects. Everyone who could be convinced of the falsity of the steady state theory by recalcitrant observations abandoned the theory for other reasons, and even the challenge of accomodating the CMBR observations has not triggered apostasy among the remaining stalwarts.[6] But more importantly, the CMBR convinced researchers that the early

[4]See, for example, Misner (2001); Peebles (2002). Misner and other theorists were much less cautious than more observationally oriented researchers in taking the early results to establish a nearly isotropic thermal background; as Peebles (2002) emphasizes, measurements of the short wavelength portion of the spectrum (which would test the turn-over to the Wien law characteristic of the black body spectrum) were not well established until much later, and ultimately COBE provided the first fully convincing evidence for a thermal spectrum in 1992.

[5]Kragh (1996) convincingly debunks this myth.

[6]Kragh (1996), Chapter 7, gives a detailed account of the response of various advocates of the steady state theory to the CMBR observations. Among those who explicitly converted to the big bang theory, such as Sciama and McCrea, the CMBR did not play a decisive role. On the other hand, many one-time advocates of the theory drifted out of cosmological research and into other fields, without ever explicitly

universe, once thought to be the subject of wild extrapolations, unfounded speculations, and little else, could be brought within the realm of responsible theorizing. Within two years after the discovery of the CMBR, several research groups began to develop detailed theoretical accounts of the first few minutes of the universe's history. Writing in 1977, Steven Weinberg commented that

> [Prior to discovery of the CMBR]...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, pp. 131-132)

Taking the early universe seriously led to a flurry of research in cosmology in the late 60s and early 70s.

This chapter describes the field of cosmology in the late 60s and 70s, focusing on the development of early universe cosmology shortly after the discovery of the CMBR. Several factors contributed to the rapid growth of interest in this speculative field. As I have already emphasized, the CMBR offered a rich, accessible source of information about the early universe; it is typically described as *the* secure observational fact in cosmology. The impact of this new discovery is similar to several other cases in the history of astronomy, such as the development of spectroscopy in the late 19^{th} and early 20^{th} century, when advances in instrumentation and technology opened up entirely new research areas. Ever more precise observations of the CMBR have been the empirical touchstone for early universe cosmology, and the possibility for detailed comparisons between these observations and early universe theories has been an important factor in originally inspiring and sustaining interest in the field. A second important factor was the

endorsing the big bang theory. Hoyle, Narlikar, and a handful of collaborators have continued to propose versions of the steady state theory, but I agree with Kragh's assessment: this research has had no impact on mainstream cosmology.

development of a widely accepted framework for cosmology at later times, the hot big bang model. In 1972 Weinberg dubbed this the "Standard Model" of cosmology (Weinberg 1972), and it drew on the ideas discussed in the previous chapter, combined with a number of insights from subsequent research. I will focus on the development of the standard model in §2.1 below. The overall success of the standard model encouraged theorists to consider more speculative extensions of it approaching the "absolute zero of time" (borrowing Misner's phrase). These extensions involved a wealth of interesting theoretical problems in classical general relativity, quantum theory, and quantum gravity, and they also addressed inadequacies of the standard model. Finally, the discovery of the CMBR came in the midst of a renaissance in the study of general relativity. General relativity was not an active part of mainstream physics for several decades, due in part to the fast-paced and well-funded development of post-war particle physics. The late 50s and early 60s saw a gradual end to this period of isolation and stagnation. Many of the physicists responsible for this resurgence of interest in relativity, such as Hawking and Misner, also focused on early universe cosmology. The participation of these prominent figures, along with the increasing interest in early universe cosmology among particle physicists (such as Weinberg), legitimated the field for a generation of physicists.

In general terms, a number of research groups pursued two different approaches: one explored possible relativistic effects due to more realistic (less symmetric) models of the early universe as it emerged from the initial singularity, while the other approach focused on the results of introducing more realistic models of the matter content of the early universe (usually in a fixed background spacetime). These approaches frequently overlapped—for example, in studies of the impact of a specific term in the stress-energy tensor on the evolution of a cosmological model (such as Misner's neutrino viscosity). Both approaches aimed to illuminate any one of several

pressing problems: the surprising isotropy of the universe, the formation of galaxies, the amount of entropy in the early universe (measured in terms of the number of photons per baryon), and the asymmetry between matter and antimatter. All of these problems derive from a common source: ignorance of the nature of the universe after it emerged from the initial singularity.

2.1 Development of the Standard Model

Historians often take the development of a widely accepted common framework to be a hallmark of a "mature" science. The development of such a framework, the hot big bang model, culminated in the late 60s with the combination (and, in the case of nucleosynthesis, rediscovery) of several ideas developed decades earlier, discussed in more detail in the previous chapter.[7] According to the hot big bang model, the large scale structure of the universe and its evolution over time are aptly described by the simple FLRW models. Extrapolating these models backwards leads to a hot, primeval "fireball," the furnace that produced both the CMBR 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, although in the 60s there was little consensus regarding how to amend the FLRW models in order to give a more detailed account of structure formation. All of these ideas had been pursued prior to 1965, but review articles from 1966 onward show a new confidence that these ideas fit together as part of a consistent overall theory compatible with observations. These reviews often cite the discovery of the CMBR as the main source of confidence in the hot big

[7] Although the big bang model clearly has its roots in the work of Lemaître, Gamow and others, there are a number of differences between these earlier ideas and the big bang model accepted in the late 60s. North (1965)'s comprehensive history, written before the acceptance of the 'standard model," is instructive in this regard: although he discusses Gamow's theory of element formation and the FLRW models, these are never presented as components of a larger, comprehensive theory (cf. Kragh 1996).

bang model; several reviews echo Zel'dovich and Novikov (1967)'s conclusion that the hot big bang model seems inescapable in light of the CMBR. With the marginalization of the steady state theory, no alternative theory had widespread support among mainstream cosmologists.

After 1966 cosmologists mostly agreed that the hot big bang model outperformed rival cosmological models, and in this sense the field reached consensus on a Kuhnian paradigm. But the big bang model still constituted only a fairly loose framework, not nearly as tightly constrained by observations or internal consistency as other theories in astronomy and physics. Two early, influential presentations of the big bang model both stressed its tentative nature. Although Weinberg (1972) called this theory the "standard model" of cosmology, he emphasized its utility, rather than any conviction of its truth, as a reason for pursuing it further:

> Of course, the standard model may be partly or wholly wrong. However, its importance lies not in its certain truth, but in the common meeting ground that it provides for an enormous variety of observational data. By discussing these data in the context of a standard cosmological model, we can begin to apprecate their cosmological relevance, whatever model ultimately proves correct. (Weinberg 1972, p. 470)

Peebles made a similar point in the introductory remarks to his influential textbook:[8]

> There is the point of view that in a science as primitive as cosmology one ought to avoid orthodoxy, give equal time to all competing cosmologies. ... My own preference is to make a subjective selection of a reasonably possible cosmology, and then study it in all the detail one can muster. The end result of such an effort may be to reveal that the cosmology is logically inconsistent or even in conflict with observation, which is progress of a sort. The hope is that the development of the observations may convince us of the rough validity of the chosen cosmology, and guide us to how the cosmology should evolve. (Peebles 1971, p. viii)

[8] It is hard to imagine that many cosmologists would have disagreed with Peebles "subjective" selection of a preferred cosmological model in the 70s; most contemporary review articles discussed the steady state theory and other alternative theories only in relation to the history of the field, and not as viable alternatives to the hot big bang model.

Weinberg and Peebles both emphasize the importance of accepting the big bang theory as a common framework for organizing the relevant observational data while down-playing the degree of certainty attributed to the theory. This caution partially resulted from a keen awareness of the numerous idealizations built into the hot big bang model. In this section I will discuss two cases of "paradigm articulation" that helped to eliminate some of these idealizations. The first is the development of new tools in relativistic cosmology that allowed the study of (some aspects of) general cosmological models, rather than the maximally symmetric FLRW models. Second was the development of an account of the thermal history of the early universe stretching back to the "hadron era."

2.1.1 Relativisitic Cosmology

The surge of research work in cosmology in the late 60s drew upon and reinforced renewed interest in the general theory of relativity (GTR). Throughout most of the 40s and 50s relativity theory remained isolated from other areas of physics in American physics departments. General relativity was not even taught in the top physics departments in the country (such as MIT, Berkeley, Princeton, Columbia and Harvard), and it was definitely not included as part of the "core curriculum" for physics graduate students.[9] Instead it was regarded as a mathematical subject, as a remark by Victor Weisskopf to Robert Dicke indicates: "I asked Victor Weisskopf one time... shouldn't a graduate student pay attention to relativity? And he explained to me

[9]Peter Bergmann started an active relativity group at Syracuse University in the late 40s, but the leading physics departments did not include general relativity as part of the graduate curriculum until much later. John Wheeler began teaching a graduate course in general relativity at Princeton in 1952, and the other leading departments followed suit more than a decade later. See Kaiser (1998) for a discussion of how and where general relativity was taught during this time period.

that it really had nothing to do with physics. Relativity was a kind of mathematical discpline."[10] Much of the work on relativistic cosmology in the 40s and 50s satisfies Weisskopf's description: studies of the mathematical features of various cosmological models (such as Taub 1951) often appeared in math journals. Regardless of where it appeared, research devoted to general relativity lagged far behind the explosive overall growth of physics: during a period (1939-1956) in which the *Physical Review* grew in size by over 600% (Kevles 1977), the output of papers related to relativity stayed constant at roughly 30 papers per year (Eisenstaedt 1986).

There are several reasons for this overall stagnation in relativity theory.[11] Perhaps the most important was the combination of daunting mathematical complexity and the small magnitude of general relativistic corrections to Newtonian theory. Despite the major conceptual differences between general relativity and Newtonian theory, for almost all applications of gravitational theory the Newtonian approximation remains valid, or can be adjusted by including relativistic corrections. Many physicists were willing to join Max Born in admiring relativity from a distance rather than learning the intricacies of the theory:

> I learned it [general relativity] not only from the publications but from numerous discussions with Einstein, which had the effect that I decided never to attempt any work in this field. The foundation of general relativity appeared to me then, and it still does, the greatest feat of human thinking about Nature, the most amazing combination of philosophical penetration, physical intuition, and mathematical skill. But its connections with experience were slender. It appealed to me like a great work of art, to be enjoyed and admired from a distance. (Born 1956, p. 253)

Starting in the 50s several experimental groups set out to strengthen general relativity's connections with experience by finally improving upon the three classical tests of the theory. Luckily

[10] Dicke, recalling a conversation with Weisskopf from his graduate school days in the early 40s, in an interview recorded in Lightman and Brawer (1990), p. 204.

[11] See, in particular, Eisenstaedt (1986) and Will (1993) for discussions of this stagnation and the subsequent renaissance.

Robert Dicke did not follow Weisskopf's advice, and instead started a group at Princeton devoted to more rigorous experimental testing of general relativity. This revived interest in experimental tests of the theory prompted Schild to comment that "Einstein's theory of gravitation ... is moving from the realm of mathematics to that of physics" (Schild 1960, p. 778).

Many of the scientists in newly founded gravity groups focused on cosmology, and in contrast to the more mathematically oriented relativistic cosmology this "physical cosmology" drew heavily on other areas of physics as well as observational astronomy. For example, the wealth of physical and observational detail in Peebles' influential 1971 textbook (Peebles 1971) contrasts sharply with earlier cosmology textbooks such as McVittie (1956), which treat cosmology primarily as the study of a handful of exact solutions to EFE. As the community of researchers working on general relativity in the 50s and 60s grew, the relativist's repertoire of mathematical techniques also grew considerably larger. New techniques imported from differential geometry helped to elucidate one of the outstanding problems of relativistic cosmology: the nature of the initial singularity in FLRW models, and whether similar singularities occur in less symmetric models.

A debate about the nature and existence of singularities in cosmological models drew the attention and research efforts of some of the leading relativists of the 60s. Early research by Tolman, Lemaître and others in the 30s established the existence of an initial singular state in the FLRW models, but this was taken to indicate a limitation of the models rather than a feature of the early universe. Tolman argued that the presence of a singular state reflects a breakdown of the various idealizations of the FLRW models.[12] The first strong indication that

[12]Tolman faults both the perfect fluid idealization, and, taking a cue from Einstein, the assumption of homogeneity (Tolman 1934, p. 438 ff.). For a historical study of research regarding singularities, see Earman (1999) and references therein.

loosening the symmetries built into FLRW models would not eliminate the initial singularity appeared in Raychaudhuri (1955): Raychaudhuri showed that with $\Lambda = 0$, for a non-rotating "dust" solution (with energy density $\rho > 0$ and $p = 0$) described in a synchronous coordinate system, $g = det(g_{ab}) \rightarrow 0$ at some finite time in the past, whether or not isotropy holds. This result establishes that timelike geodesics cross, which Raychaudhuri interpreted as signalling the presence of a physical singularity.

A number of Russian researchers (Landau, Lifshitz, Khalatnikov, and their numerous collaborators) disputed the significance of Raychaudhuri's result, and set out to analyze the general form of cosmological solutions to EFE in the neighborhood of the alleged singularity, with the hope of showing that the "singular solutions" depend upon a specialized choice of the initial matter distribution and gravitational field. More precisely, the Russians 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 specific initial conditions. The Russian 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) , although they ultimately failed to show that the initial singularity is an avoidable disaster.

The Russians reluctantly abandoned this goal only after Penrose, Hawking, and Geroch established the celebrated singularity theorems. These theorems establish that singularities, as signalled by the presence of incomplete geodesics,[13] are a generic feature of solutions to EFE

[13] An incomplete geodesic is inextendible in at least one direction, but does not reach all values of its affi ne parameter.

satisfying certain plausible assumptions. Hawking and Ellis (1968) prove a singularity theorem based on the following assumptions:[14]

(1) GTR is valid.
(2) The strong energy condition holds everywhere.
(3) The spacetime is strongly causal.
(4) There exists a point p such that all the past-directed timelike geodesics through p start converging again within a compact region in the past of p.

The second and third conditions characterize the stress-energy tensor and the global causal structure of spacetime.[15] The singularity theorems generally take the form of a *reductio* proof, showing that geodesic completeness of a congruence of timelike geodesics is inconsistent with assumptions similar to those listed above. Hawking, Penrose, and others used the same geometrical methods to prove a number of theorems differing in the list of assumptions. Hawking and Ellis (1968) chose the set of assumptions above since (4) holds if an inequality expressed in terms of an integral of the stress-energy tensor (over all past-directed timelike geodesics through p) is satisfied.[16] The energy density of the CMBR alone (neglecting the other matter and energy in the universe) satisfies this inequality, so the observed universe satisfies the assumptions of the theorem (although assumption (1) presumably fails in the very early universe). Although the singularity theorems famously do not reveal the nature of the initial singularity (as Russian critics emphasized), they establish that the initial singularity is not an artifact of symmetry assumptions. Thus, from 1965 onward relativists interested in the early universe had ample reason

[14] Recall that the strong energy condition is satisfied if the stress-energy tensor for each matter field satisfies $T_{ab}U^aU^b \geq -1/2T^a_a$ for any unit timelike vector U^a. A spacetime is strongly causal if for every p in the spacetime, every neighborhood $O \ni p$ contains a subneighborhood U cut no more than once by any timelike or null curve.

[15] The need to characterize the global structure of spacetime stimulated research into the causal structure of relativistic spacetimes. See Hawking and Ellis (1973) for an early, comprehensive discussion of the singularity theorems.

[16] See Hawking and Ellis (1968) and the discussion in §10.1 of Hawking and Ellis (1973).

to take the presence of an initial singularity seriously, even though the nature of this singularity continued to inspire debate and further research.

The singularity theorems are one example of attempts to move beyond highly symmetric solutions in order to understand the general features of cosmological models. Research throughout the 50s and 60s increased the number of known exact solutions to Einstein's field equations. Although the surprising isotropy of the CMBR appeared to vindicate cosmologist's reliance on the isotropic and homogeneous FLRW models, several groups explored the behavior of anisotropic solutions near the initial singularity. For the Russian program mentioned above, this research was an essential part of attempts to understand the initial singularity. Misner's program (discussed in detail below) hoped to show that a general anisotropic solution would quickly "smooth out" in the first few moments after the big bang. Peebles (1972) contrasted this "revolutionary" approach—with its assumption that the very early universe departs dramatically from the FLRW solutions, but various physical processes smooth out this primeval turbulence—with a "conservative" approach. The conservative approach focused on the problem of galaxy formation, usually studied in terms of perturbations on an FLRW background. This work drew on inhomogeneous exact solutions, combined with an application of relativistic hydrodynamics to the early universe in order to study the evolution of small perturbations in the otherwise uniform matter distribution.

2.1.2 Thermal History

Applications of nuclear physics and statistical mechanics to the early universe led to an increasingly detailed account of the thermal history of the universe, including, in particular, an account of light element nucleosynthesis. As discussed in the last chapter, Gamow, Alpher,

and Herman's pioneering work in nucleosynthesis did not lead to a sustained research program. However, interest in nucleosynthesis had rekindled by 1965. Burbidge, Burbidge, Fowler, and Hoyle had amended the Gamow theory of nucleosynthesis with their important account of stellar nucleosynthesis, and Peebles independently developed an account of big bang nucleosynthesis in 1965. But in addition to the conceptually straightforward (though computationally difficult) application of well-understood nuclear physics to the early universe, the early universe provided a testing ground for the various possible models of the strong interaction considered by particle physicists. In the late 60s and early 70s particle physicists had already recognized the utility of the "poor man's accelerator" (Zel'dovich's term for the early universe) for testing speculative ideas in particle physics.

In the late 60s and early 70s a great deal of effort was devoted to calculating the relative abundances of light elements (^1H, ^2H, ^3He, ^4He, and ^7Li) produced in the early universe. Despite the difficulty of these calculations, the theory of light element synthesis drew almost entirely on relatively well understood nuclear physics. Unlike Gamow's theory, these cosmological theories of light element synthesis were supplemented by the theory of heavy element synthesis in supernovae presented in Burbidge et al. (1957). Hoyle and Tayler published a helium abundance estimate in 1964, shortly before the discovery of the CMBR, and Peebles independently rediscovered the basic ideas of Gamow's big bang theory at roughly the same time. Two years later Wagoner et al. (1967) published a detailed account of light element synthesis at high temperatures (Hoyle favored the idea that these high temperatures obtained in the cores of supermassive objects, rather than in a hot big bang). The calculated abundances were compatible with the weak observational limits on helium and deuterium abundances in the late 60s; continued (and quite substantial) observational work throughout the 70s, along with further refinement

of the theory, lent increasingly strong support to the theory.[17] The sensitivity of these element abundances to the properties of the early universe at the time of nucleosynthesis ($t \approx .01$ seconds to 3 minutes after the big bang) provides a probe of the universe long before the emission of the CMBR (at the decoupling time, $t_d \approx 300,000$ years).

One of the crucial ideas underlying these nucleosynthesis calculations is the connection between the expansion rate and the interaction rate (cross section) for various reactions. In the 60s this idea was applied more broadly, in particular to the possible "freeze out" of particle species. Suppose that a particle species (say, neutrinos) maintains thermal equilibrium with other particle species via an interaction with an interaction rate Γ (per particle). For a radiation dominated FLRW model, H scales with temperature as $H = -\frac{\dot{T}}{T}$, and Γ also typically depends upon the temperature.[18] As Zel'dovich emphasized in 1965, if the interaction rate falls below the universe's expansion rate, i.e. $\Gamma \leq H$, then the species will depart from thermal equilibrium as the interactions needed to maintain equilibrium become incredibly rare. More detailed calculations of the "freeze out" of a number of particles were carried out (by solving the Boltzmann equation numerically) throughout the late 60s and early 70s.

Extrapolations backwards in time eventually reached temperatures beyond the domain of applicability of nuclear physics. At temperatures of roughly 10^{12} K standard statistical mechanics and nuclear physics are expected to break down. Below this temperature, the sea of particles can be treated with some degree of accuracy as an ideal gas (for particles with masses $m << kT$; recall that $c = 1$). But at 10^{12} K statistical mechanics certainly may not apply:

[17] See, for exmaple, Peebles (1971) (Chapter VIII) for the tentative nature of the evidence as of 1971, as compared to the much stronger support reported in Peebles (1993), pp. 184-196.

[18] The temperature is inversely proportional to the expansion rate $a(t)$. For an FLRW model with a particular equation of state, one can find an expression for the temperature as a function of time.

baryons, mesons and other strongly interacting particles present in the early universe would have been tightly packed together, with an average interparticle distance on the order of the Compton length (see Weinberg 1972). The theory covering strong interactions would need to be utilized in order to understand this state of matter and the interactions between particles could no longer be neglected. When cosmologists first discussed this problem in the late 60s, particle physicists were still actively engaged in developing a theory of the strong interactions. The lack of a well established framework left a number of possibilities open, but research seems to have focused on two different approaches: the parton or quark model, which takes baryons to be constituted by elementary particles, and the "composite particle model."

Hagedorn (Hagedorn (1970) and references therein) developed a composite particle model with several interesting consequences for early universe cosmology.[19] Hagedorn aimed to replace calculations based on the dynamics of the strong interaction with a thermodynamic approach. This approach requires two important assumptions. First, Hagedorn states the "basic postulate" of his approach as follows: "there are no elementary hadrons, each of them consists of all others" (Hagedorn 1970, pp. 186, 188)—in other words, Hagedorn treats the proton as on a par with all slowly or rapidly decaying resonant states (neutrons, pions, etc.). As a consequence of this assumption, all different possible hadronic states have an entropy which depends only on mass, and there are no postulated internal structural differences between these particles (in contrast to the parton or quark model, which treats the resonances as bound states of elementary particles). Hagedorn then calculates the properties of a sea of hadrons, treated as non-interacting particles, based on the partition function. One needs to specify the particle density $\rho(m)dm$

[19] In all the discussions I have found of a composite particle model, Hagedorn's theory is the only such model discussed. See, for example, Weinberg (1972), §15.11, and Zel'dovich and Novikov (1983), Chapter 6.

(which fixes the number of particle states between m and $m + dm$) in order to use the partition function, and Hagedorn argues that

$$\rho(m) \rightarrow Cm^{-5/2} exp\left(\frac{m}{kT_0}\right) \tag{2.1}$$

in the limit as $m \rightarrow \infty$, where C and T_0 are constants to be determined by observation and k is Boltzmann's constant. An interesting consequence of a particle spectrum of this type is the existence of a maximum temperature T_0, since the partition function is defined only for temperatures $T < T_0$. This feature apparently motivated Hagedorn's approach—he refers to (controversial) evidence of a maximum temperature in recent accelerator experiments.[20] In the early universe, the diverging energy density (as $t \rightarrow 0$) cannot be accounted for as increasing kinetic energy of particle species, since $T \rightarrow T_0$; instead, given the asymptotic behavior of $\rho(m)$ the number of more massive species diverges logarithmically as $T \rightarrow T_0$. A second important consequence of Hagedorn's model is that the expansion rate during the hadron era is given by $R \propto t^{2/3} |\ln t|^{1/2}$.[21]

Since I will discuss the "elementary particle" model in more detail in the next Chapter, here I will only briefly note its differences from Hagedorn's theory. Rather than counting all hadrons and their resonances as equally fundamental particles, according to the elementary particle model all particles are constituted by a short list of elementary particles (say, photons, leptons, quarks, and the appropriate anti-particles). If the strong interactions between these elementary particles can be neglected, then for $kT >> m$, when the particles' masses can also be

[20] See Zel'dovich and Novikov (1983) for a discussion of the experimental data, and an alternative explanation which does not involve a maximum temperature.

[21] See Hagedorn (1970), pp. 196 ff., or Weinberg (1972) for a much clearer derivation.

neglected, the elementary particles have the same properties as black body radiation. The overall energy density of the universe then depends upon the number of particle species—assumed to be a small number—and their masses. In particular, the equation of state is given by $\rho = 3p$, and the scale factor in an FLRW model evolves according to $R(t) \propto t^{1/2}$. Thus these two different models led to very different consequences for the evolution of the early universe.

2.2 Fine-Tuning Problems of the Standard Model

Within less than a decade after the initial observations of the CMBR, the research efforts of a growing community of cosmologists had led to a well-developed standard model, accepted (at least in general terms) by mainstream cosmologists. Review articles from the mid-70s praise the standard model, but also emphasize a number of shortcomings. One general weakness of the standard model stemmed from a variety of implausible conditions required of the initial state of the universe: the initial state apparently needed to be incredibly "finely-tuned" in order to yield something like the observed universe.

Misner discussed in detail the most striking "finely-tuned" feature of the universe: its isotropy and homogeneity at early times. At least four other features of the universe were widely cited throughout the 70s as examples of fine-tuning. The theory of galaxy formation required stipulating a specific initial spectrum of density perturbations which have since evolved into galaxies, clusters of galaxies, and so on. Peebles (1980) emphasizes the value of understanding the evolution of large scale structure well enough to specify the features of this required initial spectrum; he regards this as a useful input for developing theories of the early universe, but he also expresses hope that the spectrum might be understood as the result of fundamental physics. A second example of fine-tuning involves entropy, measured in terms of the baryon to photon

ratio $\eta = n_b/n_\gamma$ where n_b is the number of baryons and n_γ is the number of photons.[22] In the theories of light element synthesis, η is generally treated as a free parameter. Although ^4He abundance is largely independent of η, the abundances of ^2D and ^3He are more sensitive to η; thus, observations of element abundances place fairly tight constraints on the entropy per baryon, at the time usually quoted as $\eta \leq 10^{-7}$; present estimates (based on energy density of the CMBR) indicate $\eta \approx 10^{-8}$. In the 70s interest focused on this ratio as an indication that only limited dissipation (due to viscosity or other mechanisms) could have occured in the early universe (see, e.g., Barrow and Matzner 1977). Although the finite value of η serves as an upper limit, the amount of entropy struck most theorists as remarkably high—and it provided a clue to possible dissipative reactions or other entropy-producing interactions in the hadron era. A third example of fine-tuning is the observed present asymmetry between matter and antimatter. Harrison and several others discussed the possibility that a symmetric early universe could have produced local concentrations of matter and antimatter separated by large distances (thus minimizing observational consequences of matter / antimatter annihilation). Finally, the problem Alan Guth dubbed the "flatness problem" was emphasized by Dicke in Dicke (1969) and in Dicke and Peebles (1979): roughly, the present observational limits on the energy density of the universe imply that the early universe was incredibly close to the flat FLRW model.

This section focuses on the three most influential early attempts to solve these fine-tuning problems: Misner's study of neutrino viscosity, research regarding particle creation in the early universe, and early research regarding baryogenesis. All three postulated new physics which would eliminate finely-tuned initial conditions, and thus trace the observed features of the universe to new fundamental physics rather than the nature of the initial singularity. Even though

[22] See, e.g., Chapter 15 of Weinberg (1972) for a detailed discussion.

the first two of these three attempts did not lead to a widely accepted solution of fine-tuning (and the jury is still out on the third), this *methodology* carried over to later research.

2.2.1 Misner's Chaotic Cosmology

Soon after the discovery of the CMBR, Charles Misner launched what came to be called the "chaotic cosmology"[23] program based on the idea that understanding the dynamics of the early universe may eliminate the need to stipulate initial conditions. Misner's lengthy paper devoted to the isotropy of the universe was motivated by the sense that

> [The isotropy of the CMBR] surely deserves a better explanation than is provided by the postulate that the Universe, from the beginning, was remarkably symmetric. (Misner 1968, p. 431)

Misner is quite explicit about this methodological shift:

> I wish to approach relativistic cosmology from an unfamiliar point of view. Rather than taking the unique problem of relativistic cosmology to be the collection and correlation of observational data sufficient to distinguish among a small number of simple cosmological solutions of Einstein's equations, I suggest that some theoretical effort be devoted to calculations which try to "predict" the presently observable Universe. [...] The difficulty in using relativistic cosmology for predictive rather than merely descriptive purposes lies in the treatment of initial conditions. [...] Ideally one might try to show that almost all solutions of the Einstein equations which lead to star formation also have many other properties compatible (or incompatible!) with observation. More modest but more feasible approaches would attempt to survey much more limited classes of solutions of the Einstein equations to see whether some presently observable properties of the Universe may be largely independent of the initial conditions admitted for study. (Misner 1968, pp. 432-33)

The more modest goal of Misner (1968) is to determine whether physical processes operating in the early universe could effectively smooth out initial anisotropies before the decoupling time

[23] Linde has since appropriated this term to describe his version of inflation. Throughout this chapter, chaotic cosmology refers to Misner's approach.

t_d. "Prediction" in Misner's sense would amount to showing that all solutions to EFE satisfying other observational constraints share the property of isotropy at t_d (or even earlier times).

Misner has also characterized this methodological shift in terms of its explanatory superiority. There are a bewildering variety of cosmological models (exact solutions of Eistein's field equations), and even introducing a number of reasonable physical constraints does not narrow the scope of possible models to something similar to the observed universe. As Misner puts it, "How can you say you've explained the present universe if you can just say, 'Well, there are millions of possibilities and we're one of them'?" The chaotic cosmology program would provide a more satisfying explanation, according to Misner, in that it would show that the observed universe is "the natural kind of universe to have" rather than simply one among many possibilities (Misner 2001).

The properties of anisotropic, homogeneous models were studied extensively prior to Misner's work, and the mid-60s saw a brief resurgence of interest in anisotropic models. If one relaxes the requirement of isotropy in a homogeneous cosmological model, the different spatial directions are no longer equivalent even though all points on a hypersurface remain equivalent. Anisotropic homogeneous models can be classified according to the symmetry properties of the three-dimensional hypersurfaces, a project originally undertaken in Taub (1951).[24] Prior to the discovery of the CMBR, Kristian and Sachs (1966) calculated the distortion effects of anisotropies and inhomogeneities on a pencil of light rays reaching an observer. They established that any such distortion effects were far too small to be detected with current levels of observational accuracy. They used anisotropic models only to test the validity of isotropy. Two

[24]The models are classified into Bianchi types I through IX, following a study of all three dimensional Lie groups completed by Bianchi in 1897. See Wald (1984) §7.2 for a concise introduction to the Bianchi models.

observational results (both of which proved to be spurious) led physicists to study anisotropic cosmological models in the mid 60s. Standard big-bang nucleosynthesis in an FLRW model yields a primeval helium abundance of $\approx 25 - 30\%$ by mass, but several observations suggested that the actual helium abundance was much lower (see Thorne (1967) for references). Hawking and Tayler (1966) and Thorne (1967) showed that anisotropy at temperatures $\approx 10^9 K$ would substantially decrease overall helium production. Thus, later confirmation of He abundances in the range of $\approx 25 - 30\%$ placed limits on the amount of anisotropy present during nucleosynthesis. The second hint of anisotropy was an observation of a set of 13 QSO's by Faulkner and Strittmatter (quasi-stellar objects, now called quasars), which were clustered around the galactic poles (rather than being uniformly distributed).

Misner recorded the impression that Faulkner and Strittmatter's QSO observations had on him in a research journal he kept during his year at Cambridge (for the date of 16 November of 1966):

> Last night Faulkner and Strittmatter showed me their blackboard globe... showing that the $z > 1.5$ quasars (all 13 of them) were all ... near the galactic poles. Although this appears most likely either an ununderstood observational bias, ... we felt that calculations showing its (probably extreme) implications if interpreted in terms of anisotropic cosmologies, should be carried out, and today set about this. (Misner 1994, p. 320)

Misner was already familiar with the Bianchi models (see Misner 1963), although he had regarded this earlier work as purely mathematical. Later that year (1966) he was able to apply this extensive background to the physical problems suggested by the QSO observations, namely the isotropization of initially anisotropic models (Misner 2001).

A slight variant of the Kasner solution offers a good example of the "isotropization" which Misner hoped would hold quite generally. The Kasner solution is a relatively simple

anisotropically expanding homogeneous model, with the following line element:

$$ds^2 = -dt^2 + t^{2p_1}dx^2 + t^{2p_2}dy^2 + t^{2p_3}dz^2 \tag{2.2}$$

The variables appearing in the exponents satisfy the following relation:[25]

$$p_1 + p_2 + p_3 = (p_1)^2 + (p_2)^2 + (p_3)^2 = 1 \tag{2.3}$$

The constraints imply that at least one of the variables is non-positive. The p_i's determine the velocity of expansion along the coordinate axes of a set of co-moving coordinates, and test particles follow timelike geodesics with constant (x, y, z). The expansion along the x-axis measured between two neighboring test particles is proportional to t^{p_1}, and the volume measured in a spacelike surface changes with time proportionally to $t^{p_1+p_2+p_3} = t$. The Kasner solution is a vacuum solution, but as Schücking and Heckmann (1958) showed introducing matter into the solution causes the anisotropic expansion to smooth out into a flat FLRW model. More precisely, Schücking and Heckman found a Kasner-like solution for pressureless dust, $T_{ab} = \rho U_a U_b$. As $t \to 0$ the dust model can be approximated by the vacuum solution (since curvature effects dominate the solution at early times), but for later times the matter dominates the dynamics and drives the solution towards isotropic expansion. The matter terms cause the model to "smooth out" as $t \to \infty$.

Misner's proposal relies on a similar property which he thought would hold for more general anisotropic cosmologies: introducing a term in the stress-energy tensor to account for

[25] See Misner et al. (1973), p. 801, Hawking and Ellis (1973), pp. 142-44, or Ryan and Shepley (1975), pp. 159-162 for discussions of the Kasner metric. The constraint follows from EFE, assuming that T_{ab} is homogeneous or zero.

"neutrino viscosity" causes the solutions to isotropize and approach the $k = 0$ FLRW solution.[26] One important problem facing Misner's proposal was that any initial anisotropies had to disappear incredibly quickly in order to meet stringent limits on isotropy implied by big bang nucleosynthesis and CMBR observations. Calculations in Thorne (1967) showed that significant anisotropies at about 300 seconds after the big bang would lead to insufficient He production, and the CMBR observations placed more stringent limits on isotropy at the later decoupling time, $t_d \approx 10^{13}$ seconds. Misner expected that correctly incorporating the effects of neutrino viscosity would lead to rapid isotropization at temperatures of roughly $10^{10}K$, when neutrinos decouple from the matter in the early universe, prior to nucleosynthesis. The overall effect of neutrino viscosity is to slow down expansion by converting the energy associated with expansion into thermal energy. Misner (1968) treats the effects of anisotropy in terms of an "anisotropy energy density" which measures the energy "stored" in the shear anisotropy. As a somewhat simplified case, suppose that a radiation-filled region expands anisotropically, with contraction along the x axis and expansion along the y axis.[27] For a collision-dominated fluid, the effects of anisotropic expansion will be negligible since the particles exchange energy through frequent collisions. Such a fluid will simply cool adiabatically as the universe expands. In contrast, when neutrinos decouple from the surrounding matter and radiation (at around $10^{10}K$) for a brief period they will have a scattering cross section that is small enough so that the mean free path is relatively

[26] Misner focuses on Bianchi type I cosmologies, with a metric given by $ds^2 = -dt^2 + e^{2\alpha}(e^{2\beta})_{ij}dx^i dx^j$, where β_{ij} is a symmetric, traceless, 3×3 matrix and both α, β_{ij} are functions of time, but not of spatial cordinates. The Kasner solution is a special case of Bianchi type I.

[27] This discussion is based on Misner and Matzner (1972). The calculations in Misner (1968) utilize an approximate form of the stress-energy tensor in order to calculate the effects of viscosity.

long, and anisotropic expansion will lead to significant differences in the temperature of neutrinos moving along different axes.[28] However, the scattering cross section is still large enough so that the probability of collisions with electrons is high. These collisions serve as a dissipative mechanism, converting anisotropy energy to thermal energy, and the increase in energy density damps and isotropizes the expansion.

Within a year of Misner's first paper, Stewart (1968) and Doroshkevich et al. (1968) argued that neutrino viscosity could only smooth out small initial anisotropies. A small literature sprang up concerning the effectiveness of dissipative mechanisms, leading to a consensus that neutrino viscosity could not damp all possible initial anisotropies. In addition to these criticisms of Misner's specific proposal, the "chaotic cosmology" approach faced three more general problems. First, Barrow and Matzner (1977) recognized that the observed finite photon to baryon ratio ($\approx 10^8$), which serves as a measure of entropy, gives an upper limit on the amount of dissipation which could have occurred in the early universe (via *any* dissipative mechanism). Neutrino viscosity dissipates anisotropy energy, but it also produces entropy in the process, and dissipation of large anisotropies would lead to a massive overproduction of entropy. Second, Collins and Stewart (1971) argued that Misner's methodological goal of completely eliminating dependence on initial conditions conflicts with standard existence and uniqueness theorems from the theory of ordinary differential equations. Collins and Stewart (1971) wrote down EFE for an anisotropic homogeneous model as an autonomous system of non-linear ordinary differential equations and showed that one can always pick an arbitrarily large anisotropy at a given time t_0 and find a solution of this system of equations as long as there are no processes which could

[28]The cosmological redshift of radiation in an expanding universe is proportional to a^{-4} (where a is the scale factor). Due to the difference in a along different axes in an anisotropic model, the radiation along a contracting axis is blue-shifted relative to the radiation along an expanding axis, and thus has an energy distribution corresponding to a higher temperature. See Misner and Matzner (1972).

prevent arbitrarily large anisotropies at some $t_i < t_0$.[29] Misner himself recognized a third fundamental problem, now known as the "horizon problem," which I will discuss below (along with his attempted solution).

2.2.2 Particle Creation

The research of Zel'dovich, Parker and others focused on a different mechanism which could lead to isotropization in the early universe: the creation of particles in the strong gravitational fields near the initial singularity. Zel'dovich's work was inspired by vacuum polarization effects and pair creation in QED, whereas Parker developed an account of spin-0 and spin-1/2 quantum fields evolving in a fixed background spacetime. The difference in the rate of particle creation in anisotropic and isotropic models suggested a possible explanation of isotropy—Zel'dovich and Novikov described the possible connection as follows:

> An important result is that such effects [particle creation and vacuum polarization] are strong for anisotropic singularities but virtually absent at the Friedmann singularity. Perhaps this difference has something to do with Nature's apparent preference for the Friedmann model? (Novikov and Zel'dovich 1973, p. 390)

Elsewhere Zel'dovich stated much stronger conclusions:

> We conjecture that quantum effects prohibit the most general solutions of the general relativity equations as candidates for the initial cosmological state. [...] [A]nisotropic expansion at the singularity leads to infinite quantum effects and to infinite particle creation. This is considered to prohibit anisotropic singularities. (Zel'dovich 1974, p. 331)

Below I will briefly review the formal results meant to support these conclusions, focusing on Parker's more clearly articulated approach. Subsequent research cast doubt on Zel'dovich's bold

[29] More precisely, they write EFE in the form $\dot{x} = f(x,t)$. On the assumption that f is continuous and satisfi es the Lipschitz condition, i.e. that for some constant c, $|f(x,t) - f(y,t)| \leq c|x-y|$ for all x, y in the set of solutions G, they show that EFE have a unique solution for any set of initial conditions in G.

conclusions, leading instead to an appreciation of the difficulties with defining number operators for quantum fields in an expanding cosmological model.

Parker's approach begins with the simplest generalization of the Klein-Gordon equation, following the usual "rule" of rewriting ordinary derivatives as covariant derivatives to move from special to general relativity:

$$(g^{ij}\nabla_i\nabla_j - m^2)\phi(x^i) = 0 \tag{2.4}$$

For a flat FLRW metric with a scale factor $a(t)$ this equation takes the form:[30]

$$\ddot{\phi} + 3\frac{\dot{a}(t)}{a(t)}\dot{\phi} - \frac{\nabla\phi}{a(t)^2} + m^2\phi = 0 \tag{2.5}$$

Parker then finds that the solution of eqn. (2.5) is a generalization of the special relativistic result. Switching from Minkowski space to an expanding, flat spacetime leads to replacing the usual term $\omega(k,t) = [k^2 + m^2]^{1/2}$ appearing in the expansion of the Heisenberg field operator ϕ in terms of creation and annihilation operators with a more general expression,[31]

$$W(k,t) = \left[\frac{k^2}{a(t)^2} + m^2\right]^{1/2} + \lambda(k,t), \tag{2.7}$$

[30]The flat ($k = 0$) FLRW metric is given by $ds^2 = -dt^2 + a(t)^2[dx^2 + dy^2 + dz^2]$. $\dot{\phi}$ denotes the derivative of ϕ with respect to t.

[31]The special-relativistic expression for the Heisenberg operator ϕ is

$$\phi(x^i) = \int \frac{d^3k}{(2\pi)^3(2\omega_{\mathbf{k}})^{1/2}}[exp(i\mathbf{k}\cdot\mathbf{x} - i\omega_k t)a_{\mathbf{k}} + exp(-i\mathbf{k}\cdot\mathbf{x} + i\omega_k t)a^\dagger_{\mathbf{k}}]. \tag{2.6}$$

In this expression $a_{\mathbf{k}}$ and $a^\dagger_{\mathbf{k}}$ are time independent, but in an expanding universe these operators vary with time.

with the stipulation that the last term vanishes when $a(t)$ is constant. Parker also imposes a so-called "minimization postulate" on $W(k,t)$, meant to insure that a well-defined particle number operator exists when $a(t)$ is nearly constant (see Parker 1969, pp. 1064-65). Based on this postulate, Parker shows that no spin-0 particles are created in a flat FLRW model in two cases: (1) the limiting case as mass goes to infinity (equivalent to modeling the particle content as pressureless dust), and (2) if the particle content consists only of massless particles in equilibrium.[32]

Parker glimpsed a "far-reaching consistency of nature" (Parker 1969, p. 1067) in the connection between these results and the isotropy of the early universe. Explicitly, he derived the standard equations for the evolution of $a(t)$ in a flat FLRW model based on the results above and the following "minimization" hypotheses:

> Hypothesis A: In an expansion of the universe in which a particular type of particle is predominant, the expansion achieved after a long time will be such as to minimize the average creation rate of that particle.
>
> Hypothesis B: The reaction of the particle creation (or annihilation) back on the gravitational field will modify the expansion in such a way as to reduce the creation rate (Parker 1969, p. 1066)

As Parker emphasizes, the derivation of the results (1) and (2) above do not depend directly on the field equations of general relativity, so an argument based on these Hypotheses and Parker's formalism gives an independent route to the flat FLRW model. Unfortunately, Parker has little to say about physical motivations for either Hypothesis A or B, or the earlier minimization postulate required to fully specify the form of $W(k,t)$. Although I have not found a criticism along these lines in the literature, I would be surprised if cosmologists had accepted Parker's argument as

[32] In deriving these results, Parker considers only the evolution of a spin-0 field in the background space (generalized to a spin-1/2 field in Parker (1970)), coupled to no other fields.

an explanation of isotropy without a strong physical motivation for (or plausibility arguments in favor of) these two Hypotheses.

Further work on particle creation revealed an important difficulty with Parker's analysis (see, in particular, Criss et al. 1975, pp. 93-93). Steigman recognized this difficulty in a question posed to Zel'dovich (Zel'dovich 1974, p. 333)—paraphrasing Steigman's question, to what extent are results of this approach dependent on the definition of particle number at some initial time t_i? The definition of the creation and annihilation operators (and the number operator, defined as $N_k = a_k^\dagger a_k$) in field theory on Minkowski spacetime depends upon the ability to separate positive and negative frequency solutions of the field equation. For stationary spacetimes (which possess a timelike Killing vector field), one can still divide positive and negative frequency solutions. Parker argues that for a cosmological model which approaches a nearly static model, one can neglect the slow time variation of the number operator and treat it like the well-defined number operator for the static case.[33] However, in order to determine the total number of particles created one also needs to determine the particle content of the initial state—which was anything but static. Parker's calculations rely on stipulating that the field of interest was in a vacuum state at an initial time t_i, prior to the onset of expansion. One can then calculate the number of particles N produced between t_i and a later time, when the expansion has slowed. Unfortunately, N depends on the choice of t_i. The difficulty with defining particle number in a general cosmological model near the singularity presents a serious obstacle to applying Parker's approach to early universe cosmology.

[33] A static spacetime is a stationary spacetime for which, in addition, there are spacelike hypersurfaces orthogonal to the timelike curves generated by the Killing fi eld.

Results published in the early 70s failed to support Zel'dovich's bold conclusion that particle creation effects would explain the isotropy of the observed universe. However, the results derived by Parker and others opened up new lines of inquiry. In 1975, research in particle creation effects shifted away from the early universe to black holes, as a result of Hawking's prediction of black hole radiation.

2.2.3 Baryogenesis

The theory of baryogenesis serves as a good example of how developments in quantum field theory dramatically altered the understanding of the early universe. Since the late 70s cosmologists have used baryogenesis as the paradigm case of a successful dynamical solution to a fine tuning problem (see, e.g., Olive 1990). Other than the trace amounts created in physics laboratories, our immediate environment is composed entirely of matter rather than antimatter. Observations in the 70s established that this asymmetry between matter and antimatter extends to much larger scales (see, in particular Steigman 1976). Baryons and anti-baryons annihilate at low temperatures, so any appreciable mixing of matter and antimatter would produce gamma ray emissions. Observational searches failed to detect gamma ray emissions characteristic of matter-antimatter annihilations. Cosmic rays also provide observational constraints on baryon asymmetry: although antimatter has been detected in cosmic rays, the rate is compatible with the assumption that the antimatter is the product of interactions between cosmic rays and the interstellar medium. The observations are consistent with a maximally baryon asymmetric universe—all baryons and no anti-baryons—but cosmologists were left wondering how this came about.

There appeared to be two ways of handling the baryon asymmetry in the mid to late 60s. First, one could skirt observational constraints and preserve baryon symmetry if the matter and antimatter "clumped" into large, non-interacting regions. Harrison and Omnes both pursued this line of thought (see Harrison 1968; Omnes 1971, and references therein); both held that in the early stages of the universe matter and antimatter would "clump" into regions which would eventually grow to super-galactic scales. The main difficulty with this approach was to rectify any clumping dynamics with the overall uniformity of the CMBR and successful nucleosynthesis calculations. This clumping would have to act in the very early universe to prevent the so-called "annihilation catastrophe": because of the large annihilation cross section, in a locally baryon symmetric universe only a very small number of baryons (or anti-baryons) would survive the universe's first moments.[34] Omnes in particular developed a sophisticated account of clumping, based on the idea that a phase transition in the early universe would produce distinct bubbles of matter and antimatter. He further hoped that his model would account for galaxy formation, and possibly even for the energy output of quasars, via the inclusion of contracting lumps of antimatter in the center of a matter-dominated galaxy. The alternative to this very speculative proposal was to assume that the universe began with a slight pre-dominance of matter. This surplus matter would survive the rapid annihilation reactions in the early universe. The amount of extra matter is related to the value of the numerical constant $\eta = \frac{n_b}{n_\gamma}$ mentioned above. From nucleosynthesis constraints, it follows that in a baryon symmetric universe the initial abundances

[34] See, for example, Kolb and Turner (1990) pp. 119-128 for a detailed discussion. They show that without some way of separating baryons and anti-baryons, the abundances would "freeze out" at $\frac{n_b}{s} \approx 7 \times 10^{-20}$, where n_b is the number of baryons and s is the local entropy density, compared to an observed value of $\frac{n_b}{s} \approx 6 - 10 \times 10^{-11}$.

of matter and antimatter would have differed only by the order of 10^{-9}. But how could this imbalance take such a precise, apparently finely tuned value?

A third, much more appealing way of handling baryon asymmetry was provided by grand unified theories (GUTs) developed in the 70s. The arguments above assumed that baryon number is conserved in all interactions, but GUTs incorporate baryon (and lepton) number non-conserving interactions. These interactions open up the possibility of generating the observed baryon asymmetry from symmetric initial conditions. In a prescient paper that preceded the advent of GUTs by several years, Sakharov argued that baryon asymmetry could be explained by a model having the following three features (Sakharov 1967):

- Baryon number non-conservation
- C (charge conjugation) and CP (charge conjugation plus parity) violation[35]
- Departure from thermal equilibrium

Sakharov's model was based on using heavy bosons mediating quark-lepton interactions, but GUTs proposed in the following decade shared these features. The second condition is needed in order for the baryon-number violating interactions to produce an excess of baryons rather than anti-baryons. Since the masses of particles and their antiparticles are identical, the final condition is needed to insure that the inverse reactions do not immediately erase the excess baryon number.

Within a specific GUT one can calculate the value of η based on the dynamics much like one can calculate element abundances in nucleosynthesis based on the underlying nuclear physics. The exciting prospect of developing an account of baryogenesis was the focus of a

[35] See, for example, Sachs (1987), Chapters 8 and 9, for a clear discusion of C, P, T (time reversal) invariance in quantum fi eld theory.

small industry in the late 70s: Dimopoulos, Susskind, Weinberg, Wilczek, and several others all contributed to this effort. This was the first application of GUT-scale particle physics, in all of its gory detail, to the early universe. In the next chapter we will turn to these theories and their other consequences for early universe cosmology.

2.3 The Horizon Problem

The chaotic cosmology program led to a recognition of the general problems faced by any attempt to eliminate the various types of fine-tuning discussed above. The most important of these hinges on a somewhat counter-intuitive feature of the FLRW models first recognized clearly by Rindler (1956).[36] Length scales shrink to zero as the initial singularity is approached, and one might expect that this shrinking would allow regions of the universe which are currently far apart to come into causal contact in the early universe. Contrary to this expectation, it turns out that distant regions could not have causally interacted in the early universe, if it is similar to an FLRW model (see figure 2.1, and Appendix A.3 for further discussion). The presence of particle horizons signals this failure:

> Particle horizons in cosmological models are limits on the possibilities of causal interactions between different parts of the universe in the time available since the initial singularity. (Misner 1969b, p. 1071)

The horizon problem arises due to the apparent conflict between the presence of particle horizons in standard FLRW models and the observed uniformity of the CMBR.

[36] Although Rindler introduced particle horizons and noted their presence in the FLRW models, Misner was the first to clearly state the "horizon problem" described below. Misner mentions particle horizons in his first paper on dissipative mechanisms, Misner (1967), p. 40: "The wavelengths affected by this mechanism are those between a γ mean free path $\approx ct_{e\gamma}$ and the size of the particle horizon $\approx ct$ (this is the maximum distance across which causal signals have traveled since the initial singularity)."

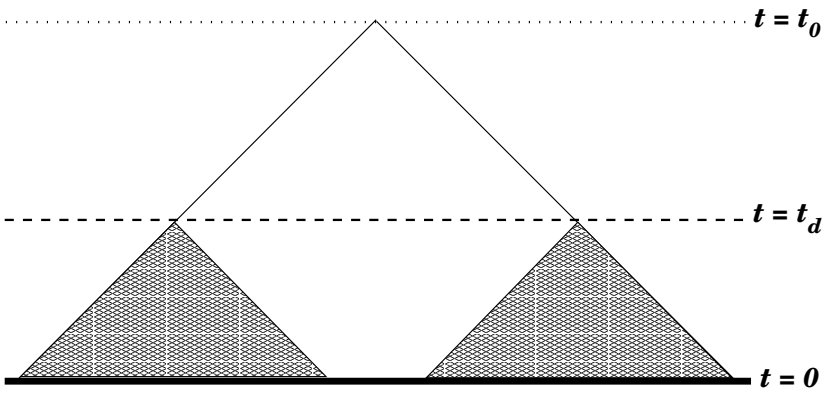

Fig. 2.1 This conformal diagram illustrates the horizon problem in the FLRW models. In a conformal diagram distance scales are distorted in order to accurately represent causal structure; light cones are all at 45°. The singularity at $t = 0$ is stretched to a line. The lack of overlap in the past light 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.

> In observational terms this effect says, for example, that if the $3^{\circ}K$ background radiation were last scattered at a redshift $z = 7$, then the radiation coming to us from two directions in the sky separated by more than about 30° was last scattered by regions of plasma whose prior histories had no causal relationship. These Robertson-Walker models therefore give no insight into why the observed microwave radiation from widely different angles in the sky has very precisely ($\leq 0.2\%$) the same temperature. (Misner 1969b, p. 1071)

Misner argued that unlike the standard FLRW models, anisotropic models may solve this horizon problem due to a complicated series of horizon-breaking oscillations in the early universe. This horizon breaking would provide world enough and spacetime for distant points on the surface of last scattering to causally interact.

Misner's own attempt to solve the horizon problem drew on the study of the evolution of vacuum solutions near the initial singularity. I will discuss this proposal in some detail since it illustrates the importance of singularities in Misner's approach. The "mixmaster universe," as Misner called his model, referring to a popular food processor, evolves through a series of stages in which it resembles a Kasner solution, equation (2.2). If we assume that the only non-zero expansion is along the x axis, eqn. (2.2) simplifies to:

$$ds^2 = -dt^2 + t^2 dx^2 + dy^2 + dz^2 \tag{2.8}$$

The two-dimensional metric corresponding to the t, x coordinates can be rewritten in terms of a new time coordinate $\eta \equiv \ln t$:

$$ds^2 = e^{2\eta}(-d\eta^2 + dx^2) \tag{2.9}$$

The homogeneous surfaces of the mixmaster universe are closed, and in order to model them in terms of the Kasner solutions we artificially close the space by identifying points on the spatial axes, say x and $x+4\pi n$. As a result, for any coordinate time interval $\Delta\eta > 4\pi$ along a worldline, the horizon distance will be greater than the length of the x axis—so that there are no particle horizons along the x axis. Misner argued that a mixmaster model would pass through an infinite number of stages similar to the Kasner solution as the singularity is approached. If each of these stages persists for a sufficient period of time, then the horizons along every axis disappear. Critics of Misner's proposal argued that the probability of the successive stages required to completely rid the mixmaster universe of horizons is quite low.[37] The techniques Misner developed to

[37] Criticisms of the mixmaster universe along these lines are voiced in MacCallum (1971), MacCallum (1979), and Collins and Stewart (1971); see also Zel'dovich and Novikov (1983), §22.4 for references to the extensive Russian literature.

study the mixing behavior of the model were used for more general analysis of the evolution of anisotropic spacetimes near an initial singularity. Misner reformulated the mixmaster dynamics in a canonical Hamiltonian form, using the ADM method. The Hamiltonian turns out to be quite simple for most of the Bianchi models—it is the same as that for a particle moving in a five-dimensional space with an effective potential. Active research in "Hamiltonian cosmology" seems to have peaked in the mid-70s, although the approach spilled over into studies of quantum cosmology and has been recently revived in studies of chaotic behavior in GTR.[38]

Misner's attempt to answer these criticisms of the mixmaster model was linked to an interesting position regarding the role of the initial singularity:

> I prefer a more optimistic viewpoint ("Nature and Einstein are subtle but tolerant") which views the initial singularity in cosmological theory not as proof of our ignorance, but as a source from which we can derive much valuable understanding of cosmology. [...] The concept of a true initial singularity (as distinct from an indescribable early era at extravagant but finite high densities and temperatures) can be a positive and useful element in cosmological theory. (Misner 1969a, p. 1329)

In particular, he speculated that taking the singularity seriously and revising the concept of time would lead to an effective mixmaster model.[39] Attentive readers may have already noticed a curious feature of the time coordinate $\eta = \ln t$ introduced in the discussion of the Kasner solution above: it diverges to $-\infty$ as $t \to 0$. Misner argues on physical grounds that $\Omega = \ln(T)$, where T is the temperature, is an appropriate time coordinate for studying the singularity. It shares the

[38] See Criss et al. (1975) and Ryan and Shepley (1975) for comprehensive reviews. Several of Misner's dissertation students actively pursued research along these lines, including Matzner, Ryan, Chitre, and Shepley (who worked closely with Misner, but did not complete a dissertation under him). See Hu et al. (1993) for a list of Misner's dissertation students, and Misner (1994) for the connection with studies of chaotic behavior.

[39] "But if the computations accept the singularity and therefore contemplate infinitely many decades of expansion, we have here a promising approach to an understanding of large-scale homogeneity." Misner (1969a), p. 1330. These ideas are also discussed in Box 30.1 of Misner et al. (1973).

divergent properties of η, blowing up to ∞ as the singularity is approached.[40] Misner comments that: "*The Universe is meaningfully infinitely old because infinitely many things have happened since the beginning*" (Misner (1969a), p. 1331, his emphasis). This saves the mixmaster by allowing ample time for the oscillations to eliminate particle horizons in all directions. This redefinition of the concept of time is coupled with an argument that quantum effects will not alter the dynamics as the initial singularity is approached (see Misner 1969c). Though this attempt to save the mixmaster was ultimately unsuccessful, it does reveal the importance Misner accorded to the proper understanding of singularities in GTR.

The horizon problem has long outlasted Misner's proposed solution of it. The presence of particle horizons presents a general problem for the chaotic cosmology program: if the horizons accurately measure the length scales over which dynamical mechanisms are effective, then dynamics cannot produce the uniform state of the early universe from arbitrary initial conditions. Misner's suggestion that an understanding of the initial singularity and the evolution of anisotropic cosmologies would alleviate this problem stimulated a significant research program. But by 1975 Criss et al. (1975) concluded a lengthy review article with the comment that: "The explosion of activity in theories of cosmology due to the new data that became available during the 1960s has to an extent run its course." Prospects for the chaotic cosmology program were dim, according to the reviewers: there was still no convincing demonstration that the dynamical evolution of anisotropic cosmologies could explain the observed isotropy of the universe. They

[40] The usual measure of local time in relativity theory is the proper time elapsed along a timelike curve, which can be interpreted as the time recorded by an ideal clock moving along the curve. Although there are some subtleties involved in treating the "time" of the initial singularity, in an FLRW model the singularity is a finite proper time to the past of all observers. The "infinity" Misner utilizes is an artifact of this particular choice of coordinates.

call horizon mixing “the ghost of an idea that failed” (p. 73). Without some modification of the causal structure of the early universe, chaotic cosmology could not solve the horizon problem.

Although responses to the horizon problem among cosmologists in the late 60s and 70s varied, based on the extensive interviews collected in Lightman and Brawer (1990), Brawer (1996) concludes that it struck a chord with most cosmologists. Misner’s close personal and institutional ties to most if not all of the major gravity groups insured that his ideas would be well disseminated.[41] Not surprisingly, the most detailed discussions of the horizon problem appeared in the literature on chaotic cosmology, but the problem was also mentioned in standard graduate textbooks on relativity. Readers of “Track 2” of the gargantuan Misner et al. (1973) would discover a discussion of the horizon problem, and chaotic cosmology more generally, in Chapter 30. Weinberg (1972) briefly mentions the problem (p. 526):

> [I]t is difficult to understand how such a high degree of isotropy could be produced by any physical process occurring at any time since the initial singularity.

The passing notice of the problem given in Weinberg’s text is fairly typical. Misner’s chaotic cosmology program was also frequently mentioned in cosmology texts (see, e.g. Peebles 1971, pp. 222-23, and Sciama 1971, pp. 200-201), but again the horizon problem did not generally receive a detailed treatment. In the extensive interviews recorded in Lightman and Brawer (1990), most of the interviewees recall first encountering the horizon problem in the 70s as one of the fundamental unsolved problems in cosmology. Those who took the problem seriously often agree with Misner’s distrust of explanations of uniformity based on specialized initial conditions.

[41] Misner maintained close ties with Wheeler at Princeton and the Caltech gravity group via Kip Thorne. In addition, the relativity group at the University of Texas included several of his former students, and the Cambridge group was well aware of the research Misner conducted there during his sabbatical year.

The response to the horizon problem contrasts starkly to the reception given to the flatness problem, first described by Dicke in a popular lecture delivered in 1969:

> Another matter [the first is the horizon problem] is equally puzzling. The constant v_0^2 in[42]
>
> $$v^2 = \frac{8\pi G\rho r^2}{3} - v_0^2 \tag{2.10}$$
>
> is very small, so small that we are uncertain with our poor knowledge of ρ as to whether or not it is zero. But the first term on the right was very much larger much earlier, at least 10^3 times as great when the galaxies first started to form and at least 10^{13} times as great when nuclear reactions were taking place in the "fireball"... The puzzle here is the following: how did the initial explosion become started with such precision, the outward radial motion become so finely adjusted as to enable the various parts of the universe to fly apart while continuously slowing in the rate of expansion? (Dicke 1969, pp. 61-62)

Guth (1981) dubbed this problem the "flatness problem" and gave it as much emphasis as the horizon problem, but prior to his paper it is only mentioned twice in the published literature (as far as I know): Hawking (1974), and Dicke and Peebles (1979). In addition, cosmologists often did not see this as a serious problem prior to Guth (1981)'s clear statement and attempted solution, which linked the flatness problem to the horizon problem. Brawer (1996) argues that part of the reason for the divergence of opinion regarding the horizon and flatness problems is the connection of the former with causality, a topic I will discuss at length in Chapter 7.

To summarize this section, the horizon problem represents one obstacle to the general approach of the chaotic cosmology program, which aims to eliminate fine-tuning by the introduction of new dynamics. But I should also briefly mention two other very different approaches to the fine-tuning problems of big bang cosmology (they will be discussed more fully in the following chapters). An approach called "quiescient cosmology," as described in Barrow (1978),

[42] This is just the Friedmann equation rewritten, with a replaced by r and $\dot{a}$ by v.

takes evidence of regularities in the early universe as evidence of regularities specified in the initial conditions, rather than as evidence for physical mechanisms which must have eliminated primordial irregularities. More generally, Penrose (1979) and others have argued that an appropriate constraint on initial conditions should emerge from an application of quantum gravity to the initial singularity. These proposals have remained speculative given the outstanding conceptual and technical difficulties facing quantum gravity and quantum cosmology, and by the late 70s (and even up to the present) there were no convincing "theories of the initial conditions" based on quantum gravity. Mainstream cosmologists have generally preferred dynamical solutions to the fine-tuning problems, perhaps because, as Misner has pointed out, "physics has no experience with" theories of initial conditions (Misner 2001). A second alternative has provoked a great deal of debate among cosmologists: anthropic reasoning may also provide a way to avoid the alleged improbability of the initial state. The weakest form of anthropic reasoning asserts simply that the presence of human observers must be taken into account in assessing cosmological theories, in the same way that (for example) the presence of stable planetary systems must be taken into account. An anthropic explanation of the apparently "finely-tuned" features of the observed universe would amount to showing that they are necessary conditions for the existence of intelligent life, and so (the argument goes) we should not be surprised to observe these features of the universe. If these conditions did not obtain, there would be no intelligent life here to wonder about the probability of the initial state. Whether anthropic reasoning of this (or stronger) forms produces satisfactory explanations has been a matter of considerable dispute, which I will discuss in greater detail in Chapter 5.

2.4 Conclusions

This chapter has traced the birth and development of the field of early universe cosmology through its first decade. The remarkable observation of primordial radiation, interpreted as the remnant of the big bang, forced researchers to take the early universe seriously. This field was a fertile intersection of interesting theoretical problems in general relativity and particle physics, and the possibility of further precision observations of the CMBR also provided some hope of putting various proposed solutions to the test. By the early 70s, cosmology had a well-articulated "standard model" accepted by a majority of mainstream cosmologists. The development of the standard model of cosmology coincided with renewed interest in general relativity, and many of the recently founded relativity groups devoted a great deal of research effort to cosmology, including early universe cosmology. In addition, the standard model incorporated a much more detailed account of the thermal history of the universe, similar in some ways to Gamow's earlier theory.

But the standard model was not without its blemishes: in particular, it required a number of apparently implausible assumptions regarding the universe's initial state. As I argued above, the most widely accepted methodology for handling these "fine-tuning problems" was to eliminate the need for specialized initial conditions by introducing new dynamics. Misner's suggested mechanism for isotropization and the mixmaster model failed to "predict isotropy" as Misner had hoped, as did the work on particle creation by Zel'dovich, Parker and others. Of the three examples of "new dynamics" discussed above, only baryogenesis was still a focus of active research by the end of the 70s.

At the end of the first decade, then, early universe cosmology was a rapidly growing field, with a number of theoretical and observational research programs devoted to its study. Although the early burst of research activity had not led to a clear resolution of the fine-tuning problems, these problems had been recognized and studied in detail, and mainstream cosmologists appear to have agreed on the proper method for solving them. Prominent particle physicists such as Weinberg contributed to the field, and also recognized that the early universe could provide an important testing ground for new ideas in high energy physics. But the real impact of particle physicists on early universe cosmology did not come until the late 70s and early 80s, a topic we will turn to in the next chapter.

Chapter 3

False Vacuum

Accounts of the development of inflationary cosmology typically present their protagonists, a small band of American particle physicists with Alan Guth in the lead, as intrepid explorers venturing into untouched territory. Driven by the need to find observational tests of Grand Unified Theories (GUTs), they were willing to boldly apply their ideas to the distant realm of the early universe. Beginning in the early 80s, they extended the frontiers of physics ever closer to the absolute zero of time and created a vibrant new field.

The last chapter undercuts this bit of conventional wisdom by showing that the early universe had been the focus of active research for over a decade before the particle physicists arrived on the scene. Early attempts at solving the fine-tuning problems of big bang cosmology relied primarily on introducing new physics from two sources: using more realistic descriptions of matter, and including gravitational effects near the initial singularity (such as mixmaster oscillations and particle creation). But the conventional wisdom is also misleading regarding the interaction between cosmology and particle physics and the provenance of one of the crucial new ideas in cosmology. Interplay between research in particle physics and cosmology began in the early 70s with the advent of the Glashow-Weinberg-Salam electroweak theory and its speculative extensions, GUTs. Soviet physicists, including Linde, Kirzhnits, Sakharov, and Zel'dovich, among others, led the way in studying the implications of these new ideas for cosmology, whereas a somewhat more rigid disciplinary divide persisted throughout the 70s in America and England.

A number of scientists proposed that the very early universe passed through a de Sitter-like phase, the characteristic feature of inflationary cosmology, before Guth's seminal paper. Finally, we will see below that the fine-tuning problems bothering Misner and others were not the sole driving force in the development of these ideas.

The discussion below focuses on the different proposals that the early universe passed through an early de Sitter-like phase, with the aim of clarifying both the motivations for developing such theories and the theoretical tools used in these accounts. Roughly speaking, there were three different reasons to introduce such a radical alteration of the early universe. First, few cosmologists of the 60s and 70s shared Misner's tolerance for spacetime singularities. Instead, their abhorrence of the initial singularity was strong enough to motivate a speculative modification of the FLRW expansion. Several authors realized independently that one could evade the Hawking-Penrose singularity theorems if the early universe somehow began in a vacuum state. As we will see below, such a state would violate one of the crucial assumptions of the theorems, a requirement that matter-energy density has a focusing effect. The second rationale bears some similarity to the reasoning behind Lemaître's "primeval atom" hypothesis, with its focus on a quantum mechanical account of the initial creation event. A number of physicists were hopeful that new ideas in fundamental physics would provide the proper theoretical framework for an account of "creation." For example, 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. Third, an early de Sitter stage emerged as the consequence of developments in two different areas in physics. Starobinsky found that the de Sitter solution is an unstable solution to the field equations in the semi-classical

approach, when one-loop quantum corrections to $\langle T_{ab} \rangle$ are included in the source term, and constructed a cosmological model based on this insight. The application of GUTs to early universe cosmology generated a great deal of interest and introduced a number of novel possibilities for early universe cosmology. In particular, several physicists independently discovered that the Higgs field (a component of the GUTs) trapped in a false vacuum state would drive a transient de Sitter-like phase.

These proposals shared two common problems. First, what was the source of an early vacuum-like state postulated to dominate the expansion in the early universe? 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? As we will see below, these problems were approached with a wide variety of theoretical tools and differing degrees of success.

This chapter proceeds as follows. Section 1 focuses on two approaches motivated by the desire to eliminate the singularity, the phenomenological approach of Gliner and Dymnikova and Starobinsky's model. The next section turns to the history of quantum field theory, for a brief discussion of symmetry breaking and the development of the Standard Model. This leads up to a discussion of the use of symmetry breaking ideas in various approaches to early universe cosmology. Section 3 focuses on research regarding early universe 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.

3.1 Eliminating the Singularity

3.1.1 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.[1] Gliner (1966) and others noted that 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.[2] The stress-energy tensor for a perfect fluid is given by

$$T_{ab} = (\rho + p)u_a u_b + p g_{ab}, \tag{3.1}$$

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

[2] Gliner takes the following requirement to be the defining property of a relativistic vacuum (what he calls the "μ - vacuum"): that interactions between ordinary matter and the vacuum cannot depend on velocity, since the co-moving frame for any particle of ordinary matter will be at rest with respect to the vacuum. Although Gliner is only concerned with local Poincar´e invariance, he does not recognize the difficulties in extending Poincar´e invariance to general relativity. As a result, in general the "vacuum" cannot be uniquely specified by requiring that it is a Poincar´e invariant state. I thank John Earman for emphasizing this point to me (cf. Earman 2002, 208-209).

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.[3] 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.[4] 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

[3]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.

[4]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 fi ner points into account.

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.[5] 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.[6]

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

[5] 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.

[6] Gliner was not alone in this preference; several other papers in the early 70s discussed violations of the strong energy condition as a way of avoiding the singularity, as we will see in the next section.

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.[7] 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 Gliner and Dymnikova (1975)'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.[8] Rather than postulating a sudden transition, Gliner and Dymnikova (1975) stipulate the following *ansatz* for the equation

[7] 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.

[8] 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.

of state during a gradual transition:[9]

$$\rho + p = \gamma\rho_1 \frac{(\rho_0 - \rho)^\alpha}{(\rho_0 - \rho_1)^\alpha}, \tag{3.2}$$

based on the idea that the vacuum matter decays as the de Sitter bubble grows. The parameter α is assumed to satisfy $0 < \alpha < 1$, and γ is the term appearing in the equation of state for a perfect fluid, $p = (\gamma - 1)\rho$. For $\rho = \rho_0$, the equation of state is just $p = -\rho$, but the transition (with the rate set by α) leads to $\rho = \rho_1$ and the equation of state for normal matter. The scale factor and the mass of the universe both grow by an incredible factor during this transitional phase, as Gliner and Dymnikova (1975) duly note; 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. There is no indication of a link with other areas of physics that might provide a more specific, well-motivated model. Instead, details of the transition are set by matching the entropy generated during the transition with 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

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

radiation density (ρ_p, ρ_{rp}):[10]

$$\sqrt{\frac{\rho_1}{\rho_{rp}}} exp\left(\frac{2(\rho_0 - \rho_1)}{3\gamma\rho_1(1-\alpha)}\right) = \frac{\rho_0}{\rho_p}\left(1 - \frac{3H^2}{8\pi G\rho_p}\right)^{-1}. \tag{3.3}$$

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.

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 30s by the likes of Sir Arthur Eddington and Willem de Sitter.[11] 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, p. 69), but the goal itself was clearly articulated.

[10] 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 Hubble's constant.

[11] Eddington (1933, 37) and de 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.

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. Recall that horizons in relativistic cosmology mark off the region of spacetime from which light signals could have reached an observer (see Appendix A.3 for a brief reminder). As we saw above, Misner (1969b) had 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. By the mid 70s a number of Gurevich's comrades (along with British cosmologists and Misner himself) had put the idea to rest. But mixmaster oscillations were unnecessary to solve the horizon problem; as Sakharov recognized, an odd equation of state would suffice:[12]

> 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
>
> $$\int_{t_0}^{t_1} \frac{dt}{a} \to \infty \quad \text{as} \quad t_0 \to 0, \tag{3.4}$$
>
> 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, although Novikov has commented that Sakharov's "wild ideas" seemed "utterly incomprehensible" at the time (Altshuler

[12] 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.

et al. 1991, p. 474). It appeared buried in the Appendix of a preprint that was only widely available following the publication of the *Collected Works* in 1982.

3.1.2 Starobinsky's Model

During a research year in Cambridge in 1978-79, Zel'dovich's protégé Alexei Starobinsky developed an account of the early universe based on including quantum corrections to the stress-energy tensor in EFE. Starobinsky was one of the main players in the development of semi-classical quantum gravity; his early work with Zel'dovich focused on particle creation in strong gravitational fields, and he had a part in discovering the Hawking effect. He clearly shared Gliner and Dymnikova's desire to avoid the initial singularity, and his model also includes an early de Sitter phase. But there the similarity with Gliner and Dymnikova's work ends.[13] Unlike Gliner and Dymnikova's sterile phenomenological approach, Starobinsky's model drew on a rich source of ideas: recent results in semi-classical quantum gravity.

Throughout the 70s 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 Hawking's alleged "no go theorem," Zel'dovich and Pitaevsky

[13] Starobinsky (1979) explicitly distances himself from Gliner: the effective equation of state in his model, while the same as in Gliner's models, follows from quantum gravity effects rather than a bald stipulation.

(1971) showed that during particle creation the effective T_{ab} violates the dominant energy condition.[14] 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 (1973)'s 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).[15]

Starobinsky (1979, 1980)'s 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 70s. 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."[16]

[14] Hawking (1970)'s 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.

[15] Bekenstein (1975) also discussed the possibility that scalar fields would allow one to avoid the singularity. Starobinsky (1978)'s 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 0.5, whereas Starobinsky finds 10^{-43}!

[16] 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.

Taking this anomaly into account, Starobinsky derived an analog of the Friedman equations and found a set of solutions to these equations.[17] 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 studying this model, Starobinsky mentions an observational constraint that simplifies the calculations considerably (Starobinsky 1980, p. 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.

In a conference paper delivered in 1981 at the Moscow Seminar on Quantum Gravity (published three years later as Starobinsky 1984), Starobinsky repeated a portion of this earlier paper with a page of new material added. This added material explains in greater detail that an extended de Sitter phase drives Ω *very* close to 1. But Starobinsky still does not present this aspect of the model as a major selling point: he comments that an extended de Sitter phase is necessary simply to insure compatibility with observations, and he does not further comment on whether an extended de Sitter phase is *natural* in the context of his model. His more detailed discussion was

---

Birrell and Davies 1982). Starobinsky was the first to consider the implications of these results for early universe cosmology.

[17] 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.

clearly motivated by Guth (1981) (which he cites), but his methodology still differs starkly from Guth's. Starobinsky's approach requires *choosing* the de Sitter solution, with no aim of showing that it is a "natural" state; as Starobinsky puts it, "This scenario of the beginning of the Universe is the extreme opposite of Misner's initial 'chaos.'"[18] 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 Guth suggests. 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, for example, Grib et al. 1984; Lapchinksy et al. 1984, and references therein).

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, p. 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 underlie later

[18] Indeed, Starobinsky acknowledges that the instability of the de Sitter solution indicates that it is *not* generic. Starobinsky (1979) describes his approach as postulating that de Sitter space is a solution of the full equations of quantum gravity (p. 684).

attempts (discussed further in Chapter 4) 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, which continued to $t \to -\infty$. But Zel'dovich noted two other important advantages of Starobinsky's model. First, it would solve the horizon problem:[19]

> 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. (Zel'dovich 1981, p. 229)

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

[19] 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 3.3 below.

danger of unemployment for theoreticians occupied with astronomical problems" (Zel'dovich 1981, p. 229).

3.2 Hidden Symmetry

The understanding of symmetries in QFT changed dramatically in the 60s due to the realization that field theories may exhibit spontaneous symmetry breaking (SSB). A typical one-line characterization of SSB is that the vacuum state of a broken symmetry theory does not share the full symmetries of the fundamental Lagrangian.[20] Symmetry breaking in this loose sense is all too familiar in physics: solutions to a set of differential equations typically do not 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. The latter notion is familiar from non-relativistic quantum mechanics: an exact symmetry S on the Hilbert space $\mathcal{H}$ preserves transition probabilities, i.e. for rays Φ, Φ', Ψ, Ψ', S maps rays onto rays such that $|\langle\Phi, \Psi\rangle|^2 = |\langle\Phi', \Psi'\rangle|^2$. A famous theorem due to Wigner demonstrates that such symmetries can be implemented by unitary transformations on $\mathcal{H}$.[21] Now consider a Lagrangian $\mathcal{L}$ invariant under a continuous global or internal symmetry. Roughly, systems for which a particular symmetry of the Lagrangian *cannot* be unitarily implemented on $\mathcal{H}$ exhibit SSB (see (B.1) for a more careful discussion). Typical characterizations of SSB focus on the consequences of this

[20] Coleman (1985), for example, characterizes SSB as the conjecture that "the laws of nature may possess symmetries which are not manifest to us because the vacuum state is not invariant under them" (p. 116).

[21] More precisely, Wigner showed that any S can be implemented by either a unitary and linear or an anti-unitary and anti-linear operator (see Weinberg 1995, Chapter 2 for a proof).

failure: observables acquire non-invariant vacuum expectation values, and the vacuum is degenerate. The idea of symmetry breaking was originally developed in relation to condensed matter systems with these features and only later imported into field theory by Nambu and others.

The study of SSB in field theory 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. SSB garnered a great deal of attention in the early 60s—at least one prominent theorist "fell in love with this idea" (Weinberg 1980, p. 515), and these research efforts (along with Weinberg's "love affair") eventually led to the idea that SSB could be used to "fix" Yang-Mills style gauge theory by indirectly giving mass to the gauge bosons. Weinberg and Salam independently proposed unified gauge theories of the weak and electromagnetic interactions incorporating SSB. On the heels of the success of these theories, physicists were willing to take SSB as something more than formal manipulations of the Lagrangian; instead, they began to speculate about the possible implications of symmetry breaking understood as a dynamical process in the early universe.

3.2.1 Spontaneous Symmetry Breaking and the Higgs Mechanism

Although the idea of broken symmetry was introduced long before the 60s, its fundamental importance for particle physics was recognized only in the late 50s and early 60s, as Nambu

and others studied SSB in field theory based on a fruitful analogy with the Bardeen-Cooper-Schrieffer (BCS) theory of superconductivity.[22] The seminal BCS paper (Bardeen et al. 1957) made no reference to symmetry breaking, focusing instead on a detailed dynamical model of superconductivity, but critics of the theory noted that calculations worked only with a particular choice of gauge (Buckingham 1957; Schafroth 1958). Following the original BCS paper, several physicists (Nambu and Anderson, in particular) explored the connection between the appealing features of the BCS theory—such as the prediction of a gap in the energy spectrum of a superconductor and the explanation of the Meissner effect[23]—and the theory's lack of gauge invariance. Nambu (1961) and Anderson (1958) both argued that taking massless longitudinal collective modes into account would insure gauge invariance, and in addition that Coulomb interactions would cause these collective modes to acquire mass. Motivated by the similarity between these features of superconductivity and earlier models of hadrons, Nambu and his collaborator Jona-Lasinio (Nambu and Jona-Lasinio 1961a,b) developed a model of pions incorporating symmetry breaking and suggested that SSB may play a fundamental role in particle physics.

A general result due to Goldstone seemed to doom symmetry breaking in particle physics barely after its inception: Goldstone conjectured and later proved (with Salam and Weinberg, Goldstone 1961; Goldstone et al. 1962) that SSB of a continuous symmetry implies the existence

[22] See Brown and Cao (1991) for a more detailed early history of SSB and discussion of its integration into QFT, as well as the first-hand accounts of Nambu, Higgs, and others collected in Chapter 28 of Hoddeson et al. (1997). The following brief account relies rather heavily on these sources.

[23] This gap is due to the correlations existing between "Cooper pairs," pairs of electrons created via interactions with the background crystal lattice in the BCS ground state. Very roughly, the distortion of the lattice caused by one electron's motion and its Coulumb attraction on the positive ions composing the lattice may persist long enough to affect the motion of a second electron (i.e., phonon exchange between the electrons results in an attraction); somewhat counterintuitively, even a very small attraction mediated by the lattice can produce pairing if the two electrons are immersed in a dense fluid of electrons. Heat capacity and other thermodynamic properties of superconductors reflect the existence of this energy gap, and the accurate prediction of the energy gap was a major success of the BCS theory. The Meissner effect refers to the expulsion of magnetic fields from the interior of a superconductor. Some more recent treatments, such as Weinberg (1996), §21.6, show that this is a direct consequence of the breaking of electromagnetic gauge invariance in a superconductor.

of a spin-zero massless boson. Goldstone et al. (1962) introduce the "effective potential" $V(\phi)$ in one of the three proofs they give of Goldstone's theorem, which I will briefly review here. $V(\phi)$ includes all terms in the Lagrangian other than the kinetic terms (terms of the form $(\partial_\mu\phi)^2$); for the simple ϕ^4 theory,

$$\mathcal{L} = \frac{1}{2}(\partial_\mu\phi)^2 - V(\phi) \quad \text{where} \quad V(\phi) = \frac{1}{2}m^2\phi^2 + \frac{\lambda}{4!}\phi^4. \tag{3.5}$$

The effective potential is identified with the effective energy density of the quantum fields, as it is in classical field theory (see Coleman 1985, pp. 138-142, and references therein). In standard ϕ^4 theory with $V(\phi)$ defined as above, the classical minimum of the potential lies at $\phi_0 = 0$.[24] However, changing the sign of the m^2 term leads to a double minimum of the effective potential at $\phi_0' = \pm\sqrt{\frac{6}{\lambda}}m$. This feature of the potential indicates that $\phi_0 = 0$ is no longer the true ground state of the field, since it has a higher energy than either of the ϕ_0' states.[25] Thus, in order to define our field operators in terms of creation and annihilation operators with respect to the true ground state, we need to "subtract off" the vacuum expectation value of ϕ_0'. To do this we introduce a shifted field variable defined by $\phi(x) = \phi_0' + \bar{\phi}(x)$. In terms of $\bar{\phi}(x)$, the Lagrangian has the following form (dropping an overall constant):

$$\mathcal{L} = \frac{1}{2}(\partial_\mu\bar{\phi})^2 - \frac{1}{2}(2m^2)\bar{\phi}^2 - \sqrt{\frac{\lambda}{6}}m\bar{\phi}^3 - \frac{\lambda}{4!}\bar{\phi}^4. \tag{3.6}$$

[24] This calculation is "classical" in that I am neglecting any higher order quantum corrections to $V(\phi)$ and simply treating ϕ as a real scalar field. These results carry over to QFT as the "tree approximation" (i.e., the approximation neglecting all Feynman diagrams with closed loops).

[25] One can also directly calculate the expectation value of the Hamiltonian for some suitably chosen states (such as coherent states); $\langle f|H|f\rangle$ as a function of the coherent state $|f\rangle$ displays the same "double minima" structure as $V(\phi)$.

The new $\bar{\phi}^3$ interaction term hides the original symmetry transformation $\phi \rightarrow -\phi$, and in addition the sign of the m^2 term is negative. The new Lagrangian appears to describe a simple scalar field with a mass $\sqrt{2}m$ and two distinct interactions; the relationship between the coupling constants of these interactions and the mass of the field provides the only hint of the hidden symmetry.

To obtain an example of Goldstone's theorem we need to generalize the *discrete* symmetry of this case to a *continuous* symmetry. Consider a theory with three scalar fields with identical masses, with the following Lagrangian:

$$\mathcal{L} = \frac{1}{2}(\partial_\mu \phi^i)^2 - V(\phi^i) \quad \text{where} \quad V(\phi^i) = -\frac{1}{2}m^2(\phi^i)^2 + \frac{\lambda}{4!}[(\phi^i)^2]^2 \tag{3.7}$$

(where $i = 1, 2, 3$—the different fields have been treated as components of a single field vector—with summation over i). The Lagrangian is invariant under the action of the three-dimensional rotation group $O(3)$ on field space. As in the case above, the minimum energy configuration corresponds to non-zero vacuum expection values of the field ϕ^i; in particular, the constant field $(\phi_0^i)^2 = \frac{6m^2}{\lambda}$ minimizes the effective potential. This condition determines only the *length* of the field vector ϕ_0^i and not its direction, so we can arbitrarily choose the direction of ϕ_0^i to coincide with the $i = 3$ field. Shifting the field variables as above,

$$\phi^i(x) = \sqrt{\frac{6}{\lambda}}m\delta_3^i + \bar{\phi^i}(x), \tag{3.8}$$

leads to the following expression for the effective potential:

$$V(\bar{\phi^i}) = m^2(\bar{\phi}_3)^2 + \sqrt{\frac{6}{\lambda}}m\bar{\phi}_3(\bar{\phi^i})^2 + \frac{\lambda}{4!}[(\bar{\phi^i})^2]^2. \tag{3.9}$$

Only the $\bar{\phi}_3$ field has a corresponding mass term—the $\bar{\phi}_1$ and $\bar{\phi}_2$ fields both appear to be massless, and the explicit $\bar{\phi}_3$ interaction term breaks the $O(3)$ invariance of the original Lagrangian. The Lagrangian written in terms of the shifted variables is still invariant under $O(2)$ rotations around the $\bar{\phi}_3$ axis, which mix only the $\bar{\phi}_1$ and $\bar{\phi}_2$ fields. The symmetries corresponding to rotations around the $\bar{\phi}_1$ and $\bar{\phi}_2$ axes have both been broken, and corresponding to each broken symmetry there is a massless scalar field.

Based on a similar calculation for a specific model, Goldstone conjectured that in general a massless scalar field appears for every broken continuous symmetry. Of the three proofs of this conjecture given in Goldstone et al. (1962), the simplest demonstrates that SSB entails the existence of a $p^2 = 0$ (zero-mass) pole in the field propagators for some of the shifted fields.[26] Suppose that we have a Lagrangian involving several fields, $\phi^a(x)$, such that given constant fields ϕ_0^a minimize the effective potential. Writing out V as an expansion around this minimum, we have

$$V(\phi) = V(\phi_0) + \frac{1}{2}(\phi - \phi_0)^a (\phi - \phi_0)^b \left(\frac{\partial^2 V}{\partial \phi^a \partial \phi^b} \right)_{\phi_0} + \cdots . \tag{3.10}$$

The coefficient of the second term (a symmetric matrix) is equal to the inverse of the momentum-space propagator,[27]

$$\left(\frac{\partial^2 V}{\partial \phi^a \partial \phi^b} \right) = \Delta^{-1}_{ab}(p). \tag{3.11}$$

A continuous symmetry transformation has the general form $\phi^a(x) \rightarrow \phi^a + i\epsilon T^{\alpha\ a}_{b} \phi^b(x)$ (summation over b understood, α labels distinct generators of the symmetry group). Assuming that

[26] See Weinberg (1996), §19.2 for two of the original proofs in updated notation, which I will follow here.

[27] See Weinberg (1996), Chapter 16 for justifi cation of this equality.

the original Lagrangian is invariant under this transformation,

$$V(\phi^a) = V(\phi^a + i\epsilon T_b^{\alpha\, a} \phi^b), \tag{3.12}$$

which can be rewritten as

$$T_b^{\alpha\, a} \phi^b \frac{\partial V(\phi)}{\partial \phi^a} = 0. \tag{3.13}$$

Finally, taking another partial derivative with respect to ϕ^c, we have

$$\left(\frac{\partial V}{\partial \phi^a}\right) T_c^{\alpha\, a} + \left(\frac{\partial^2 V}{\partial \phi^a \partial \phi^c}\right) T_b^{\alpha\, a} \phi^b = 0. \tag{3.14}$$

Evaluating this expression at ϕ_0, the first term must be zero since V is at a minimum, and thus the second term must vanish. Symmetry breaking implies that $T_b^{\alpha\, a} \phi_0^b \neq 0$, at least for some choices of α, and it follows that the matrix has a zero eigenvalue with the eigenvector $T_b^{\alpha\, a} \phi_0^b$ for these generators. This in turn implies that the propagator has a pole at zero momentum, corresponding to the existence of a massless spin-zero particle.

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, p. 516). What is to be done with the massless, spinless bosons (aka “Goldstone bosons”) produced by symmetry breaking? They rejected the possibility that the Goldstone bosons really do exist, on the grounds that such strongly coupled massless particles would have already been detected experimentally. They also did not see any way to modify the particle interpretation of the theory in order to “cancel” the Goldstone bosons, along the lines of the cancellation

of the timelike and longitudinal components of the electromagnetic field in the Gupta-Bleuler formalism in QED.[28] However, they did note that Goldstone's theorem does not apply either to discrete or gauge symmetries (Goldstone et al. 1962, p. 970):

> Goldstone has already remarked that nothing seems to go wrong if it is just discrete symmetries that fail to leave the vacuum invariant. A more appealing possibility is that the "ur symmetry" broken by the vacuum involves an inextricable combination of gauge and space-time transformations.

Several different physicists pursued this loophole in Goldstone's theorem, and independently discovered what is now called the "Higgs mechanism."

Anderson was the first to discuss the possibility that breaking a gauge symmetry might cure the difficulties with Yang-Mills theory (by giving the gauge bosons mass) without producing Goldstone bosons. Anderson argued that in several condensed matter systems with SSB 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, p. 422).[29] In addition he suggested that this "tangling" between Goldstone and gauge bosons could be exploited to introduce a massive gauge boson in gauge theories:

> It is likely, then, considering the superconducting analog, that the way is open for a degenerate vacuum theory of the Nambu type without any difficulties involving either zero-mass Yang-Mills gauge bosons or zero-mass Goldstone bosons. These two types of bosons seem capable of "cancelling each other out" and leaving finite-mass bosons only. (Anderson 1963, p. 441)

[28] Quantizing the electromagnetic field in Lortenz gauge leads to photons with four different polarization states: two transverse, one longitudinal, and one "timelike" (or "scalar"). In the Gupta-Bleuler formalism, the contributions of the longitudinal and timelike 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 Ryder (1996), §4.4 for a brief description of the Gupta-Bleuler formalism.

[29] In the case of superconductivity, the "Goldstone mode" becomes a massive plasmon mode due to the long-range Coulomb interactions.

Anderson supported these provocative remarks with neither a field theoretic model nor an explicit discussion of the gauge theory loophole in Goldstone's theorem. Peter Higgs has recently commented that "Anderson's remark was disbelieved at the time by those particle theorists who read it, myself included!" (Higgs 1997, p. 506-7) due to these shortcomings. Regardless of this disbelief, within a year of Anderson's 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). (Higgs won the naming contest.) Anderson's work and the earlier work of Nambu on superconductivity stimulated Higgs's interest in symmetry breaking. After following a discussion of the dependence of Goldstone's proof on Lorentz covariance in *Physical Review Letters*,[30] Higgs recognized that the failure of manifest Lorentz covariance for particular gauge choices (such as Coulomb gauge in QED) allows one to evade Goldstone's theorem in gauge theories. Within a week of reading Gilbert (1964), he formulated the simple field theory incorporating the Higgs mechanism presented in Higgs (1964).

In the model presented by Higgs, the massless Goldstone modes disappear from the physical particle spectrum, but in their ghostly gauge-dependent presence the vector bosons acquire mass. Higgs began by coupling the simple scalar field of the Goldstone model with the electromagnetic interaction. Take a model incorporating a two component complex scalar field, such that $\phi = \frac{1}{\sqrt{2}}(\phi_1 - i\phi_2)$ and $V(\phi) = \frac{1}{2}\lambda^2|\phi|^4 - \frac{1}{2}\mu^2|\phi|^2$ (compare eqn. (3.7)).[31] Including

[30] Klein and Lee (1964) and the response, Gilbert (1964); see Higgs (1997) for a short narrative of his discovery.

[31] See also the clear presentation given in Aitchison (1982), which I draw on here.

the electromagnetic interaction leads to the following Lagrangian:

$$\mathcal{L} = (D_\mu \phi)^\dagger (D^\mu \phi) - V(\phi) - \frac{1}{4} F_{\mu\nu} F^{\mu\nu}, \tag{3.15}$$

where $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$, and D is the covariant derivative operator defined as $D_\mu = \partial_\mu + ieA_\mu$. This Lagrangian is invariant under the following gauge transformations:

$$\phi(x) \rightarrow e^{-ie\theta(x)} \phi(x) \tag{3.16}$$

$$A_\mu \rightarrow A_\mu + \partial_\mu \theta, \tag{3.17}$$

where e is a constant introduced for convenience. Higgs' crucial insight was that a clever choice of gauge can be used to "kill" one of the components of ϕ. Explicitly, we can rewrite ϕ as follows:

$$\phi(x) = \frac{1}{\sqrt{2}} (\alpha + \beta(x)) e^{\frac{-i\xi(x)}{\alpha}}. \tag{3.18}$$

The gauge transformation (3.16) leads to $\phi' = e^{-ie\theta(x)} \phi(x)$, and the judicious choice of $\theta(x) = -\frac{1}{e\alpha}\xi(x)$ exactly cancels the exponential term in (3.18). As a result, $\xi(x)$ disappears from the Lagrangian rewritten in terms of the gauge transformed fields

$$\phi' = \frac{1}{\sqrt{2}} (\alpha + \beta(x)) \tag{3.19}$$

$$A'_\mu = A_\mu - \frac{1}{e\alpha} \partial_\mu \xi(x). \tag{3.20}$$

Neglecting an overall constant, the new Lagrangian is given by:

$$\mathcal{L} = (\frac{1}{2}(\partial_\mu \beta)(\partial^\mu \beta) - \frac{1}{2}\mu^2\beta^2) + (-\frac{1}{4}F'_{\mu\nu}F'^{\mu\nu} + \frac{1}{2}e^2\alpha^2 A'_\mu A'^\mu) \\ -\frac{1}{8}\lambda^2\beta^4 - \frac{1}{2}\lambda^2\alpha\beta^3 + \frac{1}{2}A'_\mu A'^\mu e^2(\beta^2 + 2\alpha\beta). \tag{3.21}$$

This Lagrangian displays several consequences of the Higgs mechanism: the massless Goldstone modes have disappeared (there are no fields without a mass term), the $\beta(x)$ field (aka the "Higgs field") has acquired a mass term, the $\xi(x)$ field has been "hidden" as a longitudinal mode of the vector field A'_μ, and the vector field has acquired a mass term (the fourth term above).

This list of interesting consequences of the Higgs mechanism follows from writing the Lagrangian with a particularly clever gauge choice. But conventional wisdom associates physical content only with gauge- invariant quantities and cautions against taking gauge-dependent quantities too seriously. So is the Higgs mechanism a piece of formal trickery devoid of physical content? Early studies of the Higgs mechanism (such as Kibble 1966) focused on isolating the gauge-invariant content of the Higgs mechanism by considering different gauge choices. In his 1973 Erice lectures, Sidney Coleman briefly raises worries about gauge invariance of the formalism (Coleman 1985, p. 168):

> People are sometimes worried that the formal apparatus for treating SSB ... is not gauge invariant. This is true; the vacuum expectation value of a scalar field, the effective potential, indeed the Feynman propagators themselves, are not gauge-invariant objects. This is also irrelevant. ... There is nothing wrong with [using gauge-dependent objects], as long as we are careful to express our final results in terms of gauge-invariant quantities, like masses and cross sections. The occurence of SSB does not affect this; the form of the effective potential and the location of its minimum are indeed gauge-dependent, but the values of masses and cross-sections computed with the aid of these objects are not.

We will see below that this warning to pay attention to the difference between gauge-dependent calculational tools and properly gauge-invariant physical content was not always taken to heart.

The Higgs mechanism provoked a great deal of interest; it could clearly be used to fix and combine two appealing ideas, as Anderson had speculated. Yang-Mills-style gauge theories and the general idea of SSB faced the same roadblock: the prediction of massless particles inconsistent with phenomenological constraints. The Higgs mechanism provided a way around the roadblock, and some hoped that the newly open road would lead to a gauge theory of the strong and weak interactions.[32] Three years after Higgs's paper, Weinberg incorporated the Higgs mechanism in a unified theory of the electromagnetic and weak interactions (Weinberg 1967), and a similar model was discovered independently by Salam. These theories faced another impressive roadblock, however: Weinberg, for example, tried to prove renormalizability of the theory for several years without success (Weinberg 1980, p. 518). Without a proof of renormalizability or direct experimental support the Salam-Weinberg model 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 shifting the fields in symmetry breaking would spoil renormalizability. Progress in the understanding of renormalization (due to work in the early 70s by 't Hooft, Veltman, Lee and others) revealed another important advantage of SSB: the renormalizability of a theory is actually *unaffected* by the occurrence of SSB. In the lectures quoted above, Coleman advertises this as the main selling point of SSB (Coleman 1985, p. 139).

[32]Englert and Brout (1964) explicitly mention 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 fi elds 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 62 in 1972, following 't Hooft's proof of renormalizability (Hoddeson et al. 1997, p. 16).

Testing the Higgs mechanism required a venture into uncharted territory. Although accelerator experiments carried out throughout the 70s 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, p. 274).

3.2.2 Conformal Symmetry Breaking

The proof of renormalizability of the Weinberg-Salam model in 1971 led to a dramatic increase in the study of unified gauge theories incorporating SSB, which was further enhanced by the experimental detection of neutral currents (predicted by the theory) in 1973. Part of this research effort was devoted to the study of symmetry restoration, the focus of the next section. But in addition, the success of this model encouraged the application of symmetry breaking in

[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 70s and early 80s.

other contexts, including two very different approaches to early universe cosmology. 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. In addition, Anthony Zee developed an account of gravitational symmetry breaking motivated by the desire to formulate a gravitational theory with no dimensional constants other than the mass term of a fundamental scalar field.

The members of the Brussels Consortium clearly share the desire to avoid the initial singularity we saw above, but like their countryman Lemaître decades earlier their main focus is on a quantum description of the "creation" event itself. Brout et al. (1978) begin by declaring their ambitious goal: replacing "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" (p. 78). As with Tryon (1973)'s slightly earlier proposal, this account is compatible with conservation of energy (Brout et al. 1978, put considerable emphasis on this point). Their theoretical analysis builds on a "deep analogy" between relativistic cosmology and conformally invariant QFT, which in practice involves two fundamental assumptions. First, the Consortium assumes that the universe must be described by a conformally flat cosmological model.[35] As a consequence, the metric for any cosmological model is related to Minkowksi space by $g_{\mu\nu} = \phi^2(x^i)\eta_{\mu\nu}$. The conformal factor

[35] 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, pp. 78-79).

$\phi(x^i)$ is treated as a massless scalar field conformally coupled to gravitation. The second fundamental assumption of their approach is that the "creation" event corresponds to a fluctuation of $\phi(x^i)$. This fluctuation breaks the conformal symmetry of the initial state, which is taken to be a constant $\phi(x^i)$ in a background Minkowski space.

The devil is in providing the details regarding the universe this "rational fireball" produces. According to their account, the fluctuation initially produces a de Sitter-like bubble. Particles are produced by a "cooperative" process: initial variations in the gravitational field produce massive scalar particles (via Parker's mechanism of gravitational particle creation), the particles create fluctuations in the gravitational field, the fluctuations seed further particle creation, and so on. Eventually the cooperative process comes to an end, and the primeval particles decay into matter and radiation as the universe slows from its de Sitter phase into free expansion. Although the details of these processes are meant to follow from the fundamental assumptions, a number of auxiliary conditions are introduced along the way in order to produce a universe something like our own. The malleability of the physical model is nicely illustrated by the evolution of the Consortium's thought: Brout et al. (1980) abandon the earlier ideas and instead suggest that particle production is a result of a "phase transition in which the 'edge of the universe' is the boundary wall between two phases" (p. 110). The Consortium ultimately failed to provide a concrete physical model (even when judged by the standards of contemporary particle physics!) that realized their programmatic aims.

Despite the difficulties in filling out the creation story, the Consortium does clearly explain several features of an early de Sitter phase. It will come as no surprise that an early de Sitter phase necessitates negative pressure, which Brout et al. (1978) explain as "the phenomenological expression of ... the creation of particles" (p. 85). Of greater interest is a comment buried in

an Appendix of Brout et al. (1978) (but mentioned more prominently in later papers—including the title of Brout et al. (1979)) regarding the impact of an early de Sitter phase on horizons. As with Sakharov's proposal, 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 question of *explaining* the early universe's uniformity via causal interactions. However, the absence of horizons is still mentioned as a way of solving the "causality problem"; in this model, the universe and all its contents can ultimately be traced back to the initial pointlike fluctuation of $\phi(x^i)$.

Zee (1979, 1980) proposed that incorporating symmetry breaking into gravitational theory (by coupling gravitation to a scalar field) leads to replacing the gravitation 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.[36] 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 Friedmann dynamics so that $a(t) \propto t$, which is enough to make the integral in eqn. (A.17) diverge—that is, the horizon distance goes to infinity as $t \to 0$. Zee clearly states the horizon problem and advertises this idea as a possible solution of it. According to Guth's recollections (Guth 1997a, pp. 180-81), a lunchtime discussion of Zee's paper in the SLAC cafeteria led him to discover that his own inflationary model also solves the horizon problem.

[36] 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.

3.2.3 Symmetry Restoration

In the condensed matter systems that originally inspired the concept of symmetry breaking, a variety of conditions (high temperature, large currents, etc.) led to restoration of the broken symmetry. Many of the leading researchers in QFT focused on symmetry restoration in particle physics in the early 70s (including Coleman, Weinberg, and Jackiw), perhaps due to the intricate connections between symmetry breaking, renormalization, and unified theories similar to the Weinberg-Salam model. From the outset, those studying symmetry restoration in gauge theories expected restoration to occur in the extreme conditions in the hot early universe, and nearly every paper on the subject includes speculations concerning the cosmological implications of these new results.

A cautionary note is in order before turning to calculations of symmetry restoration. The calculations discussed below all focus on the effective potential. Coleman (1985) warned against taking this quantity too seriously due to its gauge dependence, but the results below are all expressed in terms of properties of the effective potential rather than properly gauge invariant quantities. I have to admit that I haven't sorted out this murky issue. My conjecture is that there is an implicit assumption that unitary gauge (used throughout the calculations below) reflects the true physical degrees of freedom. This assumption comes into play in determining how the external heat bath couples to gauge degrees of freedom in finite-temperature field theory. Whether or not this conjecture holds historically or in terms of the physics involved, there is clearly a need for *some* justification that the quantities used in these calculations are not shifty gauge phantoms. However, I will leave the question aside for a later date.

Two Russians with backgrounds in condensed matter physics, Kirzhnits and Linde, were the first to study symmetry restoration in a model with global SSB (Kirzhnits 1972; Kirzhnits and Linde 1972). Based on a heuristic analogy with superconductivity and superfluidity, they estimate that the vacuum expectation value ϕ_0 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. 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 estimate that symmetry restoration occurs above $T_C \approx G^{-1/2} \approx 10^3 GeV$. In the standard hot big bang model, this temperature corresponds to approximately 10^{-12} seconds after the big bang. Kirzhnits and Linde (1972) further note that prior to symmetry breaking the weak interaction would have been a long-range interaction like electromagnetism, since it too would have been mediated by a massless gauge boson. This would lead to strong repulsive forces between any two bodies with unbalanced weak charges. They refrain from a more quantitative treatment of the cosmological implications of these ideas, citing the lack of a final formulation of the electroweak theory.

Much more rigorous calculations carried out over the next few years (by Kirzhnits, Linde, Weinberg, Dolan, Jackiw, Bernard and others) bolstered these initial rough results. These authors utilized and further refined the effective action formalism (a.k.a. the effective potential method) and applied it to finite-temperature field theory in order to calculate T_C and determine the nature of phase transitions in a variety of different field theories. The effective potential allows one to carry out a fully quantum calculation (i.e., one taking the possibly divergent higher-order corrections into account) along the same lines as the classical analysis sketched above. On the assumption that the vacuum state is translationally invariant, symmetry breaking occurs if the

following condition holds:

$$\frac{\partial V_{eff}(\phi_{cl})}{\partial \phi_{cl}} = 0 \quad \text{where} \quad \phi_{cl} = \langle 0|\phi(x)|0\rangle_{J(x)}, \tag{3.22}$$

where $J(x)$ is an external source.[37] $V_{eff}(\phi)$ is typically evaluated in a loop expansion, a perturbative expansion in terms of the number of closed loops appearing in the Feynman diagrams,[38] and in the case of weak couplings the second- and higher-order loops are usually negligible. Fortunately the ultraviolet divergences in this expansion can be cured with the same medicine used for an SSB-free Lagrangian; rewriting the Lagrangian in terms of the true ground state hides the original symmetry but does not change the ultraviolet-divergence structure of the theory.

These more rigorous calculations showed that symmetry restoration occurs as a consequence of the temperature dependence of the higher-order corrections to the effective potential. The full effective potential is the sum of a zero-temperature term and a temperature-dependent contribution calculated using the methods of finite-temperature field theory. Finite-temperature field theory includes interactions between the fields under consideration and a background thermal heat bath at a temperature T (neglected in conventional QFT, which treats interactions between fields in otherwise empty space). Finite temperature field theory was first developed in the 50s in order to study many-body systems in condensed matter physics. Remarkably, the difference between conventional QFT calculations and those in finite-temperature field theory lies

[37] The "classical field" $\phi_{cl}(x)$ is the vacuum expectation value in the presence of the source $J(x)$. For further discussion, see Coleman (1985, p. 132-142) or Peskin and Schroeder (1995, Chapter 11), which both use functional methods similar to those introduced in Dolan and Jackiw (1974). Cheng and Li (1984, §6.4) includes a detailed calculation of the effective potential up to one loop for ϕ^4 theory.

[38] This expansion is used rather than the usual coupling constant expansion since it is insensitive to "shifting" the fields appearing in the Lagrangian to the true ground state (and the subsequent shuffling of terms and coupling constants).

merely in a change of the boundary conditions used in evaluating path integrals.[39] In practical terms, the field theorist can use the same familiar Feynman diagrams and calculational tools of conventional QFT in finite temperature field theory, with the two-point function

$$D_\beta(x-y) = \frac{Tr(e^{-\beta H}T\phi(x)\phi(y))}{Tre^{-\beta H}} \tag{3.23}$$

replacing the usual zero-temperature two-point function. Thus the calculation of the temperature-dependent term ($\bar{V}_{eff}$) proceeds like the calculation of the zero-temperature V_{eff}, and $\bar{V}_{eff}$ generally includes a temperature-dependent mass correction, which changes the shape of the effective potential. For example, the Lagrangian in eq. (3.5) exhibits symmetry breaking only for a negative-mass term; for a positive-mass term, the global minimum of the potential lies at $\phi = 0$ and there is no symmetry breaking. If the following condition holds the negative-mass term will be exactly cancelled by a mass correction in $\bar{V}_{eff}$:

$$\frac{\partial \bar{V}_{eff}(\phi,T)}{\partial \phi^2}|_{\phi=0} = -\frac{m^2}{2}. \tag{3.24}$$

Above the critical temperature T_C (at which the equation above holds), due to the positive-mass correction in $\bar{V}_{eff}$ the total effective potential has a global minimum at $\phi = 0$ (see figure 3.1). Whether symmetry restoration occurs depends upon the nature of $\bar{V}_{eff}$ in a particular model; Weinberg (1974a) gives examples of models with no symmetry restoration and even low-temperature symmetry restoration.[40] Determining the nature of the symmetry breaking phase

[39] In particular, the Green's functions for finite temperature are periodic in Euclidean time.

[40] Symmetry restoration can also be induced by large external fields or high current densities; see Linde (1979) and references therein.

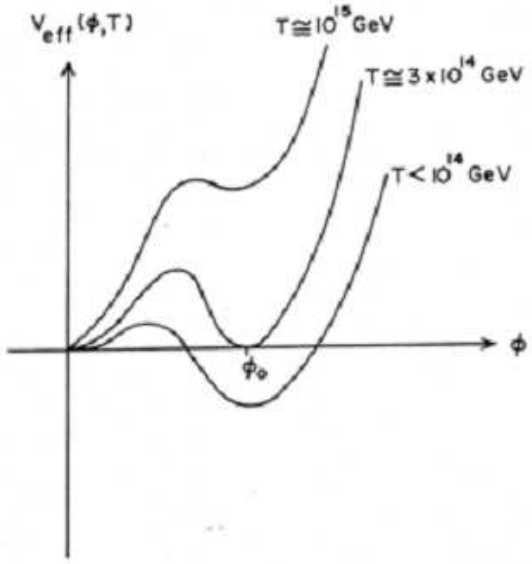

Fig. 3.1 This figure illustrates the temperature dependence of the effective potential of the Higgs field in the Weinberg-Salam model.

transition proved to more difficult than calculating the critical temperature. Weinberg (1974a) noted that the phase transitions in the models he considered appeared to be second-order but that he could not rule out weakly first-order phase transitions without a more detailed renormalization group analysis. The critical temperature is roughly of the same order of magnitude for a wide variety of models, but the nature of the phase transition depends more sensitively on the form of the effective potential and the specific values of coupling constants in the fundamental Lagrangian. The problem of determining the order of the phase transitions persisted throughout the following decade, and I will consider this in more detail in the next section. Briefly, in a second-order phase transition (such as the transition from paramagnetism to ferromagnetism), the so-called *order parameter* (i.e., the parameter which differs for the two phases) changes continuously at the critical temperature, whereas in a first-order phase transition (such as water freezing) the order paramater changes discontinuously at the transition. In the case of symmetry breaking, the order parameter is the vacuum expectation value of the Higgs field (rather than the

entropy, as in the cases above). In terms of the form of $V_{eff}(\phi, T)$, a first-order phase transition corresponds to ϕ evolving discontinuously through a barrier between the $\phi = 0$ local minimum and the true global minimum; if $V_{eff}(\phi, T)$ lacks a barrier the phase transition is second-order. The flexibility of model-building at the GUT scale leads to a number of different possibilities for the form of $V_{eff}(\phi, T)$, as we will see below.

3.3 Early Universe Phase Transitions

Explorers of the brave new early universe uncovered a number of possible consequences of symmetry breaking phase transitions starting in the early 70s. These can be grouped into roughly three different types of effects: (1) effects due to the different nature of the fundamental forces prior to the phase transition, (2) defect or structure formation during the phase transition, (3) effects of the phase transition on the cosmological evolution equations. The numerous researchers who contributed to this field reached widely divergent conclusions regarding these effects, due at least in part to the model dependence of the phase transitions and the wide variety of particle physics models under consideration.

The first works on phase transitions in cosmology (Kirzhnits 1972; Kirzhnits and Linde 1972) mention the possible consequences of long-range repulsive forces in the early universe. For example, prior to the electroweak phase transition any "weak charge" imbalance would result in long-range repulsive forces. Kirzhnits and Linde comment that "if a noncompensated charge is present (either electric or weak), a closed Universe with positive curvature cannot exist. Under the same conditions the isotropic and homogeneous Universe is also impossible" (Kirzhnits and Linde 1972, p. 474). Although this passage is not entirely clear (possibly as a result of a poor translation), I take Kirzhnits and Linde to be suggesting that long-range repulsive forces

would preclude a closed or open FLRW model, the latter presumably due to non-uniformities. Kirzhnits and Linde defer the detailed calculations meant to bolster this argument until a later paper (which I have not been able to find). In any case, this brief argument is insufficient to establish the general result. This conclusion holds only with a number of qualifications: the small horizon size prior to the phase transition limits the overall impact of such forces, and the action of these forces does not directly preclude particular models as Kirzhnits and Linde claim. Suitable qualifications might rule out sets of well-chosen initial conditions that produce a closed or open universe despite the presence of these forces. Whether this set is of small or zero measure would need to be established by considering the details of models incorporating long-range forces; as far as I know, neither Kirzhnits nor Linde pursued this line of thought. By way of contrast, later researchers hoped 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 replace Misner's neutrino viscosity and insure isotropization at very early times. They conclude that although GUT 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. Unlike the shaky speculations regarding the implications of long-range forces, physicists were able to explicitly calculate the consequences of a "domain structure" in the early universe for specific models. Other calculations depended more directly on the details of the phase transition; in contrast, defect formation depends primarily on the topological structure of the vacuum solutions to a particular field theory and is relatively insensitive to the finer details regarding the phase transition. One of the earliest calculations of the consequences of domain structure focused

on discrete symmetry breaking: Zel'dovich et al. (1975) assume that breakdown of CP invariance in the early universe would result in regions with opposite signs for CP-noninvariant effects separated by domain walls. They showed that the large energy density trapped in the domain walls would produce inhomogeneities far too large to fit observational constraints. Moreover they calculate the equation of state for this "cellular medium" (averaged over a volume containing both domain walls and the empty cells), and demonstrate that evolution dominated by matter in this state solves Misner's horizon problem. Zel'dovich et al. (1975) clearly recognize the implications of this result:[41]

> 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]. ... But it is just such a situation which Misner considered necessary, and for its sake he considered preferable an anisotropic closed model of the universe. The condition $X_c >> X_p$ denotes the possibility in principle that the conditions in the accessible part of the Universe are evening out. In the domain theory this condition is compatible with a flat or with an open or homogeneous and isotropic Friedmann cosmological solution!

Like Sakharov's discussion of the horizon problem, this apparently escaped notice—the authors' main interest was to establish the incompatibility of this model with observations, and they do not mention the horizon problem in the abstract, introduction, or conclusion.[42] In addition, the paper focused on discrete rather than continuous SSB, whereas the Weinberg-Salam theory and various unification schemes incorporated the latter.

Kibble and a number of other researchers focused on several different types of "topological defects" that may form during phase transitions in the more interesting case of gauge theories with SSB. These defects result from the domain structure of the Higgs field following a phase

[41]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.

[42]Contemporary review articles, such as Linde (1979), also do not mention the horizon problem in connection with this paper.

transition and further depend on the topology of the space of non-singular finite energy solutions to the field equations. The Higgs field develops a complicated domain structure because the symmetry is broken differently in distant regions.[43] The particle horizon at the time of the phase transition sets an upper limit on the correlation length of the Higgs field (which generally depends upon the details of the transition, and may be shorter than the horizon distance).[44] In most cases this complicated domain structure will disappear as the Higgs fields in different regions become "aligned," but for some particular models no continuous evolution of the field can eliminate all of the nonuniformities; topological defects are the resulting persistent structures.

Later work on the formation of defects in theories with SSB of local gauge symmetries also ran afoul of observational constraints. Thomas 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

[43] In the case considered above, the form of the Lagrangian determines the magnitude of $\langle\phi\rangle$ but not the direction in field space in which symmetry breaking occurs; in the absence of some correlation between two regions, one would expect the direction of symmetry breaking to differ.

[44] Kibble (1976) first argued that the correlation length of the Higgs field should be less than the horizon scale initially (based on the Landau-Ginzburg theory of phase transitions). He later stated this point as follows: "...there can surely be no correlation extending beyond the current 'horizon,' at a distance ct. More remote parts of the universe can have no prior causal contact, at any rate in the conventional picture" (Kibble 1980, p. 189). Wald criticizes the assumption that the field is uncorrelated, see Wald (1992).

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), and 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![45] The story was the same for other types of defects: Zel'dovich et al. (1975) and Kibble (1976) both showed that domain walls are incompatible with the observed homogeneity of the CMBR.

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. The stress energy tensor for a scalar field is given by

$$T_{ab} = \nabla_a\phi\nabla_b\phi - \frac{1}{2}g_{ab}g^{cd}\nabla_c\nabla_d\phi - g_{ab}V(\phi); \tag{3.25}$$

if the derivative terms are negligible, $T_{ab} \approx -V(\phi)g_{ab}$. 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,

[45] 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.

[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, as "Is the Lee Constant a Cosmological Constant?"; the correct title is "Is the Cosmological Constant a Constant?"

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, p. 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

[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 desribed 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."

"cosmological constant problem" (see Rugh and Zinkernagel 2001). 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 70s 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, developed a more general model of the phase transition in which the vacuum energy does not vanish immediately; instead, it drops with temperature such that $\rho_{vac} \approx \rho_0(\frac{T}{T_c})^{\beta}$ where β is a small number to be derived from particle physics (Kazanas sets $\beta = 2$ in his calculations). This decaying vacuum energy alters the standard FLRW dynamics; Kazanas shows that rather than the standard $a(t) \propto t^{1/2}$ for a radiation-dominated solution,

$$a(t) \approx a_0(t/\tau)^{1/3} \exp^{3/4(t/\tau)^{2/3}}, \tag{3.26}$$

where $\tau = (8\pi\rho_0/3)^{-1/2}$ is the characteristic time scale of the expansion. Kazanas clearly recognizes the implications of this vacuum-driven expansion (Kazanas 1980, p. 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.

In his concluding remarks Kazanas also notes the often implicit assumption of previous work that the phase transition is of second-order. If the transition is first-order, the Higgs field will

remain trapped in a false vacuum state for some time—and the huge vacuum energy density will dominate the dynamics.

When discussed at all, this assumption was motivated by a widely recognized problem with first-order transitions: the formation of large scale inhomogeneities (due to bubbles of the new phase forming within the old phase) would conflict with the observed isotropy of the CMBR. Sato (1981) considers a first-order transition in some detail, and derives constraints on various parameters (such as the nucleation rate for the bubbles) which must be met in order to produce a homogeneous universe.[50] Although Sato (1981) appears optimistic that a first-order transition can produce a homogeneous universe, he does not comment on whether the stringent constraints he derives are satisfied by any particle physics models of the phase transition. Linde also mentioned the difficulties associated with bubble formation as a justification for focusing on second-order transitions. He only considered first-order transitions in any detail in an unpublished paper with Chibisov on the "cold universe" model; in this model, symmetry breaking is induced by the increase in fermion density even though the temperature never reaches the critical temperature.[51]

Guth and Tye (1980) argued that a first-order phase transition had an appealing consequence alongside the unappealing formation of bubbles: such a transition could alleviate the monopole problem, since 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 consquence of bubble wall collisions. Thus the abundance of monopoles ultimately depends upon

[50]Sato's overall interest was in the possibility of a baryon-symmetric early universe, and he hoped that an early phase transition would effectively separate regions of matter and anti-matter.

[51]See Linde (1979, p. 433-34); 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, p. 485-86).

the nucleation rate of the bubbles. Guth and Tye (1980) argue that reasonable models of the phase transition have a low nucleation rate, leading to a tolerably low production of monopoles.[52] Shortly after submitting this paper, 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 other appealing features of an extended phase of expansion discussed above. The next chapter begins with an assessment of Guth's seminal paper, which gave a persuasive presentation of the advantages of an extended "inflationary" phase.

3.4 Conclusions

The consolidation of the suprisingly successful Standard Model of cosmology by the early 70s encouraged forays into early universe cosmology along a number of different theoretical trails. As with other cases of extending a successful theory, advocates of speculative new lines of research marshall whatever arguments they can to decide which trails lead to promising new insights and which should be abandoned, and to persuade their colleagues to share their assessments. In closing I would like to emphasize two points regarding the arguments made in favor of the new ideas discussed in this chapter.

First, many of the proposals above promised to solve problems "internal" to cosmology, although there was not complete agreement regarding the legitimacy of various problems. The previous chapter described various fine-tuning problems that motivated Misner's program, but these problems were not broadly recognized or even regarded as legitimate problems by all the researchers discussed above. Sakharov and Zel'dovich both duly note their solutions of the

[52] Guth and Tye (1980) do not suggest any mechanism for eliminating the resulting inhomogeneity following such a transition.

horizon problem, but neither places much emphasis on this consequence (it doesn't even find its way into the abstract or conclusion of Zel'dovich et al. 1975). Starobinsky explicitly contrasted his approach to Misner's chaotic cosmology: he proved that the de Sitter solution is a solution of the EFE modified to incorporate quantum effects, but in the absence of a uniqueness result simply stipulated that the early universe must have begun in this highly specialized state. Clearly among the Soviet cosmologists avoiding the initial singularity was a greater virtue for a cosmological model. On the other hand, several cosmologists (mainly described in the previous chapter) did see fine-tuning problems as important guides in developing new theories.

Second, the interest in symmetry breaking phase transitions arose from advances in particle physics that had nothing to do with cosmology *per se*. As I argued above, the introduction of symmetry breaking was one of the profound conceptual innovations leading to the Glashow-Weinberg electroweak theory. However, the Higgs mechanism proved to be nearly beyond the reach of the standard techniques of experimental high energy physics. But as Kirzhnits and Linde (1972) emphasized, the Higgs mechanism does have striking consequences for cosmology on the assumption that it describes a real phase transition occurring in the early universe. The application of the Weinberg-Salam theory and GUTs led to a second distinct group of problems in early universe cosmology: was this amalgam of the two Standard Models internally consistent and compatible with observations? Unlike the application of nuclear physics to cosmology in the development of the big bang theory, this is not a case of applying well understood physics with few doubts regarding its validity; instead, cosmological considerations were used as a substitute for accelerator experiments in validating the novel ideas incorporated in GUTs. Faced with stark conflicts between observation and the consequences of early universe phase transitions, some particle physicists such as Veltman were willing to entirely abandon the Higgs

mechanism. However, as we will see in the next chapter, the study of first-order phase transitions showed that particle physicists could have their Higgs mechanism and solve cosmological problems too.

Chapter 4

An Inflationary Field

4.1 Old and New Inflation

Various paths led to the idea that the early universe passed through a first-order phase transition. Guth and Tye (1980); Einhorn and Sato (1981) both argued that a first-order transition could ease the conflict between the abundance of monopoles calculated in the simplest GUTs and strong observational limits that were several orders of magnitude lower. But another novel consequence of a first-order transition, namely that an extended false vacuum phase would lead to a period of exponential expansion, initially received mixed reviews. From the very outset, Guth labeled his recognition of the connection between such an "inflationary" stage and the flatness problem a "spectacular realization." His seminal paper (Guth 1981) presented a clear rationale for further study of phase transitions producing an inflationary stage, and it had an immediate and lasting impact on research in the field. In light of Guth's paper and subsequent developments, it is easy to overlook contemporary work that emphasized the negative consequences of first-order phase transitions.

First-order phase transitions have the undesirable consequence of producing large inhomogeneities, due to the formation of "bubbles" of the new phase immersed in the old phase. As we saw above, this problem deterred Linde from considering first-order phase transitions, and Sato (1981) derived a number of stringent constraints a GUT would need to satify to avoid excessive inhomogeneities. Even Guth's contemporaries who recognized that a first-order transition

would produce a stage of exponential expansion regarded the idea with some skepticism. Einhorn and Sato (1981) focused on the difficulties that arise if the phase transition ends with bubble nucleation by quantum tunneling.[1] In particular, they argue that "fast nucleation" (a rapid phase transition without exponential expansion) is inconsistent with the observed baryon-to-entropy ratio, whereas "slow nucleation" (a slow transition with an extended period of exponential expansion) does not lead to full completion of the phase transition. Their paper concludes on the following cautionary note:

> 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,[2] we would speculate that the problem here is less the choice of GUT but rather reconciling gravity with quantum field theory.

Two choices have led to an apparent dead end: including the false vacuum energy as an effective cosmological constant, and describing the symmetry-breaking phase transition as a first-order transition (with a range of parameter choices based on particular GUTs). Einhorn and Sato suggest that the problem lies with the first choice and not the second. But by the time their paper appeared in print, Guth (1981) had given a persuasive argument that an inflationary stage is a desirable consequence of an early universe phase transition, rather than something to be avoided.

---

[1] The original draft of this paper was completed in July 1980, and 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 (1997a, p. 180), Einhorn and Sato hit upon the idea of false-vacuum driven exponential expansion independently.

[2] Here Einhorn and Sato cite Mohapatra and Senjanovic (1979a,b); Mohapatra (1980), which discuss gauge theories in which CP symmetry remains broken at high temperatures. Mohapatra (1980) emphasizes that in such theories the cosmological constant can be made arbitrarily small throughout the history of the universe.

Guth (1981) presents an inflationary stage as the decisive solution to two problems facing cosmology, and only secondarily as a solution to the monopole problem.[3] Following his work with Tye on the monopole problem, on the evening of Dec. 6, 1979 Guth calculated the impact of a first-order phase transition on the evolution of the universe. His work notebook from the next day (on display at the Adler Planetarium in Chicago) 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." [4] Guth's calculations showed that the false vacuum energy present (with energy density ρ_0) during a first-order phase transition, assuming it couples as an effective cosmological constant, drives exponential expansion $a(t) \propto e^{\chi t}$ where $(\chi)^2 = (8\pi/3)\rho_0$. An extended period of exponential expansion enormously suppresses the curvature term in the Friedmann equation (see Appendix A.2). 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. Guth (1981) contrasts this situation with the standard big bang theory: since $\Omega = 1$ is unstable under dynamical evolution governed by the Freidmann equation, *sans* inflation the initial value of Ω at the Planck time, for example, must be fine tuned to an accuracy of one part in 10^{59}.[5]

[3] The monopole problem is mentioned in the penultimate line of the abstract, whereas the title, abstract, introduction, and conclusion all emphasize the inflationary solution to the horizon and flatness problems.

[4] See Guth (1997a), Chapter 10 for a detailed account (quotation on p. 179). Guth attended a lecture by Dicke, in which he mentioned the flatness problem, on Nov. 13, 1978.

[5] Taking $t \rightarrow 0$ the value of Ω must be finely tuned to arbitrary accuracy. On the other hand, as Guth (1981) points out, this impressive degree of fine tuning does not depend strongly on when the initial conditions are taken to be fixed; at a more modest energy of $10^{14} GeV$ (compared to $10^{19} GeV$ for the Planck scale), Ω must still be fine tuned to one part in 10^{49}.

Guth was one of the first to clearly recognize that an inflationary stage solves the flatness problem in this sense.[6] By way of contrast, inflation's implications for horizons, understood by a number of Guth's contemporaries, appear to have come to him as a pleasant surprise that enhanced his confidence in the idea. Guth (1997a, pp. 180-81) recounts that a lunchtime discussion of Zee's work in the SLAC cafeteria led him to calculate the effects of inflation on horizon structure. He found that inflation 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 CMBR surpasses the current visual horizon. (As Guth notes, horizons do not disappear; they are only pushed beyond the observed part of the universe. See Appendix A.3 for further discussion of the horizon problem and the impact of inflation.) Guth (1981) again emphasizes the striking difference between this feature of the inflationary universe and the standard cosmology (p. 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 pre-inflationary patch of sub-horizon scale expands to encompass the observed universe.

In the previous chapter I argued that earlier, similar proposals faced two general difficulties: identifying the source of the false vacuum state, and giving an account of the transition from a de Sitter-like stage to FLRW evolution. The vacuum energy of a Higgs field trapped in a false vacuum state is the source in Guth's model.[7] Guth's only comment regarding whether current GUTs incorporate a Higgs potential with the appropriate properties is quite cautious (p.

[6] Linde (1979) cites a preprint by Rubakov (dated 1979) that apparently includes a discussion of the flatness problem and a solution of it via a stage of exponential expansion (cf. Linde 2002). This paper was never published, and I have not obtained a copy of the preprint.

[7] Guth (1981) uses a bare minimum of field theory, introducing only the energy density as a function of temperature for a false vacuum state, leaving a more detailed account of first-order transitions in the context of an $SU(5)$ GUT for Guth and Weinberg (1981).

352): "[G]rand unified models tend to provide phase transitions which could lead to an inflationary scenario of the universe." He is also admirably frank regarding his failure to account for the transition to FLRW expansion, later dubbed the "graceful exit" problem. This failure stems from the fact that 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. 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. 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.[8] 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.

Like earlier work, Guth's proposal failed to solve the graceful exit problem. But rather than abandon the idea, Guth argued that the explanatory advantages of inflation were reason enough to pursue the idea further. Guth's rationale for inflation has a familiar form: an innovative new idea is to be pursued further since it eliminates a number of unexplained coincidences required by existing theory, and in that sense offers a better explanation of observed regularities. The case of Copernican astronomy supplies a shopworn example of this type of argument, which

[8]See Guth and Weinberg (1983); Blau and Guth (1987) for clear descriptions of the graceful exit problem, and the calculations leading to the conclusion that this is a death blow to the initial model.

Janssen (2002) dubs a "Common Origin Inference" (COI).[9] In Ptolemaic astronomy, the model for motion of each planet is closely tied to the motion of the sun, although there is no reason for this connection other than fidelity with observations.[10] In Copernican astronomy, the correlated components are due to the Earth's motion around the sun. The Ptolemaic system incorporates the correlation by appropriately tuning the parameters of the planetary models; the Copernican theory is to be preferred because it traces these parameters to their "common origin," namely the simple geometry of Copernicus's heliostatic system. This style of argument figures prominently in reconstructions of other debates regarding the introduction of a new theory (see Janssen 2002, for three other case studies).

Guth's argument took a similar form: he emphasized that an inflationary stage eliminates several fine-tuning problems of standard cosmology. The standard cosmology requires that at the Planck time the 10^{83} causally disjoint patches constituting the observable universe began in a state of pre-established harmony that would shock even Leibniz: they must have been at the same temperature, with a delicately tuned overall energy density. Guth showed that taking GUT-scale phase transitions into account might eliminate the need for such unexplained coincidences. Another fine-tuning problem was added to the list in 1982, as we will see below: inflation provides a mechanism for generating the slight departures from homogeneity needed to seed galaxy formation. The study of GUT-scale phase transitions was still in the full bloom of youth, and Guth stressed the plausible hope that a more mature model would avoid the problems plaguing his initial idea. (Guth (1981) already mentions the possibility, studied by Edward Witten, that

[9] Here I will leave aside my doubts regarding whether this reconstruction of the case for Copernican astronomy can be defended as an accurate historical account.

[10] The center of the epicycle lies along the same line of sight as the mean sun for the interior planets. For the superior planets, the line from the center of the epicycle to the planet is parallel to the line of sight of the mean sun.

symmetry breaking of a Higgs field obeying the Coleman-Weinberg condition might yield an acceptable inflationary model.)

From the outset Guth's proposal was greeted as a major development. Sidney Coleman called Guth's first presentation on inflation the best seminar of the year at SLAC, where Guth was a postdoc (as Guth warmly recalls in Guth 1997a, p. 187). Word spread quickly enough that job offers began materializing within five days, and a month later (February 1980) Guth was on the lecture circuit introducing his ideas to a large cross-section of the particle physics community. By the following summer Guth was on his way to a faculty position at his *alma mater*, MIT. In addition to this quick upturn in his career path, Guth's talks and his paper (submitted in August of 1980) had two fairly immediate impacts. Guth introduced many astrophysicists and particle physicists to the very idea of early universe cosmology. Even those who had been aware of earlier work, such as Martin Rees, have commented that they only understood earlier results in light of Guth's paper.[11] By admitting the flaws of his initial model, Guth also left his audiences with a clearly formulated task: to find a working model of inflation. Paul Steinhardt, then a Junior Fellow in the Harvard Society of Fellows, describes Guth's talk at Harvard as "the most exciting and depressing talk" he had ever attended (Steinhardt 2002). The excitement stemmed from the promise of connecting the study of phase transitions to fundamental questions in cosmology. But after laying out inflation's ability to solve the trio of problems mentioned above, Guth ended by explaining the fatal flaw of his initial model. Steinhardt recalls his reaction: "Here was this great idea and it just died right there on the table. So I couldn't let that happen."

[11] 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 (see Lightman and Brawer 1990, p. 161 and Rees 2002).

Given Steinhardt's background in condensed matter physics and familiarity with phase transitions, he was ideally suited to take on the task of reviving Guth's idea. News of Guth's paper also reached Andrei Linde in Moscow, and inspired him to reconsider the models he had previously dismissed as "garbage". Steinhardt began studying early universe phase transitions almost immediately, and upon taking a faculty position at the University of Pennsylvania he found a graduate student, Andy Albrecht, eager to join in the project. Linde and Steinhardt and Albrecht independently realized that a symmetry breaking phase transition governed by a different effective potential could avoid Guth's graceful exit problem while providing sufficient inflation.

At roughly the same time, Hawking and Ian Moss proposed an alternative solution to the graceful exit problem. Although Hawking and Moss (1982) is sometimes cited as a third independent discovery of new inflation, it differs substantially from the other proposals.[12] The aim of the paper is to show that including the effects of curvature and finite horizon size leads to a different description of the phase transition. This phase transition proceeds from a local minimum at $\phi = 0$ to the global mimimum ϕ_0 via an intermediate state ϕ_1; rather cryptic arguments lead to the conclusion that "the universe will continue in the essentially stationary de Sitter state until it makes a quantum transition everywhere to the $\phi = \phi_1$ solution" (p. 36). They further argue that following this transition to a coherent Hubble scale patch, ϕ will "roll

[12] According to Steinhardt, Hawking demanded that Turner and Barrow mention his own paper along with Linde (1982); Albrecht and Steinhardt (1982) as the sources of new inflation in their review of the Nuffield conference. Hawking later alleged (in the first printing of Hawking 1988) that the discovery was not independent: Steinhardt could have been introduced to Linde's ideas in a lecture Hawking had given, and Steinhardt had attended, at Drexel University. (Hawking and Moss cite conversations with Linde as the stimulus for their own research.) Steinhardt found a videotape of the Drexel lecture, and proved that Hawking's memory was faulty: he made no mention of Linde or new inflation. Hawking made an apology of sorts and edited the offending passage, but Steinhardt regards Hawking's handling of the affair to have been "dishonorable" (Steinhardt 2002).

down the hill" (for an appropriate values of parameters in the effective potential), producing an inflationary stage long enough to match Guth's success.

Unlike Hawking and Moss's paper, Albrecht and Steinhardt (1982) and Linde (1982) both developed models of the phase transition based on a Coleman-Weinberg effective potential for the Higgs field. (Ironically Erick Weinberg, Guth's collaborator in calculating the disastrous effects of bubble formation in the original scenario, did not recognize the utility of the potential bearing his name.) The difference can be illustrated with the following general form for the effective potential, including one-loop corrections:[13]

$$V(\phi) = -\frac{1}{2}(2B + A)\sigma^2\phi^2 + \frac{1}{4}A\phi^4 + B\phi^4 ln\left(\frac{\phi^2}{\sigma^2}\right), \tag{4.1}$$

where $\sigma = \langle\phi\rangle$. Unlike the potential Guth used, for a Coleman-Weinberg model the quartic term at "tree level" vanishes, i.e. $2B + A = 0$.[14] Coleman and Weinberg (1973)'s surprising insight was that higher order corrections could nevertheless induce symmetry breaking, *without* invoking temperature-dependent terms. For example, if $B > 0$ this effective potential has extrema at $\phi = 0$ and $\phi = \pm\sigma$; for appropriate values of A, B, and σ, $\phi = 0$ is only a local minimum with the global minimum given by $\phi = \pm\sigma$. The difference between the effective potentials for old and new inflation is illustrated in Figures 4.1 and 4.2. Although this may appear to be a slight adjustment of the effective potential, it leads to a dramatically different phase transition. The most important consequence is that inflation continues after the formation of an initial bubble: rather

[13] A is a free parameter, whereas B is a constant depending on the coupling constants of the specific GUT used, and the vacuum expectation value σ also depends on the coupling constants and the masses of the gauge bosons. See, e.g., Kolb and Turner (1990), Chapter 7, for detailed discussions of the effective potential in particular models. Here I am neglecting temperature-dependent terms.

[14] The third term is the one-loop correction to the tree level (or classical) expression.

than tunnelling directly to the global minimum, in this scenario the field ϕ evolves to the minimum over a "long" timescale τ (i.e., much longer than the expansion time scale). Throughout this evolution ϕ is still displaced from the global minimum, and the non-zero $V(\phi)$ continues to drive exponential expansion. Linde (1982); Albrecht and Steinhardt (1982) both argue that for natural values of τ the expansion lasts long enough that the initial bubble is much, much larger than the observed universe; Linde estimates that the bubble radius will be $\approx 10^{3240} cm$ at the end of the inflationary stage, compared to $10^{28} cm$ for the visual horizon (see Appendix A.3 for the definition of visual horizon). Finally, as in Guth's scenario any pre-inflationary matter and energy density are diluted during the extended inflationary stage. In the new secnario, oscillations of the field ϕ near its global minimum would produce other particles via baryon-number non-conserving decay in order to "reheat" the universe to an energy density compatible with standard cosmology.

The initial proposals were quickly developed into a general account of new inflation. The features of the phase transition can be described simply in terms of the evolution of ϕ, which is determined by the form of the potential $V(\phi)$. The classical equations of motion for a scalar field ϕ with a potential $V(\phi)$ in an FLRW model are given by:

$$\ddot{\phi} + 3H\dot{\phi} + \Gamma_{\phi}\dot{\phi} + \frac{dV(\phi)}{d\phi} = 0, \tag{4.2}$$

where $\dot{\phi} = \frac{d\phi}{dt}$ and Γ_{ϕ} is the decay width of ϕ.[15] New inflation requires a long "slow roll" followed by reheating. Assume that the field ϕ is initially close to $\phi = 0$. Slow roll occurs if the potential is suitably flat near $\phi = 0$ and the $\ddot{\phi}$ term is negligible; given the further assumption

[15] One of the main differences between the initial papers on new inflation is that Albrecht and Steinhardt (1982) explicitly include the $3H\dot{\phi}$ term (aka the "Hubble drag" term), whereas Linde (1982) does not.

that the Γ_ϕ term is negligible, then the evolution of ϕ can be approximately described by:

$$3H\dot{\phi} \approx -\frac{dV(\phi)}{d\phi}. \tag{4.3}$$

(The name is due to the similarity between the evolution of ϕ and that of a ball rolling down a hill, slowed by friction.) More precisely, the following two conditions on the potential are necessary conditions for the slow roll approximation to apply:[16]

$$\frac{M_{pl}^2}{2}\left(\frac{V'}{V}\right)^2 << 1 \tag{4.4}$$

$$\left|M_{pl}^2\frac{V''}{V}\right| << 1 \tag{4.5}$$

These conditions are not sufficient, however, since the value of $\dot{\phi}$ must also be small initially for the approximation to hold. During slow roll the potential energy $V(\phi)$ dominates over the kinetic energy $\frac{\dot{\phi}}{2}$ (as a consequence of these conditions), and $V(\phi)$ drives inflationary expansion. The slow roll approximation breaks down as the field approaches the global minimum. The Γ_ϕ term is put in "by hand" to describe reheating: roughly, ϕ oscillates around the minimum and decays into other types of particles. The details depend on the coupling of ϕ to other fields, and are heavily model-dependent.

By the spring of 1982 several groups were at work fleshing out the details of the new inflationary scenario: a large group at the University of Chicago and Fermilab including Turner and Kolb, Steinhardt and Albrect at the University of Pennsylvania, Guth at MIT, Linde and various collaborators in Moscow, Laurence Abbott at Brandeis, Hawking and others in Cambridge,

[16] Here I have introduced the reduced Planck mass, defined by $M_{pl} = (8\pi G)^{-1/2}$.

and John Barrow in Sussex. With notable exceptions such as Hawking and Barrow, nearly everyone in this research community came from a background in particle physics. The framework described in the previous paragraph left ample room for innovation and new ideas: the connections with particle physics were poorly understood at best, the various approximations used were generally on shaky footing, and there were numerous hints of interesting new physics. Several of these researchers recognized the most important hint: homogeneity at all scales at the end of inflation would be incompatible with accounts of galaxy formation, which required an initial spectrum of perturbations. There appeared to be several ways to avoid *too much* homogeneity at the end of inflation; Linde (1982), for example, mentions a later phase transition without supercooling or quantum gravity effects as a possible means for generating inhomogeneities. In the next section we will see how these guesses were developed into inflationary cosmology's most fruitful connection with observations.

Guth's paper initiated a striking shift in the focus of research in early universe cosmology. The previous chapter described an exploratory phase of research: although it was clear that particle physics theories would lead to novel consequences when applied to the early universe, there were few signposts to guide theory development. Without clear observational anomalies, and great freedom in both particle theory and cosmology, efforts largely focused on determining the consequences—any consequences, even stark conflicts like the monopole problem—of plausible models of high energy physics. Guth's paper provided an agenda: rather than trying to determine the generic consequences for a broad range of particle physics models, nearly everyone in the field joined in the hunt for a workable model of inflation. The trio of problems Guth discussed become an entrance requirement: to be taken seriously, any new proposal had to solve the flatness, horizon, and monopole problems.

This situation resembles several other historical episodes in which the success of a new theory sets new standards, effectively upping the explanatory ante. Einstein's successful prediction of the anomaly in Mercury's perihelion shift raised the bar for gravitational theories: although the perihelion shift was not regarded as a decisive check prior to his prediction, after the prediction it served as a litmus test for competing theories of gravitation.[17] But there is also a striking difference between cases like this one and inflation. Long before Einstein began to formulate GTR, the obsessively precise work of Urbain Le Verrier and Simon Newcomb had established that Mercury's observed perihelion motion presented Newtonian theory with a *significant anomaly*, although opinions differed widely regarding how the anomaly should be handled (Roseveare 1982).[18] By way of contrast, the flatness problem was not even widely acknowledged *as a legitimate problem* prior to Guth's paper. In an appendix added to "convince some skeptics," Guth comments that:

> 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. (Guth 1981, p. 355)

[17] Einstein's contemporaries were aware that Mercury's perihelion motion would be sensitive to modifi cations of Newtonian gravitation, but early reviews did not use this test as a criteria for eliminating theories from consideration. For example, de Sitter (1911) remarks that Minkowski's sketch of a modifi ed gravitational theory leads to the wrong *sign* for the anomalous perihelion motion, but doesn't treat this as a reason for abandoning the theory.

[18] The essential point was established in Book I, Section 9 of Newton (1999): motion of the perihelion is the most sensitive measure of the exponent in the force law for gravitation. Almost all of the observed perihelion motion could be accounted for as perturbations due to interactions with the other planets. In light of the small remaining anomaly one could either modify the inverse square law or alter assumptions regarding the mass distribution in the solar system.

Guth's argument in the appendix may or may not have swayed many physicists, but the *existence* of a solution to the flatness problem lent the problem itself an air of legitimacy. Misner's description of the change in his assessment brought about by Guth's paper illustrates one common view among cosmologists:[19]

> 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. (Lightman and Brawer 1990, p. 240)

The widespread assessment that solving these problems is a "proper challenge" insures that only a theory that matches inflation's ante will be taken seriously as an alternative.[20]

But the contrast above brings out the troubling prospect that cosmologists have limited their explorations to theories that successfully solve pseudo-problems of their own making. The problems solved by inflation are defined against a speculative theoretical backdrop: the "questions of plausibility" Guth alludes to require assumptions regarding the initial state of the universe, which is acknowledged to be well outside the domain of current theory. The existence of these problems is based on a guess regarding the content of *future* theories: the as yet unformulated theory applicable to $t < 10^{-35}\ sec$ after the big bang is assumed to produce a nonuniform state. If this guess proves to be correct then something like inflation is needed to insure compatibility with observations, but if this guess proves to be wrong then the "problems" are red herrings. Unlike the empirical anomaly in Mercury's perihelion motion, the flatness problem is

[19] Lightman systematically asks every interviewee in (Lightman and Brawer 1990) about their assessment of the flatness problem before and after inflation, providing a useful cross section of the field. Several of the interviewees express views similar to Misner's, as did some of the cosmologists I interviewed (Barrow 2002; Ostriker 2002; Rees 2002). See also Brawer (1996) for a detailed discussion of the changing assessment of the horizon and flatness problems.

[20] This point has been clearly emphasized by several cosmologists; see, e.g., Edwin Turner's comments in Lightman and Brawer (1990, p. 317) and Peebles (1993, p. 393).

defeasible. (I will take up this concern with the nature of fine-tuning problems in more detail in Chapter 5.) Although some alternatives to inflation have been explored (discussed briefly below), overall Guth's rationale has convinced most cosmologists to focus their research efforts on developing a working model of inflation. This is a significant methodological shift away from clarifying the implications of different ideas in particle physics for cosmology.

4.2 The Nuffield Workshop: Birth of the Inflaton

The first international conference focusing on "very early universe cosmology ($t < 1$ sec)" convened in Cambridge from June 21 - July 9, 1982.[21] Nearly half the lectures at the Nuffield workshop were devoted to inflation, and the intense collaborations and discussions during the workshop led to the "death and transfiguration" of inflation (from the title of the conference review in *Nature*, Barrow and Turner 1982). One focus of the conference was the calculation of density perturbations produced during an inflationary stage: Steinhardt, Starobinsky, Hawking, Turner, Lukash and Guth had all realized that this was a "calculable problem" (in Steinhardt's words), with the answer being an estimate of the magnitude of the density perturbations, measured by the dimensionless density contrast $\frac{\delta\rho}{\rho}$, produced during inflation. Preliminary calculations of this magnitude disagreed by an astounding 12 orders of magnitude: Hawking circulated a preprint (later published as Hawking 1982) that found $\frac{\delta\rho}{\rho} \approx 10^{-4}$, whereas Steinhardt

[21] The description is taken from the invitation letter to the conference (Guth 1997a, p. 223). The Nuffield Foundation had previously sponsored conferences in quantum gravity, but shifted the focus to early universe cosmology in response to interest in the inflationary scenario. A 1981 conference in Moscow on quantum gravity also included numerous discussions of early universe cosmology (Markov and West 1984), but Nuffield was the first conference explicitly devoted to the early universe. The 30 participants included all of the cosmologists mentioned in the previous section except Einhorn and Sato; see the conference proceedings volume (Hawking et al. 1983) for a complete list of participants and their lectures.

and Turner initially estimated a magnitude of 10^{-16}.[22] After three weeks of effort, the various groups working on the problem had converged on an answer, but the answer proved to be disastrous for new inflation.

These difficult calculations promised to fill a well-recognized lacuna in existing accounts of structure formation. Although several alternative models had been developed from the 50s onward, by 1980 mainstream models of structure formation were based on Lemaître's idea that gravitational enhancement of inhomogeneity is the primary mechanism for structure formation. However, scaling arguments indicate that in the standard FLRW models these inhomogeneities must have been present from the earliest stages of the universe (see, e.g., Harrison 1968). Mainstream models of structure formation focused on the evolution of small primeval perturbations in background FLRW spacetimes, treated via the linearized EFE, although they differed on the nature of the initial perturbations. The initial perturbations were assumed to be a combination of the following distinct modes:[23]

- *adiabatic*: Fluctuations in energy density of nonrelativistic matter ρ_m matched by radiation fluctuations (also called "entropy perturbations"), $\frac{4}{3}\frac{\delta\rho_m}{\rho_m} = \frac{\delta\rho_r}{\rho_r}$,
- *isothermal*: Radiation is uniformly distributed, $\frac{\delta\rho_r}{\rho_r} = 0$, although the matter is non-uniformly distributed.

[22] The density contrast is defined by $\frac{\delta\rho(\mathbf{x})}{\rho} = \frac{\rho(\mathbf{x})-\bar{\rho}}{\bar{\rho}}$, where $\bar{\rho}$ is the mean density. An astute reader may worry about the gauge invariance of such a quantity, a point I will discuss below. These initial results are discussed in Guth (1997a, pp. 216-224).

[23] Zel'dovich introduced this terminology. The factor of $\frac{4}{3}$ arises since the energy density of radiation is $\propto T^4$, compared to T^3 for matter. These are called "adiabatic" perturbations since the local energy density of the matter relative to the entropy density is fixed. A third mode – tensor perturbations, representing primordial gravitational waves – were not usually included in discussions of structure formation, since they do not couple to energy-density perturbations.

The evolution of initial perturbations of these two types was studied throughout the 60s and 70s. In broad terms, two different schools of thought dominated the field: Zel'dovich's school focused on solutions in which large "blinis" (pancakes) formed first, fragmenting into galaxies and structures much later due to non-gravitational processes. The other school of thought developed a "bottom-up" scenario, in which initial fluctuations developed into proto-galaxies with larger structures forming later, all as a consequence of gravitational instability.

For both approaches the origin and nature of the initial perturbations were crucial components of the theory. Harrison, Peebles, and Zel'dovich independently suggested a particularly simple form for the initial perturbations: a scale-invariant spectrum of adiabatic perturbations (hereafter the HPZ spectrum) such that $\frac{\delta\rho}{\rho}|_{\lambda} = constant$ when λ, the perturbations' wavelength, is equal to the Hubble radius, $\lambda = H^{-1}$.[24] For different wavelengths the perturbation amplitude is fixed at different times: in an expanding universe, the wavelength λ increases with the scale factor $a(t)$ whereas the Hubble radius increases at a slower rate as the expansion slows.[25] The Hubble radius "crosses" various perturbation wavelengths in an expanding model; a scale-invariant spectrum deserves the name since the perturbations have the same magnitude as the Hubble radius sweeps across different length scales. The HPZ spectrum lacks characteristic length scales. Peebles, Zel'dovich and others were able to calculate the expected evolution of density perturbations (*modulo* a number of hotly debated auxiliary assumptions) and thereby

[24] The spectrum was introduced independently in Harrison (1970); Peebles and Yu (1970); Zel'dovich (1972). In general, for a scale invariant power spectrum the Fourier components of the perturbations obey a power law, $|\delta_k|^2 \propto k^n$; the Harrison-Zel'dovich-Peebles spectrum corresponds to a choice of $n = 1$ or $n = -3$, depending on the choice of volume element in the Fourier transform: $n = 1$ for a choice of $\frac{dk}{k}$, and $n = -3$ for $k^2 dk$. Finally, note that the Hubble radius does have dimensions of length: restoring c, it is given by $\frac{c}{H}$, and the Hubble constant is given in units of km per second per megaparsec.

[25] Since the perturbations grow with time, at a "constant time" the shorter wavelength perturbations have greater amplitudes for this spectrum. The difficulty with defining the spectrum of density perturbations in terms of "amplitude at a given time" is that it depends on how one chooses the constant time hypersurfaces.

constrain the initial spectrum based on observed large scale structure and, much more stringently, by the isotropy of the CMBR. Estimates of the magnitude of density perturbations when length scales associated with galaxies crossed the Hubble radius fell within the range $\frac{\delta\rho}{\rho} \approx 10^{-3} - 10^{-4}$. Finally, the initial perturbations were often assumed to be a form of random noise, which holds if the mass found within a sphere of fixed radius has a Gaussian distribution (for different locations of the sphere).[26]

Two features of the initial spectrum were particularly disturbing to cosmologists. The Hubble radius is equal to the length scale associated with a galaxy at around $t \approx 10^9$ seconds; at earlier times the perturbation was coherent on scales far larger than the Hubble radius. This would require trans-Hubble radius coordination, and this appeared to be in conflict with the presence of particle horizons. Bardeen concludes a study of the evolution of density perturbations as follows (Bardeen 1980, p. 1903):

> The one real hope for a dynamical explanation of the origin of structure in the Universe is the abolition of particle horizons at early times, perhaps through quantum modifications to the energy-momentum tensor and/or the gravitational field equations which in effect violate the strong energy condition.

But Bardeen's focus on particle horizons as the fundamental obstacle was not shared by others in the field; Peebles (1980), for example, frequently mentions the puzzles associated with horizons, but apparently takes this to be one of many indications that we do not sufficiently understand physics near the big bang. Secondly, it was difficult to imagine how to "spontaneously generate" an initial perturbation spectrum with an amplitude such that linear growth of the perturbations satisfies the constraint above. An early suggestion, that the density perturbations were due to

[26] Equivalently, for a Gaussian perturbation spectrum the phases of the Fourier modes δ_k are random and uncorrelated.

thermal fluctuations at early times, suffered from the following limitations (Peebles 1980, §4). Suppose that at some time t_i the matter and radiation in the universe is assumed to have thermal fluctuations away from uniformity, given by a Poisson distribution $\left(\frac{\delta\rho}{\rho}\right)_i \approx N^{-1/2}$, where N gives the number of particles. For a galaxy-sized lump of particles, say $N \approx 10^{80}$, the density contrast is $\frac{\delta\rho}{\rho} \approx 10^{-40}$. This is far too large for very early choices of t_i, such as the Planck time, if one takes the universe to begin at t_i with an "imprint" of thermal fluctuations (setting aside the question of how to arrange a coherent galaxy-mass fluctuation).[27] And it seems inappropriate to treat t_i as a free variable, choosing when to "imprint" a spectrum of thermal fluctuations such that the amplitudes match observations.

Cosmologists hoping to explain the origin of initial perturbations plucked ideas from the ample storehouse of speculative physics: Planck scale metric fluctuations, gravitational particle production, primordial black holes, "gravithermal" effects, primordial turbulence, non-equilibrium dynamics, and so on.[28] Sakharov (1966) was the first to propose a detailed quantum description of the initial perturbations, but this early paper drew no attention, partially because it was an extension of Zel'dovich's "cold bang" proposal that fell from favor following the discovery of the CMBR. From the mid-70s onward the newest item in the storehouse—early universe phase transitions—was pulled from the shelf and pressed into service. Zel'dovich (1980) proposed that string-like topological defects formed during a phase transition could provide an initial spectrum of density perturbations. Vilenkin also suggested strings as the origin of the initial spectrum in a series of seminal papers starting in 1981 (although his proposal differed

[27] Blau and Guth (1987) compare this density contrast imposed at $t_i = 10^{-35}$ seconds to the fluctuations required by the constraint mentioned in the last paragraph; evolving backwards, this constraint implies $\frac{\delta\rho}{\rho} \approx 10^{-49}$ at t_i, nine orders of magnitude *smaller* than thermal fluctuations.

[28] See Barrow (1980) for a brief review of some of these ideas and references, and Peebles (1980); Zel'dovich and Novikov (1983) for more comprehensive overviews of the field.

from Zel'dovich's), and the study of defects as the seeds of galaxy formation continued until the late 90s. But studies of the impact of phase transitions did not focus exclusively on the formation of defects. Press (1980) argues that the "tangled" nature of the gauge fields in different domains leads to inhomogeneities in the energy density appropriate to seed galaxy formation, but not topological defects.[29] In Press's scenario, inhomogeneities in the vacuum stress energy are converted into fluctuations in the energy density of matter and radiation. This "conversion" only works if the vacuum stress energy does not itself gravitate; Press notes the speculative nature of this suggestion, but argues that the other possibility – an incredibly precise cancellation of vacuum energy density – is equally unappealing.[30] Although I do not have space to discuss other proposals, in the spring of 1982 the origin of the initial perturbations remained one of the fundamental mysteries in cosmology.

The several groups studying the inflationary scenario faced a clear problem: does inflation succeed where these earlier proposals had failed? *Prima facie* inflation exacerbates the problem: as Barrow and Turner (1981) noted, inflation dramatically reduces the amplitude of any pre-existing fluctuations. But prior to Guth's paper Mukhanov and Chibisov (1981); Lukash (1980) had both argued that a de Sitter phase could generate perturbations by "stretching" zero-point fluctuations of quantum fields to significant scales. Hawking rediscovered the idea, and argued that initial inhomogeneties in the ϕ field would imply that inflation begins at slightly different times in different regions; the inhomogeneties reflect the different "departure times"

[29] Press justifies this neglect of topological defects as follows: "In this paper we will *not* need to postulate the formation of thin-walled domain structure ...," which are associated with discrete symmetries; later he comments that "We are interested in gradients of the [gauge fields] on scales many orders of magnitude larger than the horizon size at $T = T_c$, so there is no possibility of domains this large forming before the massive bosons have decayed away to lighter particles." But Press gives no argument to establish that other defects would not form. This is odd, especially since his argument depends on the same features that lead to defect formation: the gauge fields take different values in different domains, and the resulting "tangle" of values as the domains coalesce is the source of inhomogeneity.

[30] See Appendix B.2 for a brief discussion of the cosmological constant problem.

of the scalar field. Hawking's preprint claimed that this results in a scale-invariant spectrum of adiabatic perturbations with $\frac{\delta\rho}{\rho} \approx 10^{-4}$, exactly what was needed in accounts of structure formation. But others pursuing the problem (Steinhardt and Turner; Guth and his collaborator, So-Young Pi) did not trust Hawking's method; Steinhardt has commented that he "did not believe it [Hawking's calculation] for a second" (Steinhardt 2002, cf. Guth 1997a, pp. 222-230). There were two closely linked concerns with Hawking's method (beyond the sketchiness of his initial calculations): it is not clear how this approach treats the evolution of the fluctuations in different regimes, and it is also not gauge invariant.

The "gauge problem" in this case reflects the fact that a "perturbed spacetime" cannot be uniquely decomposed into a background spacetime plus perturbations. Slicing the spacetime up along different surfaces of constant time leads to different magnitudes for the density perturbations. The perturbations "disappear," for example, by slicing along surfaces of constant density. In practice, almost all studies of structure formation used a particular gauge choice (synchronous gauge), but this leads to difficulties in interpreting perturbations with length scales greater than the Hubble radius.[31] Press and Vishniac (1980) identify six "tenacious myths" that result from the confusion between spurious gauge modes and physical perturbations for $\lambda > H^{-1}$. This problem is significant for the inflationary account because over the course of an inflationary stage perturbations of fixed length go from $\lambda << H^{-1}$ to $\lambda >> H^{-1}$. Length scales "blow up" during

[31] Synchronous gauge is also known as "time-orthogonal" gauge: the coordinates are adapted to constant time hypersurfaces orthogonal to the geodesics of comoving observers. All perturbations are confined to spatial components of the metric; i.e., the metric has the form $ds^2 = a^2(t)(dt^2 - h_{ij}dx^i dx^j)$, with $i, j = 1, 2, 3$. The coordinates break down if the geodesics of co-moving observers cross.

inflation since they scale as $a(t) \propto e^{Ht}$, but the Hubble radius remains fixed since H is approximately constant during the slow roll phase of inflation.[32] For this reason it is especially tricky to calculate the evolution of physical perturbations in inflation using a gauge-dependent formalism. The first problem mentioned in the previous paragraph is related: determining the imprint of initial inhomogeneties requires evolving through several regimes, from the pre-inflationary patch, through the inflationary stage and reheating to standard radiation-dominated evolution.

Hawking and Guth pursued refinements of Hawking's approach throughout the Nuffield Workshop.[33] The centerpiece of these calculations is the "time delay" function characterizing the start of the scalar field's slow roll down the effective potential. This "time delay" function is related to the two-point correlation function characterizing fluctuations in ϕ prior to inflation, and it is also related to the spectrum of density perturbations, since these are assumed to arise as a result of the differences in the time at which inflation ends. However, these calculations treat the perturbations as departures from a globally homogenous solution to the equations of motion for ϕ, and do not take gravitational effects into account. How this approach is meant to handle the gauge problem is also not clear. Starobinsky's approach lead to a similar conclusion via a different argument: as in the first approach, the time at which the de Sitter stage ends is effectively coordinate dependent (Starobinsky 1982). The source of these differences is not traced to the production of "scalarons" during the de Sitter stage rather than a "time delay" function for the scalar field (see, in particular Starobinsky 1983, p. 303). Finally, Steinhardt and Turner enlisted James Bardeen's assistance in developing a third approach; he had recently formulated a fully gauge invariant formulation for the study of density perturbations (Bardeen 1980). Using

[32] This is often referred to as "horizon exit and re-entry," but the Hubble radius H^{-1} should not be confused with the particle horizon. I will return to this point in Chapter 6; cf. Appendix A.3.

[33] These efforts were later published as (Hawking 1982; Guth and Pi 1982).

Bardeen's formalism, the three aimed to give a full account of the behavior of different modes of the field ϕ as these evolved through the inflationary phase and up to recombination. The physical origin of the spectrum was traced to the qualitative change in behavior as perturbation modes expand past the Hubble radius: they "freeze out" as they cross the horizon, and leave an imprint that depends on the details of the model under consideration.[34]

Here I will not give a more detailed comparison of these three approaches. Despite the conflicting assumptions and different underlying methodology of the three approaches, the participants of the Nuffield workshop apparently lent greater credibility to their conclusions due to the rough agreement they achieved. Guth and Pi defended the various approximations used in their approach (see Blau and Guth 1987, for references and discussion), but the method developed in Bardeen et al. (1983) has been the basis of most subsequent work.[35] During the three weeks of intense collaborative effort at Nuffield these different approaches converged on the following results. In Bardeen et al. (1983)'s notation, the spectrum of density perturbations is related to the field ϕ by:

$$\frac{\delta\rho}{\rho}|_{\lambda} = AH\frac{\Delta\phi}{\dot{\phi}}, \tag{4.6}$$

where $\lambda \approx H^{-1}$, and A is a constant depending on whether the universe is radiation ($A = 4$) or matter ($A = 2/5$) dominated when λ "re-enters" the Hubble radius. The other quantities on the RHS are both evaluated when λ "exits" the Hubble radius: $\Delta\phi$ is the initial quantum fluctuation in ϕ, on the order of $\frac{H}{2\pi}$. The value of $\dot{\phi}$ is given by (from 4.3) $\dot{\phi} \approx \frac{V'(\phi)}{3H}$, and V' depends on the coupling constants appearing in the effective potential. For a Coleman-Weinberg

[34] The sub-Hubble radius modes evolve like the modes of a damped harmonic oscillator, whereas super-Hubble radius modes evolve like an overdamped oscillator.

[35] See, in particular, the comprehensive review of structure formation given in Mukhanov et al. (1992).

effective potential with "natural" coupling constants, $\dot{\phi} < H^2$; plugging this all back into the initial equation we have:

$$\frac{\delta\rho}{\rho}|_\lambda > A\frac{H^2}{2\pi H^2} \approx .1 - 1 \tag{4.7}$$

Inflation naturally leads to an HPZ spectrum, which is also Gaussian (see, e.g., Bardeen et al. 1983). But reducing the magnitude of these perturbations to satisfy observational constraints requires an unnatural choice of coupling constants. In particular, the conflict with observations can be evaded if the term B in the effective potential is very small, which is equivalent to requiring an incredibly small self-coupling for the Higgs field, on the order of 10^{-8}. In the context of simple GUTs, this coupling constant is expected to be on the order of 1, and there is no straightforward modification that leads to such a small coupling constant.[36] New inflation appeared to replace the fine-tuning of the big bang model with fine-tuning of the effective potential for the field driving inflation.

Calculations of the perturbation spectrum culminated in a Pyrrhic victory: a Coleman-Weinberg potential provided a natural mechanism for producing perturbations, but it could be corrected to give the correct amplitude only by abandoning any pretense that the field driving inflation is a Higgs field in $SU(5)$ GUTs. However, it was clear how to develop a newer "new inflation" model; before the conclusion of the conference Bardeen, Steinhardt, and Turner had suggested that the effective potential for a scalar field in a supersymmetric theory (rather than the Higgs field of a GUT) would have the appropriate properties to drive inflation. At roughly the same time Ellis et al. (1982) argued that inflation "cries out for supersymmetry," since a scalar field responsible for supersymmetry breaking would naturally have a potential flatter than the

[36] See Steinhardt and Turner (1984, pp. 2165-2166) for a clear discussion of this constraint, which is also discussed in detail in Kolb and Turner (1990); Linde (1990).

Coleman-Weinberg potential. 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. The introduction of the "inflaton" field, a scalar field custom made to produce an inflationary stage, roughly a year later illustrates this methodological shift.[37] In his recollections of the Nuffield conference, Guth writes:

> [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 1997a, pp. 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. Experiments carried out throughout the early to mid 80s failed to detect proton decay on time scales predicted by the minimal $SU(5)$ GUTs (Blewitt et al. 1985). Models of inflation have been based on a number of theoretical ideas that became popular following the demise of the minimal GUTs.

4.3 The Baroque Era

Following the Nuffield workshop, inflation turned into a "paradigm without a theory," borrowing Mike Turner's phrase, as cosmologists developed a wide variety of models bearing a loose family resemblance. The models share the basic idea that the early universe passed through an inflationary phase, but differ on the nature of the "inflaton" field (or fields) and the form of the effective potential $V(\phi)$. Keith Olive's review of the first decade of inflation ended

[37] Several researchers studied scalar fields with the appropriate properties to drive inflation, but the term seems to have appeared first in Nanopoulos et al. (1983); see Shafi and Vilenkin (1984) for a similar model. I thank Keith Olive for bringing the first paper to my attention.

by bemoaning the ongoing failure of any of these models to renew the strong connection with particle physics achieved in old and new inflation:

> A glaring problem, in my opinion, is our lack of being able to fully integrate inflation into a unification scheme or any scheme having to do with our fundamental understanding of particle physics and gravity. ... An inflaton as an inflaton and nothing else can only be viewed as a toy, not a theory. (Olive 1990, p. 389)

The title of this section derives from Dennis Sciama's complaint (in 1989) that inflation had entered "a Baroque state" as theorists constructed increasingly ornate toy models (Lightman and Brawer 1990, p. 148). The sheer number of versions of inflation is incredible; Guth (1997a, p. 278) counts over 50 models of inflation in the neary 3,000 papers devoted to inflation (from 1981 to 1997), and both numbers have continued to grow. Cosmologists have even complained about the difficulty of christening a new model with an original name, and a partial list of the inflationary menagerie has been used to good effect as comic relief in conference talks.[38]

Inflationary model-building has proceeded with very different programmatic aims. Three broad approaches have characterized the field over the last two decades. What I will call the "new inflation paradigm" aims to eventually re-establish a strong connection between inflationary cosmology and particle physics. The persistent problem is then to identify a fundamental scalar field in a believable model of GUT scale physics, and establish that it "naturally" has the appropriate properties needed to drive inflation. The other two approaches downplay the importance of this link. Since the early 80s Linde has argued that "chaotic inflation" evades the "fine-tuning" problems of new inflation. Later work on "eternal inflation" develops a similar approach: on this approach inflation is a well-defined theory even *without* a link to particle physics, and an (often

[38] See Edward (Rocky) Kolb's talk at the Pritzker Symposium; the slides from his talk are available online at http://www-astro-theory.fnal.gov/Conferences/psw/talks/kolb/.

confusing) combination of probability arguments and invocations of the anthropic principle are taken to provide a sufficient response to the fine-tuning problems of new inflation. Finally, in the 90s a number of authors have pursued a phenomenological approach intended to clarify the links between the effective potential and the spectrum of temperature anisotropies in the CMBR, in preparation for an eliminative program based on satellite observations.

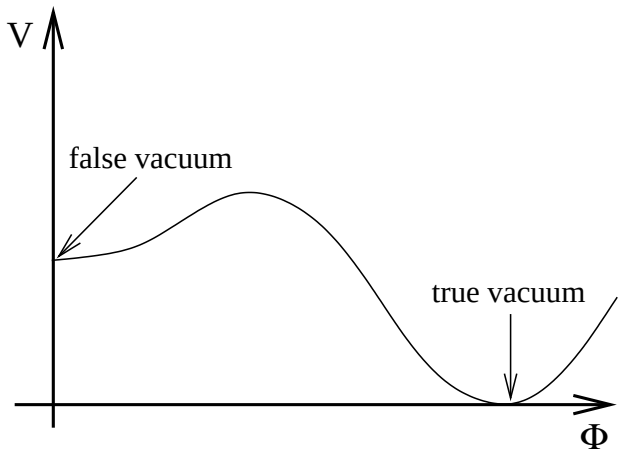

Fig. 4.1 This figure illustrates the effective potential of the Higgs field in original inflationary model.

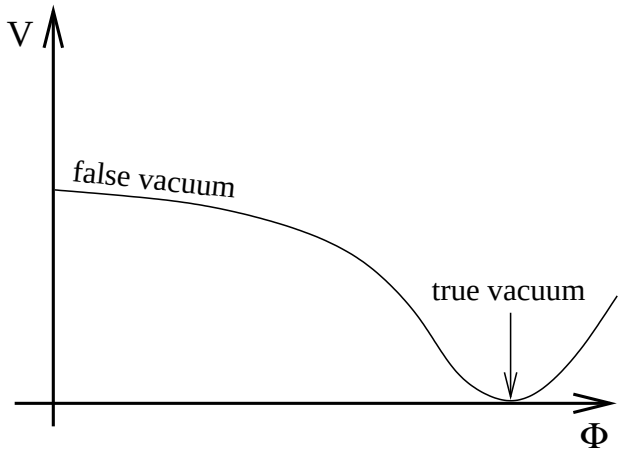

Fig. 4.2 The effective potential of the Higgs field in the new inflation scenarios differs from old inflation in that evolution from the false to true vacuum state need not involve quantum tunneling through a potential barrier. Inflation occurs during the "slow roll" phase down the flat part of $V(\phi)$.

Part II

Philosophy of Cosmology

Chapter 5

Cosmological Principles

From its early days in the 20s and 30s, modern cosmology has faced a number of distinctive methodological problems. In broad terms, these problems reflect the difficulty of using observational evidence from a small patch of the unique universe as a guide in developing and refining cosmological theories. Remarkable progress in cosmology has been accompanied by a sense that these old methodological debates have been laid to rest. In particular, arguments that cosmology differs from other areas of physics have been pushed to one side; one looks in vain for a discussion of the implications of the uniqueness of the universe in many cosmology textbooks (such as Kolb and Turner 1990). The fundamental question underlying the second part of this dissertation is whether it is possible to develop and refine cosmological theories using the same methodology as other areas of physics.

In this chapter I will introduce the methodological issues by focusing on the threat of underdetermination in cosmology and the general principles introduced in response to it. The distinct underdetermination problems can perhaps best be introduced in terms of three inferences. First, cosmologists often aspire to draw conclusions regarding the overall global structure of the universe on the basis of observations restricted to our (relatively small) cosmic neighborhood. The cosmological principle underwrites local to global inferences in that it restricts the allowed space of models to the simple FLRW models, and in these models the local behavior of idealized “fundamental observers” accurately reflects the global shape of space. Second, how

does one determine which observational regularities reflect the laws of nature in cosmology? Part I recounted the importance of the "indifference principle" (using McMullin's phrase) in the development of early universe theories. The indifference principle licenses an inference from the improbability of an observed regularity to the need for novel physics, typically in the form of new dynamics in the early universe. A theory satisfying this principle is "indifferent" to some features of the contingent initial state, which are washed away by subsequent dynamical evolution. Finally, all of the evidence used in cosmology is inherently parochial: we observe the universe from a location suited for our existence. This straightforward point undercuts inferences based on treating our location as somehow "typical," but it is not clear how to appropriately account for this selection effect. Discussions of this worry have gone under the rubric of the anthropic principle, which has been given a number of different formulations.

As with other high-level principles in physics, the epistemic status of these principles is difficult to assess. For their proponents, once clearly formulated these principles are often taken to be self-evidently true, or to be straightforward consequences of the Scientific Method. After characterizing the principles more carefully below, I will argue that they each have a very different status. The cosmological principle (hereafter, CP) falls in with other broad claims regarding the uniformity of nature used to license inductive inferences. As with other uniformity principles, the CP carries epistemic risk, but unlike other cases, the consequences of accepting the CP are remarkably unproductive. I will argue that the indifference principle is on considerably shakier footing, despite its important (and usually implicit) role in current research. The indifference principle is a response to a different, and more fundamental, underdetermination threat, due to uncertainty regarding how to theoretically describe the initial singularity. The "probable initial state" serves as a background template, with important problems defined by the contrast

between the template and observations. But this approach is questionable precisely because it involves strong physical and metaphysical assumptions. The description of the probable initial state is a guess based on the extrapolation of current physics well beyond its domain of applicability. Even if the "template" based on this guess is accurate, the indifference principle makes the additional strong metaphysical claim that all the problems defined by the contrast should be solved by introducing new laws. Finally, my brief discussion of the anthropic principle defends a deflationist account. In the standard anthropic argot, I argue that WAP (the "weak anthropic principle") represents a legitimate concern, but the stronger versions of the principle verge on incoherence.

5.1 Two Approaches to Cosmology

The sciences arguably fall into two camps: *nomothetic* sciences that seek to discover and characterize the fundamental laws of nature, and *descriptive* or special sciences that employ generalizations of limited scope rather than strict laws. Philosophers have frequently argued that cosmology falls into the latter camp, echoing Kant's skepticism regarding the possibility of scientific cosmology.[1] Yet the research described in part I belies this classification: cosmologists have been developing and apparently testing novel physical theories in cosmology for several decades. Here I will briefly argue that the modest view that cosmology is a descriptive science is an unstable position, and identify two forces pushing cosmologists into the more ambitious project of pursuing novel physical laws in cosmology. To formulate these points I will begin

[1] This is one of the perennial topics in philosophy of cosmology; see, e.g., Munitz 1962 and several of the essays in Agazzi and Cordero 1991. My approach here is closest to Torretti (2000)'s, in that I aim to give a richer description of the contrast between mathematical modelling in cosmology and other areas of physics.

with a thumbnail sketch of dynamical theories in physics, and introduce the methodological constraint that these theories should be dynamically and empirically sufficient.[2] Here I will not have space to defend this understanding of dynamical theories; rather my aim is to set the stage by contrasting this picture with cosmology.

The two basic components of a dynamical theory are a space representing dynamically allowed states, a topological space Γ (aka phase space), and a dynamical law that specifies time evolution of the system.[3] The dynamics is given by a map $f : \Gamma \times \mathbb{R} \times ... \rightarrow \Gamma$, which takes a point in Γ to another point representing the state after an elapsed time $t \in \mathbb{R}$. The "$\times$..." may be filled with a number of other real-valued parameters, such as masses of particles or coupling constants for interactions. Often Γ is a differentiable manifold, with the dynamics described by a set of differential equations governing the evolution of a set of fields defined on the manifold. For example, the phase space for a collection of n point particles moving in three-dimensional Euclidean space is a $6n$-dimensional manifold (three position and three momentum coordinates for each particle, q^i and p_i respectively). The state of the system is picked out by a point in this phase space, $\omega = (q, p) = (q^1, ..., q^{3n}, p_1, ..., p_{3n}) \in \Gamma$, and Hamilton's equations specify the dynamical evolution of the system:

$$\dot{q^i} = \frac{\partial H}{\partial p_i}, \qquad \dot{p_i} = -\frac{\partial H}{\partial q^i}, \tag{5.1}$$

[2] The terminology is borrowed from Lewontin (1974)'s clear discussion of the difficulties faced in population genetics, which I think George Smith for bringing to my attention. The discussion also draws on Laura Ruetsche's approach to interpreting theories; see Ruetsche (2002) for a concise presentation.

[3] At a minimum Γ is a topological space, with the idea of proximity provided by the topology roughly corresponding to physical similarity between states, but it is often provided with additional mathematical structure such as a symplectic form or a coordinate chart.

where H is the (usually time independent) total energy of the system. Solutions of these equations specify dynamical trajectories through phase space representing the evolution of the particles' position and momentum over time. A complete specification of the state of the system includes these time-dependent quantities (called dynamical variables) along with any other properties (such as masses of the particles) that are time-independent. A model for the theory consists of a specification of the function H along with the initial or boundary conditions of the dynamical variables. The semantics of the theory clarifies the truth conditions for the theory by identifying the set of possible worlds that serve as models of the theory. In our simple example, the semantics specify how to relate claims about observable properties of a physical system to regions of phase space. The total kinetic energy of a system of particles, for example, is represented as $T = \sum_{i=1}^{n} \frac{\mathbf{p}_i^2}{2m}$ for n particles. This is simply a function from phase space to real numbers: $f_T : \Gamma \to \mathbb{R}$. The experimental question "Is the total kinetic energy within the range Δ?" will receive a positive (negative) answer for a state ω such that $f_T(\omega) \in \Delta$ (respectively, $f_T(\omega) \notin \Delta$).

In a dynamically sufficient theory the state space and dynamical laws are well-matched, in the sense that the theory admits a well-posed initial value formulation. The first requirement of such a formulation is that fully specifying the dynamical variables at an initial time singles out a unique trajectory through phase space. Second, the system should be relatively insensitive to small changes in the dynamical variables: the bundle of trajectories passing through an open

set around a point in Γ should not "spread out" in phase space too quickly.[4] For systems satisfying these two requirements, measurements compatible with some subset of Γ combined with the dynamics governing the system yield various predictions regarding the system's future state. Practical limitations in measuring a system's state combined with the magnification of errors under dynamical evolution limit the theory's predictive power. The standards of predictive accuracy depend upon the context, but often it is sufficient to locate a system's state within a subset of the phase space such that answers to all the experimental questions of interest fall within some acceptable range Δ.

The ambition to include *all* of the physical degrees of freedom for a given system obviously runs up against practical concerns with developing a tractable theory as well as the daunting complexity of real systems. The compromise—simple, idealized models—results from dropping features that are negligible at the desired level of accuracy, or arguing that various features not incorporated in the model *should* be negligible. The nature of idealizations is a contentious topic in philosophy of science, but I assert that (at least in some cases) physicists have "control" over the idealizations involved in applying a theory in the following sense. Hempel (1988) argued that deriving predictions from a theory typically requires a "proviso" that the theory *is* complete, even when it is acknowledged that the theory offers only a partial description of the phenomena within its domain. The application of Newtonian gravitational theory to planetary motion, for example, requires a proviso stating that the only forces acting on the bodies

[4] Filling in the details of what counts as well-posed requires considering specifi c theories. But two further points are worth noting. Earman (1986) emphasized that contrary to the conventional wisdom many theories of classical physics admit a well-posed initial value formulation only if various supplementary conditions are imposed. Second, for relativistic theories a well posed formulation is also required to satisfy a constraint on causal propagation: changes to initial data associated with some spacetime region Σ should not affect the solution in regions that are causally inaccessible from it (i.e., regions outside of $J^{\pm}(\Sigma)$, in the language of Appendix A.4).

are those due to their mutual gravitational attraction, ruling out electromagnetic, vegetative, angelic ... forces. Treating the theory as a complete description in this manner allows one to draw strict inferences regarding planetary motion (based on a differential equation derived from the laws of motion), whereas without such a proviso the theory literally has no implications for future planetary motions with given initial positions.[5] The further advantage of this "control" is that discrepancies between the conclusions of theoretical inferences and observational or experimental results are particularly informative in developing successively more complicated models. Returning to the example of planetary motion, discrepancies between the simple two-body case and more complicated motions were used to refine the theory by (among other things) including perturbing forces from the other planets. On this account, physical theories aim not to capture the full complexity of real systems all at once, so to speak, but rather to develop a series of increasingly sophisticated models, in the process developing a more detailed and informative body of evidence.[6] The long term success of the theory hinges not on capturing the full complexity of a system at the first stroke, but rather on giving careful attention to the nature of the provisos needed to make exact inferences and successfully accounting for discrepancies that arise without abandoning the framework provided by the theory.

---

[5]Hempel argues that provisos of this sort must be stated in the language of the theory at hand (V_C) rather than the antecedently available vocabulary V_A, and notes that his claim regarding provisos is stronger than the Quine-Duhem underdetermination thesis, in that *no* set of auxiliary hypotheses formulated in V_A can play the role of a proviso. The provisos play a crucial role in applying theory: within the subset of models picked out by a proviso ruling out all but gravitational forces, the machinery of Newtonian mechanics links logically contingent propositions (formulated in V_A) describing an initial matter distribution to future motions. See Earman and Roberts (1999) for a clear discussion of Hempel that carefully distinguishes his position from more recent worries (sometimes claimed to derive from Hempel's paper) about ceteris paribus clauses.

[6]The clearest statement of this methodology is given by George Smith, in a sophisticated account of Newton's methodology in the *Principia*, Smith (2002a,b); I have also benefi tted from discussions with him on this topic.

To qualify as empirically sufficient the theory's predictions must pass muster by falling within some acceptable error range of observational results. But at a deeper level the parameters appearing in the evolution laws and the dynamical variables must be measurable, even though measurements of them typically are very indirect and theory-mediated. Some observational evidence related to a dynamical theory serves a primarily "diagnostic" role, in that it must be used to constrain the various parameters occurring in the theory. But ideally the theory generates a wide range of predictions once these values have been set. Otherwise one runs the risk that the theory employs sufficient degrees of freedom to model the phenomena in question regardless of the accuracy of the dynamical theory. (I will return to this issue in Chapter 7.) Furthermore, various different phenomena covered by the theory can be taken as independent diagnostics for the parameters occuring in the theory, and achieving convergent measures of the parameters of the theory by a variety of independent methods is in itself an important component of empirical success.[7]

Finally coming back to the discussion of cosmology, in the modest or descriptive approach the laws of cosmology are given by extrapolations of dynamical theories developed in fundamental physics, such as GTR and the Standard Model of particle physics. The goal is the development of a consistent cosmological model based on these ideas with no attempt to justify the underlying physical theories, although the cosmological domain may provide a weak consistency check on the tremendous extrapolations involved in applying these theories. In the terms used above, judging the dynamical and empirical sufficiency of these underlying theories is left aside, and the project of developing a consistent and detailed cosmological account resembles

[7]This point is also taken from studies of Newtonian methodology, see, e.g., Harper (1997) and references therein.

research in phenomenological physics or the special sciences rather than fundamental physics. As in other special sciences, cosmologists may discover various generalizations, such as regularities regarding the formation and evolution of galaxies. Although these general claims may draw on a wide variety of physical theories, they will be contingent in that they do not hold true in all the possible models admitted by the underlying fundamental theory. On this view the attempt to develop novel physical theories in cosmology would be as misguided as, say, attempts to develop new fundamental physics that accounts for all aspects of the Krebs cycle. Although the laws of physics are clearly relevant in both cases, some aspects of the observed regularities reflect the consequences of various contingencies rather than features of the laws.[8]

This brief sketch brings out how little the "modest approach" resembles the practice of modern physical cosmology described in Part I. The modest approach more aptly describes Sandage's observational program to determine the best fit FLRW model. Sandage's two numbers—the Hubble constant H and the deceleration parameter q—are both features of a model describing the universe as a whole, defined in terms of the scale factor $a(t)$ and its first and second derivatives (as is another important parameter, the critical density ρ_c):[9]

$$H(t) =: \frac{\dot{a}}{a}, \qquad q(t) =: -\frac{\ddot{a}a}{\dot{a}^2}, \qquad \rho_c =: \frac{3\dot{a}^2}{8\pi G\, a^2} \tag{5.2}$$

The observerational repertoire of Sandage and his colleagues consisted of different ways of indirectly measuring the scale factor $a(t)$, the fundamental dynamical variable in the FLRW

[8] See also Beatty (1995)'s argument for the "evolutionary contingency thesis," namely that there are no "distinctively biological" laws governing the contingent products of evolution (such as the details of cellular function), although there may be relevant lawlike regularities derived from physics and chemistry.

[9] For the FLRW models, $q_0 = \Omega_0/2$. G is Newton's constant, and $\dot{a} = da/dt$, where t is the time coordinate in the FLRW line element.

models. Yet the scale factor is a well-defined global quantity *only* in the FLRW models. The scale factor characterizes the time variation of the distance between "fundamental observers," who are by definition at rest with respect to the spatial coordinates of the FLRW line element (eqn. A.7). The symmetries of the models also fix a preferred cosmic time, corresponding to the proper time measured by clocks carried by the fundamental observers. The definition of $a(t)$ takes full advantage of the symmetries of the FLRW models. In more general solutions lacking the unrelenting symmetry of the FLRW models the "global" scale factor is replaced by a quantity characterizing only local expansion rates (see Appendix A.2 for a brief discussion).

This brings out the first motivation for moving beyond a modest approach: justifying the idealization employed in the FLRW models. In the study of planetary motion, Newton's gravitational theory provided the tools needed to build a complex model starting from the simple and well understood idealization of the two body problem. By contrast, in the case of the FLRW models the idealization is justified not by utilizing a proviso to the effect that the theory is complete, but in terms of approximate compatibility with observations and/or a stipulation that the overall matter distribution in the universe can be treated as a perfect fluid. Cosmologists have found this situation unsatisfactory, particularly since the symmetry of the FLRW models is suggestive of some deeper underlying physical principle. The development of inflation illustrates the advantage of a physical jusitification for the FLRW models: the deviations from the exact symmetry of the FLRW models have become the primary source of further evidence for inflationary theory. Without some understanding of why the nearly exact symmetry of the FLRW models obtains, it is unclear whether the small nonuniformities in the CMBR have any physical import. The second force pushing cosmologists away from a modest approach is that in

the standard big bang model, the early universe reaches arbitrarily high energies as $t \to 0$, and it thus serves as a natural testing ground for new ideas in particles physics.

5.2 Global Aspects of Cosmology

Classical mechanics and other areas of physics typically describe the properties and evolution of local systems. One of the novel aspects of relativistic cosmology is the introduction of global features of spacetime models that cannot be reduced to local properties. Earlier discussions of the global aspects of cosmology focused on the "logic of cosmology" (the title of Munitz 1962): Harré (1962), for instance, classifies two ways of arriving at the global claims characteristic of cosmology. Cosmologists can either ratchet up the concepts of local physics by "type elevation," characterized as application of predicates (like entropy) defined for members of a class – isolated subsystems – to the class as a whole, namely the whole Universe (conceived of as a single object comprising all that there is). On the other hand, global claims may be arrived at by indefinite generalization of claims regarding our cosmic neighborhood to the entire Universe. Harré uses this distinction to draw a line of demarcation: type elevation leads to "unintelligible propositions," and cosmologists must stick to indefinite generalizations if their theories are to be scientific.[10] Worries about global properties have not been confined to philosophers attempting to adjudicate cosmological debates from their armchairs; recently Smolin has argued that a cosmological theory must satisfy the following principle:[11]

[10] Harr´e argued in support of the steady state theory: "cosmogonic" theories incorporating a creation event were on the wrong side of this line, whereas postulating new local physical processes (such as the creation of matter) was an acceptable part of "cosmophysics."

[11] He takes this to be a straightforward consequence of the standard epistemology and methodology of dynamical theories, "[A] theory of cosmology must be falsifiable in the usual way that ordinary classical and quantum theories are. This leads to the requirement that a sufficient number of observables can be determined by information that reaches a real observer inside the universe to determine either the classical history or quantum state of the universe" (p. 23). Smolin also introduces a second principle – "Every

> Every quantity in a cosmological theory that is formally an observable should in fact be measurable by some observer inside the universe. (Smolin 2000, p. 3)

Except in unusual cases such as a spacetime with compact spatial sections, an observer "inside" the spacetime will see a finite region, from which it is impossible to measure a global property (as I will discuss in more detail in section 5.3). Qualms such as these about whether global property ascriptions are compatible with a broadly empiricist methodology should come as no surprise.[12] What is surprising, however, is that in GTR a leap to the global is neccessary to define several fundamental quantities and to characterize the causal structure of spacetime. Below I will briefly review two cases that illustrate what is gained by "going global" in general relativity.

One of the remarkable shifts in moving from special to general relativity is the loss of a straightforward local definition of energy and its associated conservation law. (I have relegated a slightly more detailed discussion of this topic to Appendix A.6.) Consider, for example, a binary pulsar releasing gravitational waves as the pulsars' orbits shrink. Intuitively the system is losing gravitational energy, which is carried off by gravitational waves, and one might expect conservation laws similar to those in classical or special relativistic physics to bolster these intuitions. However, in GTR a gap opens up between integral and differential formulations of conservation laws. Differential conservation laws guarantee that energy flows continuously through individual spacetime points. In special relativity the background inertial structure can be used to turn a differential law into an integral conservation law, which describes the flow of energy through finite regions. Integral conservation laws underwrite the intuition appealed to above, namely

---

formal mathematical object should be constructible in a finite amount of time by a computer which is a subsytem of the actual universe." Unfortunately Smolin does not give a more detailed formulation of the principle, and it is not clear to me that one *can* formulate a criterion of constructibility that draws the line where Smolin does: with much of modern mathematics qualifying as acceptable, but with constructions of the configuration space for some systems beyond the pale.

[12] Sklar (2000, pp. 24-32) sees a "retreat to the local" in both Einstein's formulation of GTR and the development of the local algebraic formulation of quantum field theory.

that the gravitational energy lost by the binary pulsar is carried away by gravitational radiation. Although GTR incorporates a differential conservation law, $\nabla^a T_{ab} = 0$, this can only be turned into an integral conservation law in stationary solutions, a special class of solutions possessing a time translation symmetry. This failure reflects the fundamental difference between gravitational energy and the energy-momentum carried by matter fields: the stress-energy tensor for the latter gives a fully covariant description of the energetic quantities, but the equivalence principle guarantees that *locally* one can always choose coordinates to "transform away" the gravitational field.[13]

A definition of energy can be recovered *globally* for some spacetimes; intuitively, "far away" from the binary pulsars and the complicated geometry they produce, the spacetime approaches Minkowski spacetime, which possesses the necessary symmetries to recover an integral conservation law. Since the early 60s physicists have defined mass, energy, and momentum in terms of the asymptotic structure for such asymptotically flat spacetimes. More recently interest has also focused on *quasilocal* energy, which is the energy associated with a compact two-surface (see Appendix A.6). The global approach has led to a number of important theorems in classical general relativity, such as the positive mass theorem (i.e., that isolated gravitational systems have non-negative total energy) and arguments that gravitational waves extract energy from radiating systems (see, e.g. Wald 1984, §11.2).[14]

[13] Standard textbook treatments of GTR often introduce t_{ab}, a quantity representing the stress-energy tensor of the gravitational field, but this quantity differs from T_{ab} in that it is only a (coordinate-dependent) "pseudo-tensor."

[14] Hoefer (2000) argues that the lack of a local definition of energy and its conservation law blocks a substantivalist move: the substantivalist cannot criticize the relationist's apparent inability to attribute energy and momentum to empty spacetime, since she cannot herself give suitable definitions of these quantities. Whether or not this is actually a good argument, unlike Hoefer I think that existing definitions of quasi-local and global energy do provide sufficient grounds for the substantivalist. In the case of the binary pulsar, the global definitions *do* license claims regarding the energy carried away by gravitational radiation, which I would argue is all the substantivalist needs.

A second case more clearly illustrates the importance of global properties of spacetime in general relativity. As discussed in more detail in Chapter 2, starting in the 60s relativists developed a number of new tools to study spacetime singularities and determine whether they occur in generic spacetimes or are an artifact of strong symmetry assumptions. Even defining what one means by a "spacetime singularity" has turned out to be an incredibly intricate conceptual and technical issue that has yet to be fully resolved.[15] I will focus on one aspect of these difficultes, namely the question of whether singularities can be analyzed locally as a property of a specific spacetime region.

Physicists have widely adopted the view that incomplete geodesics signal the presence of a singularity. The path from the intuitive notion of a singularity as a "blow up" of some field quantities to their definition in terms of geodesic incompleteness runs roughly as follows. In classical and special relativistic physics, one can meaningfully speak of singularities of a solution to the appropriate field equations as occuring at a particular point—if, for example, field quantities increase without bound on curves approaching the point. In GTR, one can construct a number of scalar quantities from the Riemann curvature tensor and use the behavior of these quantities as a signal of singular behavior. In the classical case singular behavior can be located against the fixed background spacetime, but since we are interested in singularities of the gravitational field itself in GTR, the "blow-up" of the curvature invariants cannot be directly used to "locate" the singular points. One typically assumes that spacetime is modeled by a differentiable manifold M equipped with a metric g_{ab} defined and differentiable everywhere on the manifold; *ex hypothesi* there are no singular points in M. One might still hope to use bad behavior on the part of the curvature invariants as a criterion for spacetime pathology. But this proposal, even when the

[15] Earman (1995), Chapter 2 gives the most comprehensive discussion of these issues; cf. Curiel (1999).

details have been filled in, faces a number of objections (see, e.g., Wald 1984; Earman 1995, for discussion). The most telling is that this criterion fails to capture a broad class of singular spacetimes in which an observer moving along a geodesic does not encounter unbounded increase of curvature invariants or other pathologies, yet is still "snuffed out of existence" within a finite proper time.[16] Identifying singularities via geodesic incompleteness captures these cases, and perhaps more importantly this identification proved to be fruitful: the seminal singularity theorems of Penrose, Hawking, and Geroch demonstrate that geodesic completeness is incompatible with a number of other plausible, physically motivated assumptions.

Intuitively an incomplete geodesic corresponds to a "missing point"; spacetime unnaturally runs out before reaching it. This intuition can be made precise for a manifold M equipped with a Riemannian metric, a non-degenerate, symmetric tensor h_{ab} that is positive definite. A compact manifold includes all the points that it possibly can, in the sense that, roughly speaking, it contains all the limit points of sequences in the set and it is "small" in an appropriate sense.[17] For a space with a Reimannian metric there is a clear link between geodesic incompleteness and "missing points" provided by the notion of a Cauchy sequence. A Cauchy sequence is a set of points p_i such that for any given positive ϵ, $\exists I(\forall j,k > I : d(p_j, p_k) < \epsilon)$, where d is the distance function obtained from h_{ab}. If every Cauchy sequence converges to some $p \in M$, the space is Cauchy complete, and also compact; moreover, for the Riemannian case a theorem guarantees that a Cauchy *incomplete* space has incomplete geodesics. Missing points can be naturally added to the space via a "boundary construction," provided by an isometric imbedding of the Cauchy

[16] See Wald (1984, p. 214) for a clear example, a "conical singularity" constructed from a wedge in Minkowski spacetime such that $R_{abcd} = 0$ everywhere even though curves terminating on the "vertex" of the cone are incomplete.

[17] More precisely, a topological space is compact iff every collection of open sets whose union coincides with the space itself has a finite subcollection that also coincides with the space. See, e.g., Geroch (1985), for definitions and for a clear discussion of compactness (in §30).

incomplete space into a complete space. The boundary points in the complete space correspond to equivalence classes of non-convergent Cauchy sequences in the original space that (roughly speaking) approach the boundary point.

This nice correspondence between incomplete geodesics and "missing points" breaks down for a pseudo-Reimannian metric (as in GTR), since zero-length null curves confound any attempt to define a positive distance function (and Cauchy sequences). In the spacetime case, an incomplete geodesic is a curve that is inextendible in one direction with finite affine length.[18] Misner (1963)'s "counter-example to almost everything" nicely illustrates that the connection between geodesic completeness and compactness doesn't carry over to relativistic spacetimes: the general-purpose counter-example is a compact solution that nonetheless contains incomplete geodesics. One might still try to introduce boundary points by analogy with the Riemannian case, as equivalence classes of incomplete geodesics (see Appendix A.4 for a brief discussion and references). The appeal of such boundary constructions is that they would allow for "local" analysis of singular structure, similar to the local analysis of isolated systems achieved by the conformal completions of asymptotically flat spacetimes. There are a wide variety of different boundary point constructions (a-boundaries, b-boundaries, and g-boundaries, to name a few), relying on different definitions of incompleteness and in some cases (such as the a-boundary of Scott and Szekeres 1994) substantial new formalism.

[18] "Affine length" is a generalization of elapsed proper time to include null and spacelike geodesics; an affine parametrization s of a curve is one such that $u^a = (\frac{\partial}{\partial s})^a$ satisfies the geodesic equation, $u^a \nabla_a u^b = 0$. The limit to geodesics is motivated by examples such as Born-accelerated motion in Minkowski spacetime: the curve for such motion has finite proper time, but it would be ridiculous to brand Minkowski spacetime singular as a consequence. The restriction to geodesics can be eased by introducing "generalized affine length" and the corresponding notion of "b-completeness"; see, e.g., Hawking and Ellis (1973, pp. 258-261) for further discussion. A curve $\gamma(s)$ is inextendible iff it does not possess an endpoint, a point p such that for every neighborhood O of p there exists a parameter value s_0 such that the image of $\gamma(s)$ in M remains in O for all $s > s_0$.

These boundary completions are the closest one comes to a local characterization of spacetime singularities. However, they all have a number of counterintuitive consequences for some of the cases to which they have been applied. For example, Schmidt's b-boundary for the FLRW models represents *both* the initial and final singularities as a single point that is not Hausdorff separated from any point in M (see Clarke 1993, pp. 40-45). Geroch et al. (1982) conclude a discussion of similar counterintuitive consequences with the following remark: "Perhaps the localization of singular behavior will go the way of 'simultaneity' and 'gravitational force'." Even a retreat to identifying singularities with curvature blow-up does not preserve a "local" characterization of singularities: whether a region contains curvature blow-ups depends upon which curve one considers.[19] Following Geroch et al. (1982)'s advice, we should construe "singular" as an adjective characterizing the global structure of a spacetime rather than as a property of a particular region.

Claims about causal "good behavior" of spacetimes are also characteristically global. Specifying the causal structure of spacetimes precisely is one of the crucial components of the singularity theorems. There are a number of causality conditions that relativistic spacetimes can satisfy, which can be roughly characterized as specifying the extent to which various causal features characteristic of Minkowksi spacetime hold globally (see Appendix A.4). For example, a globally hyperbolic spacetime possesses a Cauchy surface, a spacelike surface Σ intersected exactly once by every inextendible null or timelike curve. This is properly understood as a global property of the entire spacetime; although submanifolds of a given spacetime may be

[19] See, in particular Hawking and Ellis (1973), pp. 290-292 for a brief discussion; this point is also mentioned by Curiel (1999).

compatible or incompatible with global hyperbolicity, it cannot be directly treated as a property of local regions which is then "added up" to deliver a global property.

5.3 The Cosmological Principle

The status of the CP has been a subject of ongoing debate since its formulation in the early days of relativistic cosmology. The content of the principle is no longer the focus of active debate: it requires that a cosmological model is homogeneous and isotropic (see §1.2 and Appendix A.2).[20] But assessments of the CP run the gamut, from the position that it qualifies as an *a priori* truth to Ellis's characterization of it as an "unverifiable" principle. Before turning to this question, in 5.3.1 I characterize the CP as the strongest of a number of ampliative principles that make local to global inferences possible. In light of these results, I will argue in 5.3.2 that the CP and similar weaker principles play the role of a general principle supporting eliminative induction. I will further argue that unlike other inductive generalizations in physics, surprisingly little is gained by taking on the additional epistemic risk associated with such ampliative principles.

5.3.1 Underdetermination

The CP puts stringent limits on the space of allowed cosmological models, and as a consequence reduces the threat of theoretical underdetermination. In cosmology this threat can be precisely formulated; it results from a combination of two features of general relativity. First, the field equations of relativity specify a *local* relation between the various tensors that appear in EFE, but this local relation is compatible with a wide variety of global structures. Second, due

[20] The CP is sometimes taken to require only spatial homogeneity (with isotropy presumably guaranteed by CMBR observations), and it is also sometimes confused with the weaker Copernican principle introduced below. These differences in usage do not reflect deep disagreements, although there certainly were important differences among early formulations of the principle.

to the finite maximum signal speed, an observer taken to be located at a point $p \in M$ is in causal contact with only the region of spacetime marked out by the causal past $J^-(p)$.[21] The physical state at points outside of $J^-(p)$ is not fixed by observations on $J^-(p)$ in conjunction with the laws of physics (*modulo* a few caveats, discussed in section 6.2). Even fully specifying the state on $J^-(p)$ places few constraints on the global features of spacetime, in the sense that it can be embedded in a spacetime M', g'_{ab} with different global features than the original spacetime M, g_{ab}. This is the idea behind Glymour's definition of "observational indinstinguishability" (OI): if $I^-(p)$ can be embedded in M', our observer at p would have no observational grounds to claim that she is in M, g_{ab} rather than its indistinguishable counterpart M', g'_{ab}. Any global features that are not invariant under the relation of OI cannot be observationally established by our idealized observer at p. Thus the question of observationally establishing global features of spacetime can be translated into a more precise "topological" question: what constraints are imposed on M, g_{ab} by the requirement that a collection of sets $I^-(p)$ can be isometrically embedded in it? Here I will focus on clarifying the scope of OI given different assumptions regarding the space of allowed counterparts.[22] At the lowest level—only imposing this "embedding" requirement—very little can be said about the global structure of spacetime based on observations confined to $J^-(p)$. As we will see, adding stronger physical and symmetry constraints leads to stronger local-to-global inferences (see Table 5.1).

[21] In Minkowski spacetime, this set is the past lobe of the light cone at p, including interior points and the point p itself. In the discussion below I will shift to using $I^-(p)$, the chronological past (in Minkowski space, the interior of the past lobe) for convenience, since these are always open sets. Nothing is lost since $J^-(r)$ is a subset of $I^-(p)$ if $r \in I^-(p)$, except in the case of maximal timelike curves with future endpoints. See Appendix A.4 for a brief review of the relevant definitions.

[22] My approach here is also informed by the so-called "observational cosmology" program pursued by Ellis, Stoeger, and various collaborators, with the stated aim of understanding to what extent observational evidence can or cannot justify the widely accepted FLRW models. See Matravers et al. (1995) for a recent update on some results of this program.

Table 5.1. Hierarchy of Constraints

Constraint	Scope of Underdetermination
Topological	
Embedding of $J^-(p)$ sets	Global structure underdetermined
Physical	
Stipulated form for T_{ab}, initial data	Case by case "no go" results
Symmetry	
(Almost) Isotropy & Homogeneity	Determine best fit (Almost) FLRW model

Here I will not consider a more ambitious program that would aim to eliminate rival cosmological theories in favor of relativistic cosmology. Thorne, Will and others have explored the space of possible gravitational theories that satisfy several general requirements, and have concluded that observations can be used to rule out large sections of the space of possible alternative theories (see Will 2001, for a recent review). But the application of GTR to cosmology extends far beyond the range of the solar system-scale tests used in this eliminative project. In addition, applying gravitational theory at larger scales leads to the dark matter problem: gravitational measures of mass-energy density give substantially higher estimates than other observations.[23] The typical response has been to exploit the flexibility of auxiliary assumptions regarding the matter distribution. GTR by itself places few constraints on the source term T_{ab}, leaving cosmologists

[23] I have benefited from discussions with Bill Vanderburgh regarding these issues; see Vanderburgh (2001). It is worth noting that the "dark matter" problem has two variants which differ in scale and in the nature of attempted solutions: first, a discrepancy between different mass measures applied to galaxies and clusters of galaxies—Vanderburgh calls this the "dynamical dark matter problem," which has been "solved" via the introduction of various dark matter candidates. There is also a "cosmological" dark matter problem, indicated by the (much greater) discrepancy between the value $\Omega = 1$ preferred by many cosmologists and the total contribution of baryonic matter to Ω (*including* the dark matter needed to solve the first problem); currently the most popular solution to this problem is to introduce "dark energy" due to a new fundamental scalar field distinct from the inflaton. The study of galaxies relies entirely on Newtonian gravitational theory, and thus the conflict with GTR is indirect; one has to assume that GTR reproduces Newtonian gravitational theory in the weak-field, low-velocity limit for a mass distribution like that of a galaxy.

with the option of avoiding the discrepancy by introducing new source terms (such as that for "quintessence" and/or various dark matter candidates) without modifying gravitational theory. Contrast this case with the need to modify Newtonian gravitational theory based on observations of the solar system: Seeliger's zodiacal dust cloud and Dicke's solar oblateness were both introduced as ways to account for Mercury's small anomalous perihelion motion within Newtonian gravitational theory. Evidence regarding the solar system is rich enough to cast doubt on these proposed modifications of the mass distribution without begging any questions regarding the status of Newtonian gravitation. We clearly lack similarly robust, independent evidential constraints on the matter distribution in galaxies. However, there is enough discomfort at introducing otherwise undetected dark matter to motivate the development of alternative theories of gravitation that resolve the mass discrepancy without new types of matter (see, e.g., Sanders and McGaugh 2002; Mannheim 2000).[24] These theories have been extended to the cosmological regime, and their proponents have been able to recover several features of standard big bang cosmology. Considering these alternatives in detail would take me too far afield, and here I will focus on assessing the difficulty in a choosing a particular model on the assumption that GTR applies.

The modest goal of pinning down the geometry of $J^-(p)$ observationally can be realized, at least for "idealized" observers (as Ellis 1980 describes with remarkable clarity). The relevant evidence comes from two sources: the radiation emitted by distant objects reaching us along

[24] As far as I know, *no* version of MOdified Newtonian Dynamics (MOND), originally introduced by Milgrom and developed recently by McGaugh, has been formulated that would pass the first test in Thorne and Will's program: MOND is a modification of Newtonian gravitation in the low acceleration regime, and does not treat the gravitational field as a tensor field satisfying generally covariant field equations (as Thorne and Will require). In Mannheim's theory, conformal symmetry is imposed within a metric theory, in effect replacing the Ricci scalar in the Einstein-Hilbert action with a "Weyl scalar" constructed from the (conformal) Weyl curvature tensor, and also treating the gravitational constant G as a dimensionless coupling constant.

our null cone, and evidence, such as geophysical data, gathered from along our world line, so to speak. Considering only the former, suppose that astronomers somehow have full access to ideal observational evidence: comprehensive data on a set of "standard objects" scattered throughout the universe, with known intrinsic size, shape, mass and luminosity. With these data in hand one could study the distortion or focusing effects of the standard objects as well as their proper motion. Suppose that observers report no distortion or focusing effects and no proper motions — could they then conclude that the observable universe is isotropic around the observer? Not without assuming some background dynamics, such as EFE with a particular equation of state. But coupled with fixed dynamics the ideal observational data are sufficient to determine the spacetime geometry of the observer's null cone, $J^-(p)$, as well as the matter distribution and its velocity.[25] Thus in principle one could observationally establish isotropy; in practice, the small observed temperature variations of the CMBR (once the dipole moment corresponding to our proper motion is subtracted) provide the best evidence for isotropy. Numerous practical limitations on astronomical observations make it extremely difficult to actually measure the various quantities included in the ideal data set. The idealization appealed to above sidesteps one of the most pressing sources of systematic error in interpreting observations: differentiating evolutionary effects on the objects used as "standard candles" (such as galaxies or supernovae) from cosmological effects. In any case, the difficulties with actually determining the geometry of $J^-(p)$ using real astronomical data differ in kind from the limitations on claims regarding global structure discussed below.

[25] As Ellis notes, the metric quantities that determine how the null cone is embedded in the spacetime cannot be directly measured without using the dynamical equations, but the distortion and focusing effects of standard objects can be used to directly measure the intrinsic geometry of the null cone.

The underdetermination at issue differs from two other types of underdetermination discussed in the philosophy of science literature. Absolute velocity in Newtonian mechanics is one of the stock examples of underdetermination in the literature (see, for example, van Fraassen 1980, pp. 44-47, and Laudan and Leplin 1991, pp. 457-58). The background theory of Newtonian mechanics *rules out* the possibility of gathering evidence that could decide between different models embodying different choices of a preferred inertial frame of substantival space. Since absolute velocity is defined as the velocity with respect to the preferred inertial frame, the distinction between models with different absolute velocities can be drawn *only* at the level of theory. On the other hand, in the cosmological case a given cosmological model and an indistinguishable counterpart *do* in fact differ in observational content, and describe quite different overall spacetime structures (cf. Bain 1998, §3.1). OI also differs from indistinguishability that arises from changing how objects or events are identified (as originally suggested by Reichenbach 1958). As an example, consider Minkowski spacetime in which all physical fields return to the same state periodically, on the time slices $t = 0, k, 2k, ...$[26] This spacetime is "indistinguishable" from Minkowski spacetime which is "rolled up" along the time axis, such that the periodic return to the same physical state is actually a return to the numerically identical $t = 0$ slice. A long-lived astronaut traveling along a future-directed timelike curve will repeatedly cross these identical slices in either spacetime, but she can either interpret the dull repetitiveness of her spacetime as evidence for periodic behavior in an open time or as a sign of cyclic time.[27] This interpretative move (treating the time slice as a single slice or as repeated copies of the same

[26] See also Glymour (1972), as well as Weingard (1990)'s brief critical response to Harr´e (1986, pp. 140-41); Harr´e seems to be taking his cue from Reichenbach, although this debt is not acknowledged.

[27] This example only works with the strong assumption that the astronaut's mental states supervene on the physical states, so that the astronaut will have *exactly* the same mental state each time she crosses the slice. So although she may be aware of the dull repetitiveness, the astronaut cannot count the number of times she has returned to the same state.

state) is left open even for observed regions of spacetime. By way of contrast, OI relies on the existence of a *terra incognita* hidden beyond observational horizons. A suprisingly wide variety of spacetimes have OI counterparts, whereas the type of indistinguishability described above does not generalize to more interesting cases.[28] The case of OI spacetimes cannot be given a conventionalist gloss: OI counterparts describe *different* spacetimes, which happen to agree in certain patches—without any trickery involving the identification of events.

Turning now to the definition of OI, the intuitive requirement that all observers' past light cones are compatible with two different spacetimes can be formalized as follows (Malament 1977, p. 68):[29]

> *Weak Observational Indistinguishability*: Cosmological models $\langle M, g_{ab}, O_1, ..., O_n \rangle$ and $\langle M', g'_{ab}, O'_1, ..., O'_n \rangle$ are WOI if for every $p \in M$ there is a $p' \in M'$ such that: (i) there exists an isometry ϕ mapping $I^-(p)$ into $I^-(p')$, (ii) $\phi^* O_i = O'_i$ for $i = 1, ..., n$.

The adjective "weak" distinguishes this formulation from Glymour (1972, 1977)'s original, which was cast in terms of inextendible timelike curves and stipulated that the relation is symmetric. I agree with Malament's argument that these features of the original definition fail to capture the epistemic situation of observers in cosmology. First, if observers are idealized as inextendible timelike curves, whether or not a given spacetime has OI counterparts depends upon

[28] In the example above the "unrolled" Minkowski space is the covering space of the rolled up version. The covering space is obtained by "unwinding" all noncontractible curves (see, e.g., Geroch 1967). The example also requires that the fields populating the spacetime return to the same state periodically, which is difficult to arrange in more reasonable models (see Tipler 1980, for a "no-recurrence" theorem for spatially closed models).

[29] My definition differs slightly from that given by Malament, in that I am requiring that the source fields (rather than only the T_{ab}) are diffeomorphic in the indistinguishable counterparts (cf. Malament 1977, pp. 74-76). Although T_{ab} inherits the symmetries of the metric, the source fields $O_1, ..., O_n$ do not necessarily share the symmetries (see, e.g., Tariq and Tupper 1992, for discussion of cases in which the source fields do not inherit symmetries of the metric). The source fields are tensor fields defined everywhere on M, such as the Maxwell tensor F_{ab}, which satisfy the appropriate field equations.

the nature of future infinity. Second, surely the epistemic situation of an observer in M does not depend on that of observers in M'—undercutting the symmetry requirement.[30]

The familiar de Sitter solution provides a nice illustration of a spacetime with indistinguishable counterparts (following Malament 1977): even observers idealized as inextendible timelike curves in de Sitter space only see a finite "strip" of the full spacetime, and the initial spacetime is indistinguishable from other spacetimes that differ beyond this strip. The de Sitter solution can be visualized in a reduced model (suppressing two spatial dimensions) as a one sheet hyperboloid H imbedded in a flat three-dimensional space (see figure 5.1). Future and past timelike infinity are spacelike surfaces in the de Sitter spacetime (see A.4). A group of test particles moving along geodesics in de Sitter space rapidly separate with the expansion of the space. The horizons resulting from this expansion can be easily visualized in the two-dimensional covering space of de Sitter space, the t, x plane with the metric $ds^2 = dt^2 - (cosh^2 t)dx^2$. Since the light cones narrow as $|t| \rightarrow \infty$, every observer can see only a vertical strip of the spactime 2π wide in the x coordinate of this metric. This strip of the covering space is OI from the two dimensional reduced model of de Sitter space.

A second example illustrates that global properties may vary between WOI counterparts. Consider Minkowski spacetime with a closed ball O surgically removed. The pre-surgery version of Minkowski spacetime $\mathbb{R}^4, \eta_{ab}$ is WOI from the mutilated version, since the chronological

[30] This definition of WOI can be further modified by taking into account the possibility that event horizons may hide some regions of spacetime. The current definition assumes a "democracy of observers": all points in the manifold are considered in the construction of an observational counterpart. However, Penrose's cosmic censorship conjecture (see, e.g., Penrose 1979) suggests a division between two different classes of observers: those outside black hole event horizons protected from singularities (and possibly other causality violations) by the Cosmic Censor, and the poor infalling observers who may sneak a peak at a naked singularity. If the definition of OI is restricted to the former, they may remain ignorant of even the *failure* of various causality conditions. Whether this is in fact the case depends on the black hole uniqueness theorems, which suggest that distant observers can reliably distinguish different black hole spacetimes without receiving word from the infalling observers. Clarifying this weaker notion of observability requires a detailed study of black hole spacetimes, something I will not pursue further here.

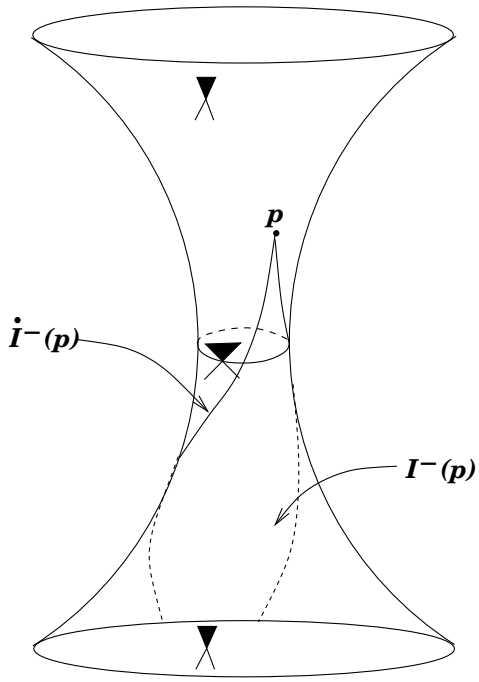

Fig. 5.1 The reduced model of de Sitter spacetime, a hyperboloid embedded in flat space, illustrating the chronological past $I^-(p)$ and the "collapsing" light cones as one moves away from the neck of the hyperboloid. See Schrödinger (1957) for a clear description of the de Sitter solution.

past of any observer in Minkowski spacetime can be embedded "below" the mutilation. Symmetry fails, since any observer in the mutilated spacetime ($\mathbb{R}^4 - O, \eta_{ab}$) whose causal past included the removed set would be well aware that she was not in Minkwoski spacetime anymore. This example illustrates that the existence of a Cauchy surface is not invariant under the relation of WOI (there are Cauchy surfaces in Minkowski spacetime, but not in the mutilated counterpart). More generally, the WOI counterpart to a given spacetime can be visualized as the sets $I^-(p_i)$ hung along a "clothesline" with space-time filler in between.[31] Here we are not concerned with whether the WOI counterpart is actually a *sensible* cosmological model in its own right; the space-time filler is allowed to vary arbitrarily between the $I^-(p)$ hung on the clothesline, as

[31] A proof due to Geroch (1968, pp. 1743-44) guarantees that one can always find a countable sequence $\{p_i\}$ such that the union of their chronological past covers M, i.e. $M = \bigcup_{p_i}\{I^-(p_i)\}$.

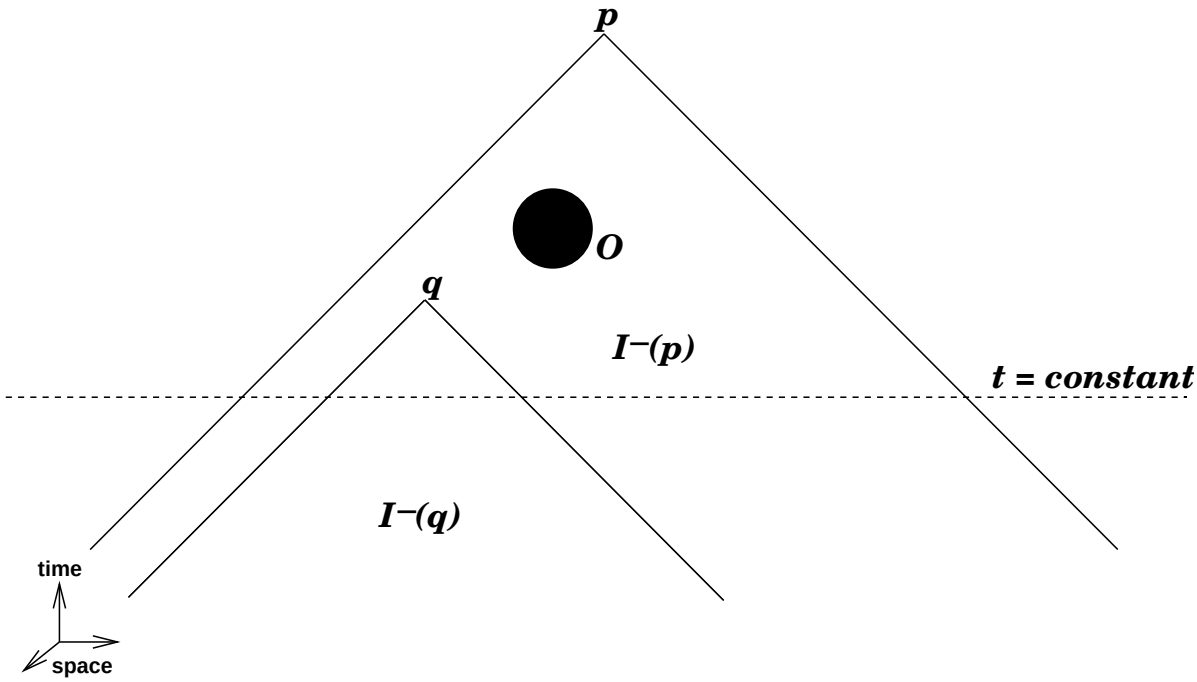

Fig. 5.2 "Mutilated" Minkowski spacetime (the set O excised), which is WOI from standard Minkowski spacetime. An observer at p can detect the causality violations associated with the excised region, but an observer at q cannot.

long as continuity holds on the boundaries. Malament (1977) presents of series of brilliant constructions to illustrate that only the *failure* of various causality conditions necessarily holds in WOI counterparts (see, in particular, the table on p. 71).[32] As Malament emphasizes, an observer may know conclusively that one of the causality conditions is violated, but no observers will ever be able to establish conclusively that causality conditions hold.

A natural objection to this line of thought is that we *should* be concerned with whether the constructed indistinguishable counterparts are sensible cosmological models in their own right. While these indistinguishable counterparts are solutions of the EFE, they are constructed

[32] I share Malament's intuition that the only spacetimes without a WOI counterpart are *totally vicious* (i.e., for $\forall p \in M, p \in I^-(p)$), although I have not been able to prove a theorem to this effect.

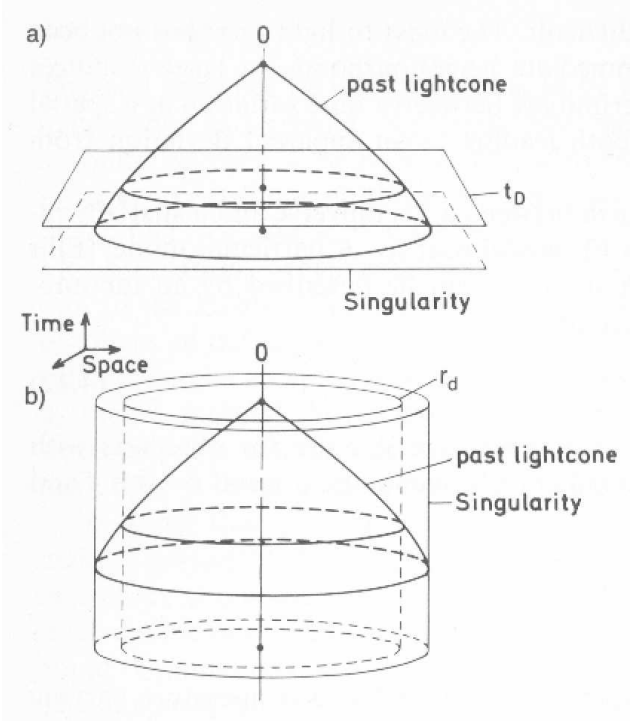

Fig. 5.3 This figure contrasts the standard big bang model (a) and Ellis et al. (1978)'s model (b); in the latter, a cylindrical timelike singularity surrounds an observer O located near the axis of symmetry, and the constant time surface t_D from which the CMBR is emitted in the standard model is replaced with a surface r_D at fixed distance from O (figure from Börner 1993, p. 130).

by stringing together "copies" of $I^-(p)$ sets and generally require a bizarre distribution of matter. This objection suggests that counterparts should be subject to a stronger constraint, namely that they correspond to solutions of the EFE that can be derived from physically motivated assumptions about the matter content.

Ellis et al. (1978)'s example of an indistinguishable counterpart to the observed universe illustrates the difficulties with satisfying such a stronger constraint. Their model incorporates isotropy for a preferred class of observers, but abandons homogeneity and the usual conception of how sources evolve. In this static, spherically symmetric model, *temporal* evolution (of, e.g., the matter density or various astronomical objects) in the standard FLRW models is replaced with *spatial* variation symmetric around a preferred axis (see 5.3). Unlike the timelike big bang

singularity of the FLRW models, this model incorporates a singularity that "surrounds" the central region at a finite distance (all spacelike radial geodesics intersect the singularity). Ellis et al. (1978) show that such a model can accomodate several observational constraints, at least for an observer whose worldline is sufficiently close to the axis of symmetry. They counter the obvious objection that it is unreasonable to expect our location to be close to the "center of the universe" with an anthropic argument (p. 447): in such a model, only the central region is (literally) cool enough for observers. However, there is a more substantial objection: it turns out to be quite difficult to match the observational constraints on the magnitude-redshift relation given EFE with a perfect fluid source. Roughly, the symmetries of the model place very tight constraints on solutions to the field equations, and of the six possible solutions none fit the observed magnitude-redshift relation without unnatural modifications (see Ellis et al. 1978, §6 and §7).[33] But this is precisely the point of the exercise: the model is suspect *not* because it violates spatial homogeneity, but rather because of the difficulty in satisfying both the EFE for a reasonable equation of state and observational constraints.

The major difficulty with replacing the definition of WOI given above with a physically motivated constraint along these lines also appears in other areas, such as attempts to prove

[33]To be more precise, Ellis et al. (1978) note that the for the solution to remain static the gradient in the gravitational potential as one moves out along the radius must be matched by a pressure gradient. But this implies that the present era is radiation dominated in the alternative model (rather than matter dominated, as in the standard models), since "dust" uncoupled to radiation does not satisfy the equation of hydrostatic support. Hence the alternative model uses an equation of state with $p = \rho/3$, with a non-zero Λ thrown in for an added degree of freedom. They conclude that that if $\rho > 0$ (satisfying the strong energy condition), there is no choice of the parameters of the theory that fits the observed magnitude - redshift relation. There are a few ways to avoid this conclusion, such as considering much more complicated equations of state or alternative gravitational theories, but Ellis et al. (1978) dismiss the alternatives as not "immediately compelling."

Penrose's cosmic censorship conjecture: what exactly should be required of a "physically reasonable" solution of the field equations?[34] Requiring that the source term T_{ab} satisfies various energy conditions will not do in this case, since a clothesline-constructed counterpart satisfies any energy conditions satisfied in the original spacetime; other more restrictive constraints on T_{ab} fail for the same reason. Ignorance of the space of solutions of the EFE also makes it difficult to imagine how one could formulate a "naturalness" or "simplicity" requirement in terms of initial data specified on some Cauchy surface Σ that would rule out WOI counterparts. The WOI counterparts certainly look like Rube Goldberg devices rigged up to be indistinguishable from a given spacetime. However, exact solutions with high symmetry are also "unnatural," and it is hard to see how to formulate a criterion that would rule out WOI constructions but not solutions such as the FLRW models. Without an entirely general formulation, we have instead the piecemeal approach of Ellis et al. (1978): construct a model without spatial homogeneity and a given equation of state, then see whether it can accomdate various observational results. Failure to construct a workable model may reflect lack of imagination rather than a fundamental feature of GTR, and so this only provides slight evidence for the claim that the FLRW models are the only physically reasonable models incorporating isotropy.

Adding information from multiple observers reduces the freedom in constructing indistinguishable counterparts. Spatial homogeneity is the strongest form of this requirement: it stipulates exact symmetry between every fundamental observer. More precisely, homogeneity holds if there are isometries of the spatial metric on each Σ—three-surfaces orthogonal to the tangent vectors of the fundamental observers' worldlines—that carry any point on the surface

[34] See Earman (1995), Chapter 3 for a comprehensive discussion of cosmic censorship and an extensive list of references to the physics literature.

into any other point. Suppose that we amend the definition of WOI to include the requirement that homogeneity must hold in M, g_{ab} as well as M', g'_{ab}. Pick a point $p \in M$ such that p lies in Σ, and choose an isometric imbedding map ϕ such that the point $\phi(p)$ is in M'. If homogeneity holds, then M' must include an isometric "copy" Σ' of the *entire Cauchy surface* Σ along with its entire causal past. Take ξ to be an isometry of the spatial metric defined on Σ, and ξ' an isometry on Σ'. Since $\phi \circ \xi(p) = \xi' \circ \phi(p)$, and any point $q \in \Sigma$ can be reached via ξ, it follows that Σ is isometric to Σ'. Mapping points along an inextendible timelike curve from M into M' eventually leads to an isometric copy of our original spacetime. If both cosmological models are inextendible, there are no indistinguishable counterparts (up to isomorphism) under this amended definition.[35]

Even a weaker requirement than the exact symmetry of spatial homogeneity reduces the scope of indistinguishable counterparts. Exact isotropy around two or more distinct spacetime points entails homogeneity, and it is also true that (with appropriate qualifications) "near isotropy" entails "near homogeneity." The Copernican Principle (whether appropriately named or not) is typically characterized as requiring that "our location is not distinguished." Here I will take the Copernican Principle to make the stronger requirement that, roughly put, no point $p \in M$ is distinguished from other points q by spacetime symmetries. Ellis et al. (1978)'s model would fail to meet this requirement, since there are points distinguished by their proximity to the axis of symmetry.[36] The Ehlers-Geren-Sachs theorem (Ehlers et al. 1968) shows that if all

[35] An *inextendible* spacetime cannot be imbedded as a proper subset of another spacetime. This qualification is needed to rule out spacetimes such as a "truncated" FLRW model, in which there is an end of days—a "final" time slice at an arbitrary cosmic time t_{end}. Such a model would be WOI (in the amended sense) from its extension, in which time continues past t_{end}.

[36] What is lacking here is a precise way of stating that there should be an "approximate symmetry" obtaining between different fundamental observers separated by some length scale L, in that they see a distribution of galaxies and fluctuations of temperature in the CMBR that differ only due to the random

fundamental observers in an expanding model find that freely propagating background radiation is exactly isotropic, then their spacetime is an FLRW model.[37] Recent work has clarified the extent to which this result depends on the various exact claims made in the antecedent. The fundamental observers do not need to measure *exact* isotropy for a version of the theorem to go through: Stoeger et al. (1995) have further shown that *almost* isotropic CMBR measurements imply that the spacetime is an *almost* FLRW model.[38] There are, however, counterexamples showing that the theorem does not generalize in other respects. Given the assumption that the matter content can be characterized as pressureless dust completely decoupled from background radiation, the fundamental observers travel along geodesics. Clarkson and Barrett (1999) show that non-geodesic observers can observe an isotropic radiation field in a class of inhomogeneous solutions. In addition, observational constraints confined to a finite time interval may not rule out more general models which approximate the FLRW models during that interval but differ at other times (Wainwright and Ellis 1997).

5.3.2 Status of the CP

The previous section clarified the extent to which claims regarding global structure require an appeal to a general principle of uniformity. Bold global extrapolations certainly may fail: if inflation occurred, for example, then the CP would only apply to the interior of post-inflationary "bubbles" rather than to the universe on the largerst possible scales. In practice

processes generating them. See Stoeger et al. (1987) for a proposed definition of "statistical homogeneity" along these lines, defined with respect to a given foliation.

[37] "Freely propagating" means that the radiation is decoupled from the matter; the stress energy tensor can be written as two non-interacting components, one for the dust-like matter and another representing the background radiation.

[38] Wainwright and Ellis (1997) introduce various dimensionless parameters defined in terms of the shear, vorticity, and Weyl tensor to measure departures from the exact FLRW models; a spacetime is *almost* FLRW if all such parameters are $<< 1$ (see, in particular, pp. 62-64).

these principles have generally played the role of defining the space of possible models considered by observational cosmologists. This strategy is a familiar one in science, and it has been dubbed (among other things) eliminative induction.[39] In an eliminative induction very general premisses delineate the scope of possible theories or hypotheses. The epistemic risk associated with an inductive argument is shifted entirely to these premisses; the *deductive* argument to the conclusion that some subset of possible hypotheses is correct proceeds by ruling out the competitors. Observational programs that aim to determine the "best cosmological model" only stand a chance when the list of possible models has been trimmed down using the CP or a similar principle. Of course this strategy shifts the epistemic risk from the inference to the principles invoked in it, rather than eliminating that risk entirely. The advantage consists of replacing an inductive generalization with a careful characterization of the "uniformity of nature" appealed to in a particular context.

In comparison to other cases of inductive generalization, taking on the epistemic risk associated with the CP is remarkably unproductive. Consider (again) a brief contrast with Newtonian gravitation: the latter half of Book III of the *Principia* illustrates the potential payoff of universal gravitation, as Newton gives preliminary accounts of the tides, the motion of the moon, the shape of the earth, and so on. The inductive step (however it is characterized) to universal gravity leads to a host of further empirical problems that present an opportunity to refine and develop the theory, and indeed these problems spurred the development of celestial mechanics throughout the eighteenth century. In the cosmological case, invoking the CP to make global extrapolations does not lead into similarly rich empirical territory. These extrapolations may

[39] I thank John Norton for emphasizing the importance of eliminative induction to me in several discussions; see his Norton (1994), for example.

satisfy the urge to speculate about the overall global structure of the univers and its eventual fate, but there are no immediate testable consequencs of accepting or rejecting the CP. This is not to say that there are no issues of theoretical interest that depend on the truth of the CP: in particular, the singularity theorems require assumptions regarding the global causal structure of spacetime.

5.4 The Indifference Principle

In slogan form, the indifference principle holds that we should prefer a theory that does not require "special" initial conditions (or "special" parameter values). Momentarily we will turn to the difficulties of making the slogan precise, but in any case there is no shortage of sloganeers.[40] In the *Discourse on the Method* Descartes stated a preference for indifference as follows (Descartes 1985, pp. 132-34):

> I therefore supposed that God now created, somewhere in imaginary spaces, enough matter to compose such a world; that he variously and randomly agitated the different parts of this matter so as to form a chaos as confused as any the poets could invent; and that he then did nothing but lend his regular concurrence to nature, leaving it to act according to the laws he established. [...] So, even if in the beginning God had given the world only the form of a chaos, provided that he established the laws of nature and then lent his concurrence to enable nature to operate as it normally does, we may believe [...] that by this means alone all purely material things could in the course of time have come to be just as we now see them.

Descartes' hope that the new mechanics would render teleology unnecessary was as unfounded as it was bold.[41] Closer to hand, we find cosmologists displaying similar rationalistic leanings (Sciama 1959, pp. 166-67):

---

[40] McMullin (1993) gives an insightful overview of the history of the idea from the Greeks to contemporary cosmology.

[41] Several of Descartes' contemporaries (and later generations of natural theologians) reached exactly the opposite conclusion. For example, Newton held that the action of Divine Providence (whose mysterious ways include the judicious use of comets) was needed to insure the stability of the solar system (Kubrin 1967).

> [We must] find some way of eliminating the need for an initial condition to be specified. Only then will the universe be subject to the rule of theory. ... This provides us with a criterion so compelling that the theory of the universe which best conforms us to it is almost certain to be right.

Sciama was discussing the steady state theory, and within a few short years he would admit that observational accuracy provides an even more compelling criteria for theory evaluation. But his target was clearly the big bang model, which apparently requires a highly specialized initial state very different from Cartesian Chaos. Ironically, indifference was also accepted as an important theoretical virtue by fans of the big bang model: Misner and later Guth hoped that any imprint of an initial "chaotic" beginning could be erased by subsequent evolution. As we saw in Part I, these lines of research have dominated early universe cosmology since the late 60s.

The great appeal of these ideas is that they resolve an apparent conflict between a widely accepted assumption regarding the universe's initial state and the observed universe. I will call this assumption the "Creation Hypothesis" (CH): the initial state of the universe is "chosen at random" from the space of physically possible models. Perhaps the Creator tossed a dart at the Cosmic Dart Board representing cosmological models without aiming for anything in particular. Even without a good understanding of the space of solutions to EFE or how one is chosen to be "actualized," it seems clear that one of the maximally symmetric FLRW models must be an "improbable" or "finely-tuned" choice: for any reasonable choice of measure, these models are presumably a measure-zero subset of solutions to EFE. (More precisely, models lacking symmetry form a dense, open subset of the space of solutions to EFE; see Isenberg and Marsden 1982 .) An "average" cosmological model lacks the global symmetries of the FLRW models, and instead features bewildering inhomogeneities "as confused as any the poets could invent." Furthermore, dynamical evolution governed by the EFE appears to enhance initial inhomogeneities rather than

suppress them. Thus CH immediately leads to a dramatic conflict with observations: the most "probable" initial state develops into nothing like our universe.[42]

There have been three different responses to this apparent conflict (cf. Hartle 1986):

1. Dynamics – New dynamics is introduced to turn a "typical" initial state into the observed universe, in the process (partially) erasing the imprint of the initial conditions. The apparently "finely tuned" features are a consequence of these dynamics.
2. "Theories of initial conditions" – The assumption that the initial state is chosen randomly from among the states allowed by classical GTR should be rejected, since the full theory of quantum gravity may incorporate principles (in the form of global constraints) that trim down the space of physically possible models.
3. Anthropic principle – The "special" features of the initial state are necessary preconditions for our existence. As Collins and Hawking (1973) put it, "the answer to the question 'why is the universe isotropic?' is 'because we are here' " (p. 334).

I will turn to the anthropic principle in the next section. After discussing the first two responses, I will argue in favor of a skeptical response. I will focus on two aspects of the set-up for the apparent conflict. All the talk of probabilities above has been carried out at the level of rough intuitions. In the following subsection, I will review serious obstacles to putting these intuitions on firmer footing. Even granting for the moment that these probabilistic arguments make sense, the perception of a conflict depends on an implicit assumption that the space of physically possible cosmological models should directly match the observed universe. But what exactly is the problem with unrealized physical possibilities? Lawlike generalizations in other sciences, such as biology, hold only in a subset of physically possible models, and do not *define* the space of physically possible models. Commitment to the indifference principle reflects a demand that cosmology should resemble the methodology of physics in seeking fundamental laws.

[42] Of course, given the probabilistic nature of the hypothesis there is not an *outright* conflict here, but CH renders the observed history of the universe *incredibly* improbable.

As we saw in part I, early universe cosmology has been dominated by attempts to implement the dynamical approach. Supplementing the standard big bang model with new dynamics satisfies the indifference principle, since the theory no longer requires special initial conditions. In slogan form the big bang model with new dynamics offers a more robust dynamical explanation of the regularities at issue. Here a word of caution is in order: the suggestion that introducing new dynamics *eliminates* dependence on initial conditions is false advertising. This point, first spelled out by Collins and Stewart (1971) in response to Misner's work, bears repeating since it is often ignored. In the case of inflation, even if inflation occurs one can choose initial conditions that lead to an arbitrarily non-uniform universe with any value of Ω, despite inflation's "preference" for a uniform universe with $\Omega = 1$. Thus the advantage of new dynamics lies in *enlarging* the range of initial conditions compatible with observations rather than *eliminating* dependence on initial conditions.[43] A second point will be familiar from the discussion of the fine-tuning problems of new inflation in Chapter 4: current versions of inflation typically involve a trade-off between fine-tuning of the initial conditions and fine-tuning of the dynamics (in the form of specifying the potential of the inflaton field). Clarifying and assessing the demand for robustness will be the focus of Chapter 6, but for the moment I will grant that introducing new dynamics renders the observed universe "more probable" by enlarging the range of compatible initial conditions.

Consider the following general argument in favor of introducing new dynamics. The overall uniformity of the universe is one of its most striking and fundamental features. The probabilistic arguments above indicate that classical general relativity only accomodates this

[43] Guth (1997b) acknowledges this point: "... I emphasize that *NO* theory of evolution is ever intended to work for arbitrary initial conditions. ... In all cases, the most we can hope for is a theory of how the present situation could have evolved from *reasonable* initial conditions" (pp. 240-241, emphasis in the original).

fundamental feature of our universe as a remarkably improbable, contingent feature of a particular model. Surely it would be more reasonable to assume that this feature directly reflects the laws of nature, rather than some contingent, global constraint on the initial data? Perhaps this odd fact is a significant hint of how the laws of quantum gravity (presuming such a theory exists) differ from those of classical general relativity; namely, they insure that uniformity results from dynamical evolution for almost any initial state. The striking uniformity of the early universe looks like a clear signpost guiding the way in formulating new fundamental theories applicable to the early universe. This certainly seems more progressive than simply chalking up the universe's uniformity to special initial conditions.

But what is the force of this argument? Physicists undoubtedly focus on isolating the deep structure of their theories in the drive to reformulate and improve them. But deciding what should be caught in the web of necessary connections embodied in a physical theory is undoubtedly one of the most difficult tasks facing creative theorists. There are numerous examples of regularities once thought to be deeply enmeshed with the laws of nature that subsequently slipped through the netting: Kepler thought that the number and relative distances of the planets and their moons was as closely related to the laws of nature as his harmonic law, Leibniz took the common direction of orbital rotation of the planets to reflect a lawlike regularity accounted for by the vortex theory, and it is fairly easy to produce other historical examples. Clearly we need to avoid casting the net too broadly, without thereby curtailing the search for new ways of catching observed regularities within physical theory.

Turning to a closely related issue that has been the focus of more philosophical discussion will help to clarify matters. The difference between the past and future is one of the

most striking features of our universe.[44] The Second Law of thermodynamics characterizes this asymmetry in terms of the entropy: for a closed system, the change in entropy ΔS obeys the following inequality: $\Delta S \geq 0$. This law implies that entropy does not decrease in transitions that a closed system undergoes, and entropy reaches its maximum value for equilibrium states. Systems obeying the law evolve *toward* equilibrium states in the future, but they do not evolve *from* equilibrium states in the past. Since Boltzmann's first "derivation" of the infamous H-theorem, a tremendous amount of effort has been devoted to understanding the status of the Second Law with regard to classical (and quantum) statistical mechanics. Phenomenological thermodynamics is usually taken to be reducible in some sense to statistical mechanics (see, in particular, Sklar 1993); suppose that the laws of thermodynamics are meant to be recovered as laws of statistical mechanics.[45] This assumption leads to a fundamental conflict due to two features of the laws of classical statistical mechanics (clearly recognized by Boltzmann's contemporaries, Zermelo and Loschmidt): the laws of classical statistical mechanics are time reversal invariant, and the evolution of a closed system obeying these laws is quasi-periodic. Applying the indifference principle to this case, we should conclude that the asymmetry of time should be directly reflected in the laws of statistical mechanics, and join in the century-old hunt for new physics incorporating an arrow of time.

There is an alternative to joining the hunt (as with many ideas in statistical physics, first suggested by Boltzmann): the laws of statistical mechanics conjoined with the "past hypothesis" (using Albert 2000's terminology) are compatible with the Second Law. According to the past

[44] See, e.g., Sklar (1993); Albert (2000) for much more detailed discussions of these matters.

[45] This is not intended to be a minimal assumption required for an account of the reduction of thermodynamics to statistical mechanics. Indeed, this is a very strong requirement, and I am sympathetic to Callender (2001)'s position that lowering the sights (by requiring only that statistical mechanics recovers suitable analogues of thermodynamics laws) eases this conflict.

hypothesis, in the distant past the universe was in a low entropy state, but it will not return to a low entropy state in the distant future.[46] Due to this initial state, transitions towards equilibrium are to be expected whereas those away from equilibrium are incredibly unlikely. To those hot in pursuit of new physics, this resolution seems to have all the advantages of theft over honest toil. It solves the conflict by *fiat*, without (as the standard complaint has it) really providing a satisfactory *explanation* of the striking time asymmetry of observed phenomena. How could a contingent matter of fact regarding the global distribution of matter and energy shortly after the big bang serve as explanatory grounds for the fundamental asymmetry manifest in local phenomenon? The advocate of adopting the past hypothesis has to bite the bullet and admit that the cooling of my cup of coffee is, in some sense, explained by the universe's initial low entropy state. Those pursuing new physics expect us to share their feeling that this is inherently unsatisfactory, and join in the demand for a better explanation.

Several cosmologists have made a move analogous to Boltzmann's, in shifting focus from local, dynamical laws to global features of the initial state. But unlike Boltzmann they accept the argument that the universe's uniformity *must* reflect the laws, and aim to develop a "theory of initial conditions" that singles out a unique initial state. Uniformity (and perhaps other properties) directly reflect the universe's initial state rather than its subsequent evolution. In other words, these theories propose that we "trim down" the Cosmic Dart Board to include, say, only highly uniform initial states. Little wonder then that the "actualized" initial state is compatible with what we observe!

[46] I am glossing over two points here. First, an advocate of this line of thought needs to relate the overall low entropy boundary condition to the behavior of local systems; see Sklar (1993) for discussion. Second, the "distant past" need not refer to an early time slice in an FLRW model, as I will assume below; for example, Boltzmann thought of the initial low entropy conditions as resulting from a "local" fluctuation away from equilibrium.

Penrose (1979, 1989)'s suggestion that the initial state should have vanishing gravitational entropy provides one example of this approach. There is no widely accepted definition of gravitational entropy, but Penrose argues that the natural equilbrium state for matter is maximally "clumpy" (consisting of black holes) since gravity enhances inhomogeneities. Thus the uniformity of the early universe reflects incredibly low gravitational entropy, and it is no coincidence that this supplies the low entropy initial state required by the past hypothesis. Penrose's conjecture is that the Weyl curvature tensor approaches zero as the initial singularity is approached; his hypothesis is explicitly time asymmetric, in that the Weyl curvature increases as black holes are approached.[47] Imposing this constraint requires that the early universe approaches an FLRW solution. This idea still has the status of an imaginative, "botanic" proposal—while it classifies the nature of the initial singularity, it has not yet been derived from a theory of quantum gravity.[48]

Quantum cosmology provides another example of a "theory of initial conditions." Research in quantum cosmology has long had the stated goal of finding laws which uniquely determine the initial quantum state of the universe. Hawking has claimed that "the no-boundary proposal makes cosmology into a science, because one can predict the result of any observation" without an *ansatz* for the initial conditions (Hawking and Penrose 1996, p. 86). The no-boundary proposal is formulated using Euclidean techniques borrowed from the path integral formulation

[47] The Weyl tensor C_{abcd} is the trace-free part of the Riemann curvature tensor, and it represents the spacetime curvature due to the gravitational field itself. The FLRW models (and other conformally flat spacetimes) have vanishing Weyl curvature, which motivates the use of vanishing Weyl curvature as the measure of uniformity. Goode et al. (1992) formulate this hypothesis in terms of the limiting behavior of C_{abcd} in a conformal completion, which includes $t = 0$ as a spacelike hypersurface.

[48] The term is Hawking's, but Penrose agrees that it applies to the Weyl curvature hypothesis; see Hawking and Penrose (1996, p. 106).

of QFT. The path integral for gravity coupled to a scalar field is given by:[49]

$$Z(M) =: \int d\mu(g_{ab}) d(\phi) exp(-S_{E}[g_{ab}, \phi]) \tag{5.3}$$

The no boundary proposal then calculates the probability of a solution with an initial geometry (Σ, h_{ij}) within the "range of values" O as $\int_O |\Phi|^2 d\mu(h_{ij})$ where h_{ij} is the induced three metric on the surface Σ, and ϕ_0 is the field configuration of the scalar field. $\Phi(h_{ij}, \phi_0, \Sigma)$ is the "ground state" wave function(al) defined as the sum of $Z(M)$ ranging over compact Euclidean manifolds (hence the name) with a unique boundary Σ. Hawking and Hartle (1983) argue that this probability should be interepreted as the probability that the universe appears from nothing with the initial spatial geometry given by Σ, h_{ij}. It will perhaps come as no surprise that this flight into the formalism takes several twists and turns that others have not followed, and substantial questions remain regarding the proper criteria for choosing an appropriate wave function and then interpreting the chosen one. Vilenkin (1998) concludes a review of competing definitions on an appropriately cautionary note:

> ... the reader should be aware that all three wave functions [proposed by Hawking and Hartle, Linde, and Vilenkin] are far from being rigorously defined mathematical objects. Except in the simplest models, the actual calculations of these wave functions involve additional assumptions which may appear reasonable, but are not really well justified.

Although further research in quantum cosmology may lead to universal acceptance of a particularly "natural" wave function, the fundamental disagreements so far are not encouraging.

[49] S_E is the Euclidean action for gravity coupled to a scalar field ϕ, and the integral ranges over all Reimannian metrics on a given manifold. There is no general rigorous definition for the measures $d\mu(g_{ab}$ and $d(\phi)$, but theorists have made good use of Euclidean techniques in tackling a wide variety of problems. See Gibbons and Hawking (1993) for an overview of these techniques, and Isham and Butterfield (2000) for a philosophical assessment of Euclidean quantum gravity.

Despite their differences these proposals share one common ingredient: they both introduce *global* laws, in the form of global constraints on the initial state. Although they are often ignored by philosophers, similar "laws of coexistence" appear in other dynamical theories as constraints on initial data (discussed in more detail in §6.2.2). Both proposals in effect attempt to identify such constraints in a full theory of quantum gravity despite our ignorance; while this is certainly a risky strategy, I see no reason to rule out the possibility that quantum gravity incorporates such a global constraint.

Returning to the line of thought above, why not admit that the initial state may simply reflect a brute, contingent fact? A modern Leibnizian would respond by reiterating the Principle of Sufficient Reason. But is such a demand for further explanation—either in terms of new local physics or in terms of a global constraint law—warranted? Physicists generally recognize that laws serve as explanatory stopping points: there is no empirical answer to a question regarding why a particular law of nature obtains. Initial conditions may also function as explanatory stopping points in the same way as laws, at least in the sense that it would be inappropriate to make further explanatory demands regarding initial conditions. The force of traditional empiricist criticisms of the cosmological argument is that attempts to meet the demand carry one into metaphysics or theology. Vilenkin (1983) offers a rare confession that in discussing the quantum state of the universe he (and his colleagues) are engaged in "metaphysical cosmology," which he defines as "the branch of cosmology totally decoupled from observations" (p. 2854). Rather than attempting to clarify the sense of explanation invoked above, the point can be put in terms of the epistemic risk involved in declaring that a particular regularity reflects the laws rather than

initial conditions. In most cases, the claim that a law obtains entails a number of other consequences that do not follow if the regularity is chalked up to contingent initial conditions, but in cosmology this is a modal distinction without an empirical difference.

Up to this point I have argued that invoking metaphysical principles to justify a demand for either further dynamical theories or a theory of initial conditions is unconvincing. But I expect that many cosmologists would recoil at being classified as metaphysicians, and instead insist that other considerations motivate research in early universe cosmology. I will turn to these other considerations in the next two chapters.

5.4.1 Probabilities

The discussion above has assumed that the intuitively compelling fine-tuning arguments can be put on firmer footing. The force of these arguments derives from the astounding numbers involved; typical presentations of the flatness problem end with a dramatic punchline: in the classical cosmological models, $|\Omega - 1| < 10^{-59}$ at the Planck time! But arguments like these are notoriously shaky as long as the ensemble and probability distribution assigned over it remain unspecified.[50]

The sting of the flatness problem derives from the assumption that the probability for initial values of Ω is "spread out evenly" over an interval of values $(0, a) \in \mathbb{R}$ with $a > 1$. In other words, one takes a uniform probability with respect to the standard Lebesgue measure on a set of real numbers. But is there any reason to use this particular measure in assigning probabilities to the initial values of Ω? Several authors have argued that the flatness problem disappears when

[50] George Ellis in particular has frequently emphasized the diffi culties with applying probabilistic arguments in cosmology, see Ellis (1990); Coles and Ellis (1997) and references therein.

this measure is replaced with a more appropriate choice.[51] For example, Hawking and Page (1988) argue that a dynamically invariant measure (introduced in Gibbons et al. 1987; Henneaux 1983) "solves" the flatness problem, in the sense of showing that almost all solutions to the field equations for classical FLRW models coupled to a massive scalar field have negligible spatial curvature. The FLRW models with a massive scalar field can be treated as a constrained Hamiltonian system, where the trajectories through phase space correspond to the dynamical evolution of a cosmological model. The phase space naturally possesses the structure of an even-dimensional symplectic manifold. Gibbons et al. (1987) showed that the symplectic form of this phase space can be used to define a volume element μ on a cross section of the constraint space. This volume element is invariant under the Hamiltonian phase flow, which means that the measure of a given trajectory (or set of trajectories) is independent of the cross section used to evaluate the measure. Hawking and Page (1988) show that according to this canonical measure, all but a finite subset of solutions behave like the "flat" FLRW models (i.e., they expand to arbitrarily large size with negligible spatial curvature). Thus a probability distribution that is uniform with respect to μ would assign a zero probability to the set of non-flat solutions; in this sense there is no flatness problem.

This result exploits the fact that a probability distribution that is absolutely continuous with respect to μ assigns a zero probability to any measurable set whose complement has finite measure, in the case where the full space has infinite measure. But as Hawking and Page (1988)

[51] In addition to the "canonical" measure introduced by Gibbons, Henneaux, Hawking, and Stewart, Evrard and Coles (1995) derive a "minimal information" measure using Jaynes's principle. Their argument based on this measure leads to a similar conclusion, namely that "*there is no flatness problem* in a purely classical cosmological model" (original emphasis, p. L96). Cho and Kantowski (1994) introduce a kinematic measure based on De Witt's field metric (apparently compatible with the canonical measure), and reach similar conclusions to the work of Hawking and Page. For a review of the results outlined in the text, see Coule (1995).

go on to show, several other questions do not admit unambiguous answers based on measure alone: in particular, assessing the probability that inflation occurs in their model requires taking the ratio of two sets of infinite measure, and is thus ambiguous (cf. Hollands and Wald 2002a). The ambiguity could be resolved by introducing an objective probability distribution, if only there were grounds for doing so.[52] Whatever one might think of efforts to justify probability distributions in statistical mechanics, the strategies employed in that case do not carry over to cosmology. The phase space of cosmological models is clearly not ergodic since the trajectory corresponding to a particular cosmological model does not cycle through the phase space. Furthermore, there is no widely accepted definition of entropy for the gravitational field. The introduction of "tychistic" probabilities (single case objective chances) associated with the creation of the universe also faces important obstacles. The proper theoretical context for assigning single case probabilities to a "creation event" has not yet been (and may never be) formulated. Calculating probabilities based on current versions of quantum cosmology requires choosing and interepreting the wave function of the universe, and I briefly described the fundamental difficulties facing these two tasks above (see, e.g. Unruh and Wald 1989; Isham and Butterfield 2000, for further discussion). The nature of the probabilities assigned to various "initial states" is one of the central contentious issues in quantum cosmology, and any appeal to tychistic probabilities awaits its resolution.

This raises a second issue regarding Hawking and Page (1988)'s results: how should quantum effects expected to dominate near the singularity be taken into account? As Coule (1995, p. 456) aptly puts it, having an appropriate measure "is tantamount to having a correct

[52] See also Earman and Mosterin (1999); Sklar (1993) for discussions of the difficulties with introducing probability in this context.

quantum gravity theory." Attempts to finesse this issue have taken two different tacks. Belinsky and Khalatnikov (1987); Belinsky et al. (1988) introduce a measure defined at the "quantum boundary," the surface where the energy density of the scalar field reaches m_p^4 ($m_p \approx 2.2\times10^{-5} g$ is the Planck mass). The classical initial data are assumed to be equiprobable—equal areas on the quantum boundary are assigned equal probabilites. Belinsky and Khalatnikov (1987) argue that this is the simplest proposal for assigning probabilities in the absence of further knowledge regarding the quantum initial state. Since the proposed measure is not invariant under dynamical evolution (Hollands and Wald 2002b, pp. 6-7), choosing a different boundary surface would result in different probabilities. A second more ambitious approach aims to incorporate projected principles of Planck-scale physics prior to assigning a measure. For example, according to the *Planck equipartition proposal*, at the Plack time ($t_p \approx 10^{-43}\ s$) all energy densities are roughly equal to the Planck energy density ($E_p \approx 10^{19}\ GeV$) (Barrow 1995). The underlying idea is that gravitational interactions prior to t_p will effectively transfer energy between gravitational degrees of freedom (such as the anisotropy energy density in gravitational waves) and matter-energy density. Although Barrow (1995) does not explicitly define a measure, he argues that (restricting attention to anisotropic, homogeneous models) the PEP eliminates models with large initial anisotropies, leaving it more probable that the observed isotropy is compatible with an "aribtrary" initial state without invoking inflation.

To sum up, the attempt to find firmer foundations for the assessments of probability invoked in the fine-tuning arguments has instead revealed shifting sand. It is impossible to avoid the theoretical uncertainties associated with the initial state in attempts to apply probabilities in cosmology. I should add that many cosmologists would presumably reject this demand for

firmer foundations as a mathematical nicety, and insist that the various fine-tuning problems are important clues whether or not they satisfy philosophers.[53]

5.5 Anthropic Principles

From its first explicit formulation in Carter (1974), the anthropic principle (in its many guises) has sparked considerable controversy among cosmologists. Many leaders in the field endorse some version of anthropic reasoning as an appropriate response to fine-tuning problems, yet other equally astute and respected scientists and philosophers dismiss the whole idea out of hand. Here I will argue briefly for a deflationary account. The term "anthropic principle" is as much a misnomer as "U.S.S.R.": the weak anthropic principle simply highlights the importance of observational selection effects, and involves nothing particularly "anthropic" and invokes no new "principles," and stronger versions of the principle make a number of broader (and questionable) explanatory demands. [54]

Dicke (1961)'s response to Dirac's cosmological speculations has been widely hailed as exemplifying succesful anthropic reasoning. Carter characterized Dicke's argument as an application of the "weak anthropic principle" (WAP), which I will define as follows (compare Carter 1974, Barrow and Tipler 1986, p. 16):

> *Weak Anthropic Principle*: What we observe is restricted by the necessary conditions for our existence.

[53] Charles Misner, for example, acknowledged that "mathematicians" find the fine-tuning arguments unconvincing, but still argued that they have intuitive force and heuristic value (Misner 2001).

[54] The anthropic arguments of Carter and other cosmologists in the 70s has spawned a vast literature; see, e.g., Barrow and Tipler (1986); Leslie (1989); Bertola and Curi (1993). My aim is to give a concise version of the deflationary account, which has been considered in greater detail and depth elsewhere (see, in particular Earman 1987b; McMullin 1993).

All parties to the anthropic debates have endorsed Dicke's use of the WAP, although in some cases the endorsement carries over to stronger anthropic principles and in others the WAP is treated as a corollary of confirmation theory or plain common sense. Dicke's argument undermined the surprising coincidence Dirac saw in order of magnitude equality between "large numbers" constructed out of fundamental constants:

$$\frac{t_0}{e^2/m_e c^3} \approx 10^{39}, \qquad \frac{e^2}{G m_p m_e} \approx 10^{39}. \tag{5.4}$$

Thus the ratio between the age of the universe t_0 and a natural "atomic time scale" (where e, m_e are the charge and mass of the electron and c is the speed of light) and the ratio between the electrostatic and gravitational forces between a proton (with mass m_p) and electron both have approximately the same value. Dirac's wonder at this coincidence (and others) inspired a new theory which he called the "Large Number Hypothesis": all such large dimensionless numbers constructed from the fundamental constants "are connected by a simple mathematical relation, in which the coefficients are of the order of magnitude unity" (Dirac 1937, p. 323). Since the first number includes the time t_0, so must they all; Dirac assumed that the masses and electric charge remain constant, forcing him to accept time variation of the gravitational "constant" G.

While Dicke was no foe of time-varying "constants," he trenchantly criticized Dirac for failing to take selection effects into account. Surprise at the fine-tuning coincidences might be warranted if t_0 could be treated as "a random choice from a wide range of possible values" (Dicke 1961, p. 440), but there can only be observers to wonder at the coincidence for some small range of t. Dicke (1961) argued that if main sequence stars are still burning and an earlier generation of red giants had time to produce carbon in supernovae—surely two minimal requirements for the

existence of observers like us—then the value of t must fall within an interval such that Dirac's coincidence automatically holds. The "evidence" provided by eqn. (5.4) bears no relation to the truth or falsity of Dirac's hypothesis, since these relations are guaranteed to hold.

The WAP thus appears to be nothing more or less than a selection effect. Rather than guiding scientific theorizing as a principle might be expected to do, WAP effectively neutralizes some evidential claims. Bayesians can account for this by explicitly conditionalizing on the presence of complex systems such as astrophysics PhD.'s: $P_s(\cdot) = P(\cdot|A)$, where P_s is the probability measure with the selection effect taken into account, and A is the proposition that astrophysicists exist. Dicke's argument shows that with this new probability measure, $P_s(LN|H_D) \approx P_s(LN|H_B) \approx 1$, where LN is the large number "coincidence" in eqn. (5.4), H_D is Dirac's cosmological theory, and H_B is the standard big bang theory. Thus the coincidence is neutral with regard to Dirac's idea or the standard cosmology. More generally, a selection effect is in force if the truth value of a given hypothesis is irrelevant to the evidence obtained in a given test. In other words, conditionalizing renders an originally "informative" piece of evidence E useless in that $P_s(E|H) \approx P_s(E|\neg H)$. Here I have cashed this idea out in Bayesian terms, but the idea of a selection effect should be captured in any adequate confirmation theory.[55] The imaginative and difficult part of arguments like Dicke's comes in recognizing the connections between our existence and a number of striking features of the universe—ranging from its overall near uniformity to the existence of a resonance level of the C^{12} nucleus at

[55] Clearly I have only given a sketch of how a selection effect might be accounted for, and two recent lines of work challenge this approach. Roush (1999, Chapter 2) argues that taking selection effects into account requires considering all evidence that is nomically related to our observational procedures, rather than only the evidence available. Treating selection effects by conditionalizing on given evidential claims may miss matters of fact that nonetheless produce biases in experimental or observational procedures, and Roush argues in favor of a more general account of selection effects partially motivated by Nozick's idea of "tracking". Bostrom (2002) develops a theory of observational selection effects, the centerpiece of which is the claim that all evidence statements should be understood as *indexicals* ("these observations are made by us"), where "we" are "randomly" picked members of a specified reference class.

around $\approx 7.7MeV$ (along with numerous other examples; see, e.g., Barrow and Tipler 1986). But throughout this argument we could have conditionalized on the existence of rutabagas or cockroaches and reached the same conclusions; what matters is the existence of complex carbon based systems and not *homo sapiens*, or even living things, in particular. This is not to say that more detailed information about our species never comes into play in considering selection effects, but rather that the general point regarding the use of evidence is not in any way anthropic.

Carter (1974)'s more provocative strong anthropic principle (SAP) was directly inspired by the fine-tuning worries discussed in the previous section. Most of the formulations of the principle share a family resemblance to this one:[56]

> *Strong Anthropic Principle*: The universe must have properties such that life develops within it.

This formulation clarifies little due to the ambiguity of "must," but in practice Carter and others wield the SAP as an explanatory demand. McMullin (1993) explicitly formulates the SAP as such: "Evidence of cosmic 'fine-tuning' ought to [or, in a weaker formulation, may be] given an anthropic explanation" (p. 377). All that separates this demand from the indifference principle discussed above is the qualifier "anthropic."

To clarify the explanandum and the form of "anthropic" explanation it will be helpful to consider one of the famous anthropic defenses of the big bang model, that of Collins and Hawking (1973). Homogeneous, anisotropic solutions of the EFE fall into three classes: those with an expansion rate lower than, equal to, or greater than the "escape velocity" (the rate of

[56] Other anthropic principles have been formulated, including the "final anthropic principle" – that life must not only exist but persist – and even crazier variants, but I will not discuss them here (see, e.g., Barrow and Tipler 1986).

expansion needed to avoid recollapse). The bulk of their paper is devoted to proving two theorems: first, that for homogeneous solutions of EFE which obey the dominant energy condition and positive pressure criterion, the set of initial data which isotropize as $t \to \infty$ is of measure zero; in other words, anisotropic modes dominate in generic models.[57] On the other hand, Collins and Hawking (1973) also showed that for particular models (Bianchi type VII_0) there is an open neighborhood of initial data including the exactly flat case, such that every element of this neighborhood approaches isotropy at late times. They respond to these results as follows (Collins and Hawking 1973, p. 319, cf. p. 334):

> We shall now put forward an idea which offers a possible way out of this [fine-tuning] difficulty. This idea is based on the discovery that homogeneous cosmological models *do* in general tend toward isotropy if they have exactly the escape velocity. Of course, such "parabolic" homogeneous models form a set of measure zero among all homogeneous models. However, we can justify their consideration by adopting a philosophy that has been suggested by Dicke (1961) and Carter (1968).[58] In this approach one postulates that there is not one universe but a whole infinite ensemble of universes with all possible initial conditions. From the existence of the unstable anisotropic mode it follows that nearly all of the universes become highly anisotropic. However, these universes would not be expected to contain galaxies [...]. The existence of galaxies would seem to be a necessary precondition for the development of any form of intelligent life. Thus there will be life only in those universes which tend toward isotropy at large times. The fact that we have observed the universe to be isotropic is therefore only a consequence of our existence.

Some commentators have characterized "anthropic explanations" as primarily aimed at removing puzzlement in the face of fine-tuning—in this case, isotropy shouldn't be puzzling since it

[57] Collins and Hawking (1973) define isotropization of an expanding model as the conjunction of the following three properties: (1) $T^{00} > 0$ and $\lim_{t\to\infty} \frac{T^{0i}}{T^{00}} = 0$, (2) $\lim_{t\to\infty} \frac{\sigma}{\dot{\alpha}} = 0$, where σ is the shear and $\dot{\alpha}$ is the volumetric expansion rate, and (3) the "cumulative distortion" $\beta \equiv \int_t \sigma dt$ approaches some constant as $t \to \infty$. If T^{ab} is diagonalizable, it can be written in the form $T_{ab} = \rho t_a t_b + \sum_{i=1,2,3} p_i x^i_a x^i_b$ where ρ corresponds to the energy density and p_i are the principal pressures. The dominant energy condition then states that $\rho \geq |p_i|$, and the positive pressure criterion holds that $\sum_{i=1,2,3} p_i \geq 0$.

[58] Here Collins and Hawking (1973) refered to an unpublished Cambridge University preprint by Brandon Carter, and I am unsure whether that paper ever appeared in print.

is a necessary condition for our existence. Galaxies develop only in those universes with an expansion rate close enough to the escape velocity that anisotropic modes do not dominate. As Earman (1987b) has emphasized, such an explanation is besides the point if the demonstration that isotropy is a necessary condition for our existence is precisely the source of puzzlement, and in any case Collins and Hawking (1973) do not stop there. They further invoke an ensemble of actually existing universes. Within this setting SAP sounds more like WAP: the surprise at fine-tuning is mitigated by a "selection effect," in that among physically possible worlds (initial data sets) the subset of worlds compatible with observers share various features such as isotropy. Even though this subset is of measure zero, it happens to be the only place where observers could be located within the ensemble.

Fans of the SAP such as Leslie (1989) have argued that the explanatory demand posed by fine-tuning can be answered either along these lines, via an actually existing Multiverse, or an appeal to Design. First I should emphasize that these arguments often exploit the ambiguity of the word "explanation": it is important to distinguish anthropic explanation from explanations related to the laws of a successful theory (however this notion of explanation is further characterized). This ambiguity encourages the idea that the availability of anthropic explanations in itself provides evidence for either the Multiverse or the Designer. But without substantial independently motivated additions, either idea is completely uninformative. We already knew that a universe like ours exists; *if* the Multiverse is developed within the context of a particular physical theory (such as chaotic inflation), there is some chance of deriving new results based on a physically motivated probability measure over the ensemble. Existing accounts are not encouraging in this regard: although inflationary models naturally accomodate a multiverse scenario, the mechanism introduced to produce variation in, say, values of the fundamental constants in different

regions of the multiverse is typically not independently motivated or constrained.[59] Likewise, opening the door to a theological explanation may do more than remove our puzzlement only if it is accompanied by a robust sense of what the Designer desired (as McMullin 1993 also emphasizes). Accounts of the Multiverse or Designer do not need to be informative in this sense to be appealing to metaphysicians or theologians, but they *do* need to be informative in this sense to fall within the realm of empirical inquiry.

In summary, the WAP cautions against taking evidence obtained for cosmological theories at face value, since our inherently parochial perspective acts as an indirect filter. Although handling this selection effect is a subtle issue, the subtlety derives from the complicated relationship between cosmological theories and our observational procedures, and not from any new anthropic addition to confirmation theory. The SAP offers a way of soothing worries regarding fine-tuning, but only in the sense of reducing puzzlement by appealing to extremely speculative physical accounts of the Multiverse or a theological account of Design.

[59] GTR without inflation can also accomodate a "multiverse," in which vastly separated different regions have varying degrees of homogeneity and isotropy.

Chapter 6

Explanations in Cosmology

Philosophers of science have often pondered cases of theory choice between empirically equivalent rival theories, such as the choice of Copernican astronomy over its Ptolemaic rival or Einstein's special relativity over (a portion of) Lorentz's electron theory. Kuhn (1970) famously described such cases of theory change as more like a "conversion experience" (p. 151) than rational deliberation, although in later work he retreated from the rhetoric of *Structure* and argued that—far from being irrational—scientists have often found a number of criteria for evaluating competing theories, namely judgements of accuracy, consistency, scope, simplicity, and fruitfulness.[1] While few historians or philosophers would argue that these criteria (and perhaps a few others) have not played a role in theory choice, empiricists insist on sharply dividing empirical adequacy, taken to be the only legitimate ground for believing a theory to be true, from the other criteria, which are "pragmatic virtues" of a theory relevant merely to its acceptance by working scientists. Van Fraassen argues for a sharp distinction between belief and acceptance of a theory combined with epistemic "voluntarism": while "freedom from conflict with evidence is the bottom line" regarding belief in a theory, the "quasi-political process of decision" by which scientists choose (or provisionally accept) a theory, partially based upon judgements of its pragmatic virtues and vices, is a rationally permitted "leap of faith".[2] For the constructive empiricist

[1] Kuhn (1977, p. 321-22) claims no originality for this collection of virtues, and similar lists crop up throughout the literature.

[2] The quotations are from van Fraassen (1985, p. 281 and 296), and these themes are developed at greater length in van Fraassen (1980, 1989).

the explanatory adequacy of a theory may set it apart from empirically equivalent rivals, but only in terms of grounds for accepting the theory—regarding grounds for belief there can be no distinction between empirically equivalent rivals. On this account there is no over-arching algorithmic Methodology which serves to guide theory choice in cases where empirical adequacy alone does not render a clear verdict.

In opposition to this empiricist position, several realists have argued that explanatory adequacy (characterized in different ways) can and must legitimately serve to justify belief in a theory, as opposed to mere acceptance of it. Broadly characterized, the realist goal is to find rationally compelling grounds for theory choice based on the criteria demoted to "pragmatic virtues" by empiricists such as van Fraassen. For example, Glymour (1980a, p. 31) acknowledges that if successful explanations are to provide reasons for belief in a theory above and beyond its empirical adequacy, they must do "something more to the phenomena, or say something more about the phenomena, than merely entail their description." Glymour's own suggestion (further developed in Glymour 1980b, 1985) is that explanations eliminate contingency, in that regularities follow as "mathematical necessities" once additional theoretical structure is introduced, and that they unify, in that diverse regularities are shown to exhibit a common pattern. The explanatory advantages usually claimed for inflationary cosmology over standard cosmology fit nicely with Glymour's suggestion: first, inflation is usually presented as a natural consequence of unifying particle physics with general relativity, and second, it apparently eliminates many of the "brute facts" which must be stipulated to hold as features of the initial conditions in standard cosmology. I will take up these two alleged explanatory advantages of inflation in turn: §1 below focuses on unification and §2 focuses on the "robustness argument," namely that inflation

offers a more robust explanation of the early universe's regularities in that it does not depend on "finely-tuned" initial conditions.

6.1 Unification

Contemporary physics abounds with talk of unification: proponents of the various GUTs of the 80s, and more recently of string theory, have emphasized their ability to account for the fundamental forces in a single framework as one of their most appealing features. Maudlin (1996) calls this strong emphasis on unification a "velvet revolution in the conception of the aim of physical theory," and argues that the desired degree of unification falls somewhere between mere consistency of different theories and complete unification (exemplified by historical cases such as special relativity, discussed below). Researchers in early universe cosmology often describe the connection between cosmology and fundamental particle physics in terms similar to Edward Kolb's:

> Nowhere is the inherent unity of science better illustrated than in the interplay between cosmology, the study of the largest things in the Universe, and particle physics, the study of the smallest things. (Kolb 1994, p. 362)

Successes in "deriving" various features of the universe from fundamental physics (e.g. baryogenesis) bolster such claims. For the last two decades the goal of ongoing research is often presented as completing this unified picture: the aim is to show that even more features of the universe can be explained as consequences of particle physics. As the first half of this dissertation illustrates, the commitment to this picture of how the "inner space – outer space" connection functions has been widely accepted and continues to shape research in the field. Early universe cosmology seems to be caught up in the "velvet revolution."

However, glowing reports of the unification achieved in early universe cosmology have been exaggerated, according to several critics:

> So far, there is little or no observational evidence which motivates the idea that particle physics is intimately related to cosmology. ... [I]t has yet to be established that particle physics is *relevant* to cosmology. (Zinkernagel 2002, original emphasis, p. 18)

Zinkernagel further emphasizes a point discussed above: as long as the inflaton potential is treated as a free function, rather than as a feature of a field identified in a particle physics theory, no substantive unification has been achieved. Earman and Mosterin (1999) emphasize this difficulty (and many others) in their sustained critical assessment of inflation. Torretti (2000) criticizes the standard big bang model on the grounds that it *lacks* any satisfying sense of unification: he argues that the incompatibility of quantum mechanics and general relativity undercuts even the status of the CMBR as evidence in favor of the big bang model—to say nothing of the more speculative applications of particle physics to even earlier times.

Below I will focus on the nature and force of unification arguments, first considering general accounts of unification and then turning to the case of cosmology. In order to function as a criterion of theory choice on a par with empirical adequacy, one would need to formulate a definition of "unification" clear enough to differentiate between competing theories, which also is not based on controversial commitments regarding the course of future theory. I will argue that the most detailed account of unification available (due to Kitcher) fails to satisfy this requirement. The case of early universe cosmology aptly illustrates the highly defeasible arguments regarding "unification" that occur in the development of novel theories.

Before going further I should clarify two different senses of unification at play. Unification in the first sense refers to a theory's ability to bring together a number of diverse phenomena

within a single theoretical framework. This synthesis may involve a reductive claim, in which two different entities (such as electromagnetic waves and light) are shown to be essentially the same thing. On the other hand, unification may bring together two different domains of phenomena previously regarded as distinct, such as the celestial and terrestrial. (Morrison 2000 calls the former "reductive unification" and the latter "synthetic unification.") Two of the paradigm cases of unification in modern physics, both due to Einstein, fall under this general sense of unification.[3] Very briefly, special relativity achieved a unification by changing the understanding of Lorentz invariance. In Lorentz's theory, the Lorentz invariance of laws governing matter follows from what Janssen (1997) calls the "generalized contraction hypothesis."[4] This hypothesis effectively guarantees the Lorentz invariance of the laws governing matter, despite their original formulation as Galilean invariant laws in Newtonian spacetime. The theory does not provide an explanation of why the generalized contraction hypothesis holds true. In contrast, special relativity unifies classical mechanics and electromagnetism in that the Lorentz invariance of the laws governing matter and fields both follow from the structure of Minkowski spacetime, and no generalized contraction hypothesis is needed. General relativity is also based on a remarkable unification, namely Einstein's realization that gravity and inertia have "the same essential nature [*wessensgleich*]." In Newtonian theory, gravitation resembles other forces in that it merely deflects particles from inertial trajectories, but bears no other direct link to inertial structure (i.e.,

[3] Both of the following examples have been discussed extensively in the literature; discussions with a similar focus on unification include Maudlin (1996); Janssen (1997); Morrison (2000). I have also benefitted from discussions with Michel Jannsen.

[4] Suppose that a given material system produces a field configuration in a state of rest (relative to the aether). Lorentz's theorem of corresponding states shows how to translate this field configuration into a set of "fictive fields" in a frame moving uniformly with respect to the aether (with a given velocity, say v). The generalized contraction hypothesis then holds that when put into a state of uniform motion (with a velocity v relative to the aether) the material system alters so that it produces the appropriate field configuration. As Janssen (1997) discusses at length, this generalized contraction hypothesis covers the familiar Lorentz-Fitzgerald contraction along with a number of other effects which explain the null results of all second-order aether drift experiments.

the affine structure of Newtonian spacetime). However, the odd fact that gravitational "charge," unlike electric charge, is precisely equal to the inertial mass sets gravitation apart from the other forces. For the other forces one can imagine varying the ratio of electric charge (for example) to inertial mass in order to locally distinguish the electrical force from inertial effects, but the equality of inertial and gravitational mass prevents this in the case of gravitation. General relativity eliminates the distinction between gravitational force and inertial structure, replacing them with a single inertial-gravitational structure embodied in the (dynamical) metric field. Inertia and gravitation are represented in the theory by a single structure, rather than two distinct structures that are mysteriously linked.

The case of GTR further illustrates a different sense of unification, understood as an intra-theoretic relationship: the theory of GTR combines the apparently incompatible theories of STR and classical graviational theory. This suggests a definition of unification in terms of a structural relationship between theories: a theory T unifies T_0 and T_1 (identified with classes of models O, O_0, O_1, respectively, as in the semantic conception of theories) if and only if for every model $M_0 \in O_0$, there are corresponding models $M_1 \in O_1$ and $M \in O$, such that both M_0 and M_1 can be embedded in M, and vice versa.[5] The two senses of unification are typically blurred because it is assumed that successfully combining theories results in unification in the first sense, as it undoubtedly did in several historical cases (such as the development of GTR). Thus we might add a further requirement that the theory M should not only recover the models of the two prior theories, but do so by achieving unification in the first sense. Perhaps this could be characterized by requiring that (in some sense to be made precise) M_0 and M_1 are more "finely tuned" than the

[5] This formulation presumes that all three theories can be formulated in a common framework to permit the assessment of these questions of embedding. This is by no means a trivial assumption, and I see no reason to expect that it will hold in all cases where one might wish to grant a more intuitive sense of "unifi cation."

model M (e.g., they require more arbitrarily set free parameters). Something like this account may capture the sense in which many of our current theories "unify" preceeding theories. But before turning to Kitcher's detailed attempt to formulate something along these lines, I will briefly discuss two motivations for formulating unified theories.

There is often a strong motivation for attempting to combine apparently inconsistent theories: one wishes to describe systems falling into the overlapping domains of applicability of the theories. In the case of early universe cosmology, Olive (1990) calls the application of QFT to the early universe "compulsory" since classical physics breaks down at the incredibly high temperatures the early universe is expected to reach. QFT predicts a number of novel phenomena at these temperatures, most notably phase transitions. As we saw in Chapter 3, in the Standard Model temperature dependent corrections to the effective potential of the Higgs field result in symmetry breaking phase transitions at high temperature. This latter claim is based on extrapolating the FLRW expansion backwards; while this should be taken with a large grain of salt, there are (as far as I know) currently no viable alternative scenarios in which the temperature reaches a low finite maximum.[6] Without a low limiting temperature, the early universe reaches temperatures and densities which can only be treated by quantum field theory. In addition, the strong gravitational fields in the early universe demand a general relativistic treatment. Thus, any theoretical treatment of this era requires taking both quantum field theory and general relativity into account (at least to some degree)—and in this modest sense requires some degree of unification.

[6]The electroweak and quark deconfinement phase transitions are expected to occur at roughly 10^{15} K and $10^{12} - 10^{13}$ K, respectively, and the early universe should reach these temperatures in the first fraction of a second after the big bang (at $t \approx 10^{-12}$ and $t \approx 10^{-5} - 10^{-6}$ seconds, respectively). String theory apparently predicts a finite limiting temperature, but it is on the order of 10^{31} K—much higher than the temperature where phase transitions are expected to occur (see Kolb and Turner 1990, §11.5).

The argument loses its force if one adopts the view that physics offers models of specific phenomena with only a very limited domain of applicability (see, in particular, Cartwright 1999). On this view, there would be no need to worry about the overlap in the domains of applicability of QFT and GTR, since neither actually extends as far as the physicists expect: the domain of physical theories does not extend beyond the carefully shielded "nomological machines" constructed to exemplify the regularities encoded in their laws. Although I do not have the space to counter this objection in any detail here, let me briefly sketch one response. Cartwright (1999)'s view is partially based on her argument that various purported laws are actually falsified when applied outside of the carefully restricted domain of a nomological machine. But as Smith (2002c) argues, these arguments in fact target the differential equations derived from the laws conjoined with various provisos (in Hempel's sense) rather than the laws themselves. While it is undoubtedly true that the laws of motion derived from Newtonian gravitational theory for the two body problem do not accurately describe the complicated real motions of the solar system, this gives reason to develop a more sophisticated account within the framework of Newtonian gravitational theory rather than to abandon that framework entirely. This is only one strand of Cartwright (1999)'s argument, and tugging at it does not by any means unravel her position. In any case, here I will set aside these concerns and adopt the view that physicists have some warrant for taking the domains of their theories to extend beyond "nomological machines."

Returning to the main line of argument, philosophers have argued that there are stronger motivations for developing unified theories than the need for a consistent theoretical description holding in overlapping domains. In particular, Friedman (1983) argues at length that unification enhances confirmability, roughly because a more unified theory can be confirmed by a wider

range of phenomena. On this view, unified theories are to be preferred because of their confirmatory advantage: a unified theory receives a larger confirmatory boost from the successful prediction of two (or more) phenomena than distinct theories predicting the same phenomena. Friedman uses the molecular model of gases (which I will call T) as an example. The molecular model entails predictions regarding a wide range of phenomena, including the behavior of gases, chemical phenomena, thermodynamics, etc. By way of contrast, a purely phenomenological description of gases (call this P) considered separately only entails predictions regarding the behavior of gases.[7] As a result of passing experimental tests in other areas, the probability assigned to T may exceed the prior probability assigned to the purely phenomological theory P; since T entails P, the requirement of probabilistic coherence implies that the probability of P should be increased as well.[8] The same holds true for the conjunction of two theories: successful predictions of a conjunction of two theories ($T_1 \wedge T_2$) may boost $Pr(T_1 \wedge T_2)$ past the probability initially assigned to either theory individually, but since the conjunction entails both theories the degree of belief assigned to each theory must be increased to equal $Pr(T_1 \wedge T_2)$.[9] Thus the conjunction of theories allows scientists to develop more well-tested theories. This assessment of the advantages of unification fits well with research work in particle cosmology: in several cases, the conjunction of cosmology and particle physics is subject to much stronger observational tests than those provided by earth-bound accelerator experiments alone.

---

[7] By "considered separately" I mean that it is *not* treated as a consequence of T; if P is regarded as a logical consequence of T, then it would be impossible for P to be less well-confirmed than T since P is a logical consequence of T (Friedman 1983, p. 243-244).

[8] Although Friedman does not endorse a Bayesian approach to confirmation theory, this point holds given that a theory T cannot have a lower "degree of confirmation" (Friedman's term) than a theory which entails it.

[9] Friedman further argues that the conjunction of theories over time is more accurately represented in terms of a *reduction* of observational structure to theoretical structure (i.e., a literal identification of elements of the two) rather than a *representation* (merely a claim that the former can be embedded in the latter), and that the former is available only to a scientific realist. I will not pursue this question in more detail, but see Morrison (2000); Kukla (1995) for criticisms of this account.

While this may establish an epistemic advantage of unification understood as conjuction of theories, this argument notably fails to establish an epistemic advantage for any form of unification *stronger* than mere conjunction. This argument in favor of unification is a substantial retreat from Friedman's first account of unification (cf. Kukla 1995), according to which unification provides greater understanding in the sense of "reducing the total number of independent phenomena that we have to accept as ultimate or given" (Friedman 1974, p. 15). The formal treatment of "independently acceptable phenomena" in Friedman's first account suffers from a number of deficiencies (see Salmon 1989, pp. 94-101, for a concise critique). Without an account of the epistemic advantages of such "true unification", Friedman's second account provides no reason to prefer a truly unified theory to a massive conjunction of theories with the same empirical consequences. One immediate response is to note that combining two theories is rarely a straightforward logical maneuver; instead, attempts to combine two (initially incompatible) theories often result in corrections and alterations of both theories, leading to a unified, corrected theory rather than a mere conjunction. However, an empiricist may respond that acceptance of such a new theory is based on its succssful new predictions and has nothing to do with its genealogy (this is essentially van Fraassen's response to Putnam's conjunction argument, van Fraassen 1980, pp. 83-87). In the next section I turn to a detailed attempt to capture the intuition behind Friedman's initial account.

6.1.1 Kitcher's Account of Unification

Kitcher's account of unification has a central role in his ambituous project of finding a *global* methodology applicable to all sciences at all times. In particular, according to Kitcher the

fundamental aim of scientific inquiry is to provide an economical, unified systematization of our beliefs (Kitcher 1989, p. 432):

> Science advances our understanding of nature by showing us how to derive descriptions of many phenomena, using the same patterns of derivation again and again, and, in demonstrating this, it teaches us how to reduce the number of types of facts we have to accept as ultimate (or brute).

Theory change in science can then be characterized as rational and progressive based on the extent to which it advances this fundamental aim.[10] In this section, I will argue that Kitcher's proposed means of comparing unifying power do not fulfill these grand ambitions. In particular, the proposed principle for assessing unifying power applies only to a narrow range of cases, and Kitcher provides no argument for his further "principle of optimism," which would insure that only this narrow range of cases really matters. In addition, Kitcher's focus on global systematizations of knowledge neglects the important question of how to manage a trade-off between unification within a restricted domain and overall, global unification. Both of these shortcomings indicate that Kitcher's account of comparitive unifying power does not offer a complete account of explanatory progress.

Kitcher characterizes deductive systematization in terms of sets of argument patterns used to relate various statements in K, our set of beliefs. The focus on argument patterns (rather than, say, axiomatizations of a theory) stems from Kitcher's view that understanding a theory requires "internalizing" appropriate arguments, along with his interest in biological cases where mathematical axiomatizations would be incredibly remote from practice. Argument patterns

[10] I should immediately acknowledge that Kitcher (1993)'s much richer account of scientific practice and progress answers many of the objections I have to this earlier account. In this new account Kitcher still holds that "the growth of scientific knowledge is governed by a principle of unification" (Kitcher 1993, pp.171-72), but he admits a variety of other factors influencing theory choice in his "compromise model" of rationality. However, I think it is still useful to consider the account in some detail to bring out the difficulties with characterizing unification.

consist of three distinct components (Kitcher 1989, pp. 432-34). First, *schematic sentences* are obtained by replacing some of the non-logical vocabulary of a sentence with dummy letters: e.g., a sentence from the "Simple Selection" argument pattern reads "The organisms in G are descendants of the members of an ancestral population G^* who inhabited an environment E" (Kitcher 1989, p. 444). The second component, *filling instructions*, tell us how to replace the dummy variables with names; in this case, the filling instructions specify what names of species may be substituted for G, G^* and so on. Finally, the *classification* specifies the deductive relationships between the various schematic sentences included in the argument pattern. The optimal systematization of our set of beliefs is called the "explanatory store" $E(K)$; an argument counts as explanatory precisely if it is a member of this set of optimal argument patterns. Kitcher devotes considerable effort to showing that this account avoids the standard counter-examples to Hempel's D-N model of explanation. Kitcher argues, for example, that a pattern explaining the height of an object in terms of the length of its shadow and the position of the sun has no place in $E(K)$, whereas a pattern explaining the length of a shadow in terms of an object's height and the sun's position does belong in $E(K)$. Whether Kitcher's account avoids the counterexamples to the D-N model is still a subject of active debate, which I will not enter into here; instead I will focus on the prior question of whether Kitcher provides an adequate account of optimal systematization.

The best systematization of our beliefs maximizes unifying power, understood as a balance between three competing virtues: the stringency of the argument patterns, the paucity of argument patterns, and the breadth of the conclusions derived. Stringency is imposed to avoid classifying a wide variety of vaguely similar arguments under one pattern with loose filling conditions; Paracelsus' use of the "microcosm - macrocosm" argument pattern, for example, would

have a low ranking in terms of stringency. Unifying power varies directly with the breadth of conclusions derived, inversely with the number of argument patterns, and directly with the stringency of the patterns. In simple cases where two sets of argument patterns are equally matched with respect to two of the virtues, it will be clear which to choose based on the third virtue. However, such examples bear little resemblance to the genuine cases of theory change Kitcher hopes to analyze. Since judgements of unifying power are the central engine of theory change and scientific progress on Kitcher's account, he requires a general account of how to measure unifying power in cases where the comparison requires a more careful balance of the three virtues.

Kitcher's more precise proposal of judgements of unifying power uses some additional formal apparatus. A set of argument patterns, U, generates a set of derivations, S, when the dummy variables in the schematic sentences are filled in with appropriate names as per the filling instructions. A particular derivation in the set S is acceptable relative to the set of background beliefs K just in case each step of the derivation is deductively valid, and the premises of the argument are elements of K. The conclusion set $C(S)$ is the set of statments that occur as conclusions for some derivation in S (Kitcher 1989, p. 434). Kitcher's proposed comparison principle is then:[11]

> *Comparison:* U has greater unifying power than U' if one (or both) of the following conditions is met:
> (C1) $C(S')$ is a subset of $C(S)$ (not necessarily proper), and there is a one-one map $f : U \rightarrow U'$ such that for each pattern p in U, p is at least as stringent as $f(p)$, and either f is an injection, or f is a surjection and there is at least one pattern p in U such that p is more stringent that $f(p)$;
> (C2) $C(S')$ is a proper subset of $C(S)$ and there is a one-one map $f : U \rightarrow U'$ such that for each p in U, p is at least as stringent as $f(p)$.

[11] I have corrected an obvious mistake in Kitcher's original formulation of (C): Kitcher defi nes the maps f, f' in terms of S, S' rather than U, U'. However, these maps relate argument patterns (the elements of the sets U, U') and *not* derivations, which are the elements of the sets S, S' (Kitcher 1989, pp. 478-79).

As Kitcher notes, *Comparison* introduces an asymmetric, transitive relation on sets of argument patterns satisfying the conditions. Considering (C2) first, the first clause guarantees that U captures a broader set of conclusions than its competitor U'. The existence of f implies that the number of argument patterns in U' is greater than or equal to the number in U, and the last clause requires that each pattern in U is mapped into an equally stringent or less stringent counterpart. Thus (C2) applies when U has the advantage of capturing a broader set of conclusions, and fares as well as U' with regards to the other two virtues. The condition (C1), on the other hand, applies when U has an advantage with regards to either the number of patterns (when f is an injection) or stringency (the last clause), and fares as well or better than U' with regards to the other two virtues.[12]

Rather than providing a general principle applicable to difficult cases, this principle merely formalizes how to handle the simple cases. Two examples due to Daniel Steel (Steel 2002) illustrate the limitations of this principle. Suppose that U' contains a single argument pattern. One would expect that U' would lack stringency, lead to a small set of conclusions, or both, and that this should be reflected in judgements of unifying power. However, both (C1) and (C2) require the existence of a one-one map f from U to U', so neither of these conditions can be met if U includes more than one argument pattern (as it surely does). Thus *Comparison* does not allow one to conclude that U has greater unifying power than U', despite the clear intuition that the latter's victory in achieving a small number of argument patterns comes at too great a cost. Similarly, construct a new argument pattern U' by simply "tacking on" an argument pattern p to the orignal set U, such that this pattern leads to the derivation of a new conclusion not included

[12] Kitcher further clarifies comparisons of stringency by introducing two principles to characterize relative stringency. Briefly, stringency is reflected in both the "tightness" of the filling instructions and the "tightness" of the classification of the logical structure of the argument. Since my argument is not directly related to these principles, I will not discuss them further here.

in $C(S)$. Again, (C1) and (C2) would both fail since $C(S')$ is *not* a subset of $C(S)$; but U' clearly suffers in comparison to U as a result of adding an epicycle with such slim payoff. These extreme cases both illustrate that *Comparison* fails to apply when U fares better on two of the virtues despite failing in comparison to U' with regards to the third. *Comparison* can be used to define a strict partial ordering—degree of "unifying power"—on sets of argument patterns similar enough to satisfy (C1) or (C2), but this ordering does not extend to cases of genuine trade-offs between competing virtues. In these two simple cases we can clearly order U and U' without the aid of *Comparison*, but in more complicated cases these judgments would presumably not be so obvious. Kitcher owes us either an extension of his principle that defines a strict partial ordering over a broader range of sets of argument patterns, or an argument that such an extension is unnecessary.

Kitcher introduces an aptly named principle meant to assure that *Comparison* will be sufficient for judging unification without such an extension (Kitcher 1989, p. 478):

> *Optimsim:* Let U, U' be sets of patterns. Then there is a set of patterns U^* such that:
> (a) There are one-one maps $f : U^* \rightarrow U$, and $f' : U^* \rightarrow U'$ (injections or surjections) such that for each pattern p in U^*, p is at least as stringent as $f(p)$ and at least as stringent as $f'(p)$;
> (b) The consequence sets $C(S), C(S')$ are both subsets (though not necessarily proper) of $C(S^*)$

If this principle holds, then a trade-off between unifying virtues can be avoided by constructing a set of patterns which maximizes the virtues of the competing sets. Kitcher admits that he does not know whether this principle, or some more restricted version of it, is true, and offers no argument in its favor (Kitcher 1989, p. 478). I do not share Kitcher's optimism. First, in the historical cases Kitcher ultimately hopes to analyze, this principle would insure that scientists can *always*

reconcile two competing approaches without losing the unifying virtues of either competitor. This is far too strong, and Kitcher does not argue that this has actually happened in any of the historical cases he discusses. But more fundamentally, in some cases when *Comparison* does not apply we have clear intuitions that one set of patterns should be rejected in favor of another. In both of the simple examples of the previous paragraph, it would be perverse to look for a set of argument patterns U^* combining U and the clearly inferior U' rather than simply selecting U as the more unified set of argument patterns.

Two additional critical points undercut hopes of extending Kitcher's principle to a more broadly applicable ordering. First, Kitcher assumes that a given set of scientific ideas can be formulated in terms of a "canonical set" of explanatory patterns. But is it really plausible that, say, physicists who generally agree about the "content" of non-relativistic quantum mechanics would all formulate the same Kitcher-style explanatory patterns to systematize the theory? This is not an idle worry since *differences* in these formulations would be reflected in differing judgements of unifying power related to rival theories. In the case of non-relativistic quantum mechanics, would there be a single argument pattern for "One-Dimensional Problems" with dummy variables allowing different choices for the potential, or more specific patterns for "Potential Barrier," "Simple Harmonic Oscillator," and so on? Syntactical choices of this sort would lead to an overall paucity or plurality of argument patterns in the canonical set.

The second, more important critical point is that Kitcher's approach does not provide the resources to handle cases in which unification is achieved at the cost of introducing conflicts with other widely accepted theories (cf. Koertge 1992). Copernican astronomy arguably achieved a more unified description of planetary motion than Ptolemaic astronomy in that, for example, it accounted for the correlations between the sun's motion and the motion of the planets without

requiring independent assumptions relating them. But this "local" reformulation of astronomy came at the cost of abandoning Aristotelean physics: Copernicus did not provide a replacement for Aristotelean physics that would apply to a moving Earth, and offered relatively weak arguments intended to mitigate the conflict. Kitcher's account of unification generally favors the "globally" more satisfying set of argument patterns, whereas historically *local* assessments of the advantages of Copernican astronomy were one of the main attractions of the theory.

In sum, Kitcher's account is well equipped to handle rational reconstructions of historical cases by exhibiting the explanatory advantages of successful theories. Yet it flounders in providing a criterion of unification applicable in judging the merits of competing theories, in that there is no reason to expect the "principle of optimism" to hold. Retreating from the claim that unification is the central engine driving the progress of science does not mean that it cannot play a role in theory choice. In particular, Friedman's point that unification enhances testability holds regardless of the difficulties with cashing out any stronger sense of unification in terms of "reduction of brute facts." Indeed, in Kitcher's own application of his "compromise model" of rationality to the Copernican revolution (Kitcher 1993, pp. 205-211) unifying power is limited to one among many competing considerations in assessing theories.

6.1.2 Unification in Cosmology

As we saw in Chapter 3, throughout the 70s a number of researchers explored the consequences of combining QFT with cosmology. Given the speculative nature of QFT at high temperatures, a large range of models was explored with a wide variety of novel consequences. The discovery of inflation brought about a dramatic shift: research focused on the subset of models incorporating a scalar field with the right properties to produce inflation. The conviction that

inflation *must* be a consequence of the conjunction of QFT and cosmology seems to be based on two distinct arguments. First, in Guth's original model inflation was driven by the Higgs field responsible for symmetry breaking in an $SU(5)$ GUT, and the properties of the Higgs field and its potential appeared to be "natural" in the context of this theory. Thus inflation seemed to be a straightforward consequence of high energy physics applied to cosmology. Second, inflationary cosmology ties together several large scale features of the universe which otherwise bear no relation to each other. In standard big bang cosmology, the uniformity of the universe on large scales, the flatness of the universe, and the presence of small scale fluctuations in the density of matter are all independent features of cosmological models. One can construct cosmological models with large scale uniformity but a different spectrum of small scale density perturbations. Both are features of initial conditions, and can be "tuned" virtually at will. Thus inflation apparently achieves an important unification in the description of the early universe.

Balanced against this case in favor of inflation are several fundamental obstacles. Roughly put, several critics of inflation suspect that conceptual conflicts hidden in the loose amalgam of particle physics and general relativity making up the theory may undercut its positive results. Early calculations of the temperature dependence of the effective potential (see Chapter 3) were carried out in standard (finite temperature) QFT without taking cosmological effects—other than the increase of temperature—into account at all. This could hardly be called a "unification," even in the sense of a conjunction of theories. Further work incorporated general relativistic effects by replacing the flat background Minkowski spacetime of traditional QFT with one of the dynamically evolving FLRW models; this approach is a 0^{th} order approximation to a full theory of quantum gravity in that field theory is formulated over a fixed, non-dynamical background spacetime. Early universe cosmologists cannot rest easy with only a 0^{th} order approximation,

since inflation results from including the inflaton field as a source in Einstein's field equations. The semi-classical approximation (a 1^{st} order approximation to quantum gravity, so to speak) incorporates the effect of the quantum fields on the spacetime metric by replacing the classical stress-energy tensor T_{ab} with the expectation value $\langle\phi|T_{ab}|\phi\rangle$ of the (renormalized) stress-energy tensor for the quantum fields.[13] In other words, in the semi-classical approximation one treats the matter fields quantum mechanically and the metric classically. Aside from more general reasons for dissatisfaction with such a "half-and-half" theory (Callender and Huggett 2001; Arageorgis 1995), there are several problems related to inflation which depend upon strong assumptions regarding the complete theory.

The first problem relates to the *onset* of inflation and the so-called "cosmic no-hair" conjecture. The no-hair theorems aim to show that a stage of exponential expansion erases all the wrinkles of the early universe; more precisely, one hopes to show that a fairly general initial state rapidly approaches a (locally) de Sitter solution undergoing inflationary expansion (see Appendix A for a definition of "locally de Sitter"). The intuition behind these theorems is that all contributions to the energy density other than Λ decay with time in an expanding universe, so eventually Λ should dominate the expansion. With a convincing no-hair theorem in hand, an inflationary cosmologist could ignore the details regarding inflation's onset: regardless of the precise characteristics of the initial conditions or how inflation begins, the output would reliably be an inflating universe.

[13] Inflation is often treated as the purely classical dynamics of a scalar field (see, e.g., Kolb and Turner 1990, Chapter 8) with a classical T_{ab}; however, this only accounts for the first term of a perturbation expansion in powers of the coupling constant. The higher order quantum corrections will be small if the inflaton field is only weakly self-coupled and weakly coupled to other fields, and in fact the observational constraint on the magnitude of post-inflationary density perturbations implies that the inflaton self coupling is on the order of 10^{-12}. Although the weakness of the inflaton's self-coupling justifies neglecting higher order contributions, this is one of the main "fine-tuning" problems of inflation.

A no-hair theorem requires demonstrating two results: first, that an effective cosmological constant gets rid of anisotropies and inhomogeneities, and second, that for generic (or a suitably large set of) initial conditions an inflationary stage will occur which mimics a cosmological constant term. Attempts to prove the second claim more clearly illustrate the limitations of a "half-and-half" theory. Proofs of the first claim (briefly reviewed in Appendix A.5) involve only classical general relativity (in essence, determining the effect of a transient Λ), but the argument that an early universe phase transition produces such an effective Λ relies on a combination of flat space QFT with general relativity. Recall the familiar stress-energy tensor for a scalar field,

$$T_{ab} = \nabla_a\phi\nabla_b\phi - \frac{1}{2}g_{ab}g^{cd}\nabla_c\nabla_d\phi - g_{ab}V(\phi). \tag{6.1}$$

During slow roll in the "new inflation" scenario, the inflaton mimics Λ only if all but the last term are neglible, so that $T_{ab} \approx -g_{ab}V(\phi)$. But approximate mimicry may not be sufficient to relate inflation to the proofs of the first claim, which depend on imposing energy conditions which do not necessarily hold if the gradient terms are non-zero.[14] In addition, in order for inflation to occur the gradient terms must be negligible in a region larger than the horizon radius at the time of inflation's onset. In other words, the picture of an inflationary stage emerging from a chaotic, inhomogeneous initial state is inaccurate; the inflaton field and spacetime must be homogeneous in a super-horizon patch to produce an inflationary stage.[15] A more fundamental problems arises

[14] The no-hair theorems of Wald and others apply if the strong and dominant energy conditions hold for $T_{ab} + V(\phi)g_{ab}$ (i.e., stress-energy tensor excluding Λ). This assumption is essential to the theorems, and the same conclusion does not follow if it is dropped (see Goldwirth and Piran 1992, for discussion and references).

[15] Both of these points have been emphasized in the literature; see, in particular, Vachaspati and Trodden (2000) and Goldwirth and Piran (1992) for more detailed discussions. Note, however, that similar complaints do not apply to Linde's chaotic inflation—and the hope of avoiding such requirements seems to be one of his main motivations in developing chaotic inflation.

in justifying the use of the effective potential $V(\phi)$ calculated using flat space QFT for very general initial conditions. Extrapolations are expected to break down at the Planck scale due to gravitational effects, but even at sub-Planckian energy densities several effects may invalidate the flat space calculations. In anisotropic models the anisotropy energy density may reach the Planck scale well after the Planck time, and more generally negative curvature can suppress the expected phase transition.[16] In practice, detailed calculations in various inflationary models usually avoid these problems by focusing on the later stages of inflation, when the field theory calculations can be performed in a background de Sitter spacetime. But the no-hair theorems play a crucial role in justifying this approach, and I see no way to prove the second component of a no-hair theorem without some assurance that gravitational effects do not invalidate the flat space effective potential. The only alternative is to abandon any pretense of calculating rather than simply stipulating the form of the potential.

The second general problem with combining QFT and GTR important to inflationary cosmology has been called a crisis in contemporary physics by a number of theorists (including, prominently, Weinberg 1989). The problem goes under the name "the cosmological constant problem"; the main issue is how to understand a stress-energy tensor in GTR incorporating quantum fields as sources.[17] Calculations of the vacuum energy density in quantum field theory yield incredible results: according to the calculation reviewed in Appendix B.2, the vacuum energy density of the free electromagnetic field is $\langle \rho_{vac} \rangle \approx 10^{46}\ erg/cm^3$. For the sake of comparison, this is roughly 10 orders of magnitude greater than the mass-energy density of a neutron star

[16] See, e.g., Hu (1986) for an early criticism of the reliance on flat space QFT. Hu argues that curvature anisotropies and other dynamical effects render the usual effective potential inapplicable to the general case of a curved spacetime.

[17] I have benefited from conversations with John Earman regarding this topic; cf. his Earman (2001). I am also drawing on the very careful, comprehensive discussion in Rugh and Zinkernagel (2001).

(roughly $10^{36}\ erg/cm^3$)! Why treat this vacuum energy as anything other than an artifact of the formalism, to be gotten rid of via some renormalization procedure? Physicists have often argued that the Casimir effect demonstrates the reality of zero-point energy (Weinberg 1989, p. 3, for example). The availability of alternative derivations of the Casimir effect which do not appeal to vacuum energy density renders this argument inconclusive.[18] In these alternative derivations, Casimir's calculation of the force between two attracting plates (treated as a consequence of the lowering of the vacuum energy density between the plates) is replaced with a calculation in terms of the source fields of the plates themselves. Although I will not pursue the question further here, several philosophers have begun exploring the viability of interpretations of QFT which do away with vacuum energy density. Almost all of the extensive evidence supporting QFT has no direct bearing on this question, since the vacuum energy is irrelevant to the calculation of the S-matrix amplitudes compared to experimental results.

The vacuum energy leads to a crisis when $\langle \rho_{vac} \rangle$ is included as a source term in EFE. In flat space QFT, the form of the stress-energy tensor for the vacuum follows from the Poincaré invariance of the vacuum state $|0\rangle$. The physical properties of the vacuum state are also invariant; in particular, $\langle 0|T_{ab}|0\rangle$ is Poincaré invariant. Since the only Poincaré invariant rank two tensor is the Minkowski metric, $\langle 0|T_{ab}|0\rangle = \text{constant} \times \eta_{ab}$, which is the same form as a cosmological constant term. This argument is usually stated in one line, and immediately generalized to GTR (perhaps with an invocation of the $\eta_{ab} \rightarrow g_{ab}$, normal derivative $\rightarrow$ covariant derivative "rule"). If this generalization holds, then the vacuum energy density should lead to an effective cosmological constant; a comparison between the vacuum energy density calculated in QFT and observational limits on Λ reveals an incredible discrepancy of some 120 orders of magnitude!

[18] See, in particular, the discussion in Rugh and Zinkernagel (2001).

If this stunning conflict with observation does not provide enough evidence that something has gone terribly wrong, two important theoretical difficulties undercut this generalization of the flat space properties of the vacuum to GTR. First, general relativistic spacetimes generally lack the symmetries used to identify the vacuum state and to constrain the form of its stress-energy tensor. In stationary, globally hyperbolic spacetimes the construction of QFT on curved spacetimes is similar to that on Minkowski spacetime, in that a unique vacuum state can still be identified; but uniqueness fails for more general spacetimes.[19] Since the FLRW models (and presumably whatever model describes the universe) are not stationary, the "vacuum state" cannot be uniquely identified using spacetime symmetries. Second, even defining $\langle T_{ab} \rangle$ and including it as a source term in QFT on curved spacetimes requires a regularization procedure. Trying to define a "quantum" T_{ab} by replacing the classical ϕ's in eqn. (6.1) with quantum fields yields nonsense: the quantum fields are operator-valued distributions and T_{ab} includes products of these fields, but there is no way to define a product of distributions. Thus the quadratic terms in T_{ab} give rise to divergences; all of the various formal methods of "regularizing" $\langle T_{ab} \rangle$ extract a finite, reasonable quantity from these divergent expressions by subtracting off the vacuum energy (see Birrell and Davies 1982, Chapter 6, for a review of these techniques). The axiomatic approach developed by Wald (see Wald 1994, §4.6) includes an axiom that $\langle T_{ab} \rangle = 0$ in Minkowski spacetime, and in more general spacetimes $\langle T_{ab} \rangle$ is calculated using a point-splitting prescription which subtracts off a term roughly corresponding to the vacuum energy.[20] The physicists worried about

[19] A stationary spacetime possesses a timelike Killing field, i.e. a vector field ξ^a such that $\nabla_a \xi_b + \nabla_b \xi_a = 0$. Very roughly, the timelike Killing field can be used to uniquely fix the subspace of positive frequency solutions to the Klein-Gordon equation used in constructing the Hilbert space of states, and the vacuum state in the associated Fock space is the ground state of the Hamiltonian. For details of the construction of QFT in a stationary spacetime, see Wald (1994), §4.3.

[20] More precisely, the quadratic terms are defined using the "point-splitting prescription," $\langle \phi^2(x) \rangle := \lim_{x \to x'} [\langle \phi(x) \phi(x') \rangle - H(x, x')]$; in Minkowski spacetime, $H(x, x')$ is given by $\langle 0 | \phi(x) \phi(x') | 0 \rangle$ but in general spacetimes $H(x, x')$ is constructed so that the axioms are satisfied.

the cosmological constant problem apparently think of the vacuum energy density as something other than a "meaningless infinite quantity" (Birrell and Davies 1982, p. 150); but it is far from clear how to even define $\langle T_{ab} \rangle$, or incorporate it as a source term in EFE, without employing a regularization method that treats it as such.

The cosmological constant problem indicates that there is still much that we do not understand regarding the quantum vacuum. This problem has also been described as the Achilles heel of inflationary cosmology; in his review of the first workshop devoted to inflation, Frank Wilczek commented (cf. Kolb and Turner 1990, p. 314):

> It is surely an act of cosmic *chutzpah* to use this dismal theoretical failure [in understanding Λ] as a base for erecting theoretical superstructures, but of course this is exactly what is done in current inflationary models (Hawking et al. 1983, p. 476, original emphasis).

It is possible that setting the overall cosmological constant to zero (via some regularization procedure, or through the introduction of new physics) will not eliminate the shifts in vacuum energy which are thought to produce an inflationary stage. However, it is also possible that reconceptualizing the quantum vacuum and the renormalized stress-energy tensor will leave no room for Λ—or for an effective Λ driving an inflationary stage.

The third general problem is the enduring mystery: who is the inflaton? Following the failure of Guth's model, the inflaton has been identified with a number of different scalar fields postulated to exist in extensions of the Standard Model. As described in more detail in Chapters 4 and 7, the observed magnitude of density perturbations places tight constraints on the inflaton potential and gave rise to a new fine-tuning problem, namely why does the inflaton potential have these features? One response is to hope that the final theory will incorporate a scalar field which "naturally" has the right features to drive inflation. There are currently a plethora of models (over

50!) which embed inflation in various underlying particle physics theories, and the hope that one of these will emerge as the canonical realization of inflation is optimistic, but not unreasonable.

6.2 Robustness and Causality

Salmon (1998) includes an interesting discussion of the abuse of (alleged) causal principles by astrophysicists. The case involves inferring the size of an object (such as a quasar) from observed variations in its spectra or luminosity: since the early 60s, astrophysicists have applied the "$c\Delta t$" criterion to determine the size of a variable source. For a source varying on a time scale Δt, one infers that the size of the object cannot be larger than $c\Delta t$, on the grounds that the variability must be due to a signal propagating through the emitting object at a speed $\leq c$. The $c\Delta t$ criterion is repeatedly used in this way (Salmon lists several examples), and it is often presented as a direct consequence of special relativity.

Despite its popularity, this argument is egregiously false. One of the simplest counterexamples involves a large spherical shell of gas surrounding a compact source. Signals (such as a burst of radiation) from the central source could stimulate emission from the shell with a very short period of variation; for sufficiently small Δt the size of the sphere inferred from the $c\Delta t$ criteria is much smaller than the actual sphere. But there is clearly no conflict with special (or general) relativity: the signal propagates outward from the central source at or below the speed of light. It turns out that the currently accepted models of quasars *do* satisfy the $c\Delta t$ criterion. However, as Salmon argues, appeals to "causal" arguments obscure the real issues involved in assessing competing models of quasars and historically limited research to a class of models which satisfy the (unjustified) criterion. Other considerations (such as whether the model can

reproduce the spectrum of a quasar), rather than any straightforward causal argument, led to the abandonment of various models which also violated the $c\Delta t$ criterion.

Presentations of inflation often include a "causal argument" in its favor: an inflationary stage is the *only* causal mechanism which solves the horizon problem and produces appropriate density perturbations. Salmon's case illustrates that scientists are prone to misrepresent plausibility arguments as (presumably more authoritative) causality arguments, and below I will argue that this is true for the case of inflation as well. The causal arguments made by inflationary cosmologists do not involve an obvious fallacy, as in the case of the $c\Delta t$ criterion, but disentangling the implications of the presence of particle horizons in the early universe requires care (cf. Earman 1995). The horizon problem is often also called the "causality problem" or the "causality paradox" based on the intuition that horizons prevent causal explanations of the early universe's uniformity. In the next section I formulate and briefly discuss the problems with Reichenbach's principle of the common cause, before concluding that an alternative principle more clearly illustrates the intuitions underlying the horizon problem. As I discuss in §6.2.2, the horizon problem stems from a conflict between the assumption that physical laws do not place constraints on relatively spacelike events and the highly correlated, global uniformity of the early universe revealed by the CMBR observations. In the absence of lawlike constraints on initial data, cosmological models can apparently only accomodate this uniformity by stipulating that "special" initial conditions held. Cosmologists generally favor robust explanations which do not require such carefully set initial conditions, and in §6.2.3 I will formulate a criterion for "robustness" of explanation and examine when this criterion fails to apply. In particular, inflation offers a more robust explanation of the early universe's uniformity only if two important assumptions about Planck-scale physics hold true: first, that there are no lawlike constraints on

initial data which render inflation superfluous, and second, that high energy physics "naturally" incorporates a scalar field with a potential appropriate to drive inflation.

6.2.1 Reichenbach's Principle of the Common Cause

Reichenbach was the first to give a precise formulation of the intuition that a common cause underlies improbable correlations between events. For two simultaneous random events A and B with a positive correlation, i.e. $Pr(A\&B) > Pr(A)Pr(B)$, Reichenbach's principle asserts that there must exist a common cause C in the past such that (Reichenbach 1956, pp. 158-59):[21]

$$Pr(A \wedge B|C) = Pr(A|C)Pr(B|C) \tag{6.2}$$

$$Pr(A \wedge B|\neg C) = Pr(A|\neg C)Pr(B|\neg C) \tag{6.3}$$

$$Pr(A|C) > Pr(A|\neg C) \quad \text{and} \quad Pr(B|C) > Pr(B|\neg C) \tag{6.4}$$

Together these equations imply that $Pr(A \wedge B) > Pr(A)Pr(B)$. The first two equations show that the common cause "screens off" the correlation; the first equation implies that $Pr(B|C) = Pr(B|A \wedge C)$, which more clearly expresses the fact that C renders A statistically irrelevant to B (as does $\neg C$). The final equation merely defines "cause" as that which makes its effect more likely than it would be in C's absence (otherwise, the PCC is symmetric between C and $\neg C$). Reichenbach attributed physical content to the PCC in two senses: first, he thought that it clarified the concept of causation in an indeterministic universe, and second, he thought that

[21] The conditional probability is defined as $Pr(A|B) = Pr(A \wedge B)/Pr(B)$, and I will assume that none of the probabilities are zero. Here I am assuming that Reichenbach intended to rule out direct causal links between A and B; alternatively, sometimes the PCC is formulated as stating that either A and B are directly causally linked, or there is a common cause C.

it would provide a way to distinguish the past from the future since C lies to the past of A, B. The principle has also been presented as a methodological "categorical imperative": observed correlations between simultaneous events must be explained in terms of a common cause, and a hypothesis incorporating these common causes (even if they are unobservable) is to be preferred to a "separate cause" hypothesis. Every one of these claims has been the source of vigorous debate (see, for example Arntzenius 1993, 1999; Salmon 1984; Sober 1994), but I will focus mainly on the methodological issues in the following.

Before turning to criticisms of the PCC, I will first introduce some necessary refinements of the principle. As formulated above the PCC applies to events which do or do not occur, but it can be generalized to (discrete or continuous) observables. On the assumption that the instantaneous physical state of the system determines whether or not an event occurs, an event is just a partition of the phase space Γ of the system into two cells ("C occurs" and "C does not occur"). But a more general property—such as whether an observable A takes a value lying in some interval $\Delta \subset \mathbb{R}$—can also be uniquely associated with a subset of Γ.[22] Following Uffink (1999), the PCC modified to cover such observables requires that for n correlated quantities $\{A_i\}$, there exists a set of m mutually exclusive quantities $\{C_i\}$ such that the joint distribution factorizes:[23]

$$Pr(A_1 \wedge A_2 ... \wedge A_n) = \sum_{i=1}^{m} Pr(A_1|C_i)...Pr(A_n|C_i)Pr(C_i) \tag{6.5}$$

[22] In classical mechanics the observables are real-valued functions on phase space $f_A : \Gamma \to \mathbb{R}$, and the subset of phase space corresponding to a measured value of A lying in the interval Δ is just the inverse image ($f_A^{-1}(\Delta)$).

[23] This formulation treats $\{C_i\}$ neutrally as causal factors, but clearly one can add the analogues of the third equation above to stipulate that they deserve to be called causes.

This reformulation saves the PCC from counterexamples where a single event C does not screen off the correlated events, even though a combination of events (or more generally, an observable) does. A second refinement is to associate the observables with local physical states in regions of spacetime: suppose A represents the physical state in a spacetime region U and B the state in V, then C should represent a physical state associated with the region lying in the overlap of the past light cones of the two regions, i.e. within $J^-(V) \cap J^-(U)$ (the region from which a signal travelling at or below the speed of light could reach both U and V, see Appendix A.4).

The PCC clearly fails to hold as a methodological principle of unrestricted generality. In non-relativistic quantum mechanics, states such as the singlet state of two spin-$1/2$ particles exhibit correlations between spacelike-separated events (e.g., measurements of the spins of the two particles). These correlations generally cannot be screened off by local common causes which reproduce quantum statistics (see, for example Elby 1992). These features do not disappear in QFT: spacelike correlations without a prior screener-off are endemic in QFT, despite the fact that the theory is constructed to satisfy relativistic causality constraints (i.e., Lorentz invariance).

Defenders of the PCC such as Salmon have claimed that it may still apply in cases unaffected by quantum mechanical weirdness, but I will briefly review three cases in which even this more modest position does not hold. The first case involves situations where a "separate cause" explanation seems much more plausible; for example, the correlation between bread prices in Britain and the level of sea water in Venice (which have both been increasing monotonically) does not seem to cry out for a common cause explanation (this example is due to Sober 1994, 2001). More interesting examples drawn from evolutionary biology illustrate this point (Sober

1994): in ongoing debates regarding phylogenetic inferences based on similarities and differences between species, correlation between traits is not taken as decisive evidence for propinquity of descent. The comparison between a "common cause" explanation of a similar trait and a "separate cause" explanation is based on a number of background assumptions, and the PCC has not been used in this debate as a methodological silver bullet to kill the latter option. Sober argues that whatever force the PCC has derives from the more general methodological claim, the "likelihood principle", which states that evidence E favors a "common cause" theory T_{cc} over a "separate cause" theory when $Pr(E|T_{cc}) > Pr(E|T_{sc})$. The PCC appears to have general validity only because its defenders focus on cases where T_{cc} has higher probability than a competing theory that postulates separate causes.

Second, in a deterministic universe correlations are screened off by quantities both to the past *and* to the future of the events to be explained (Arntzenius 1990). In other words, the PCC fails if we allow *any* subspace of Γ to count as a potential screener off, since there will in general be screening off "common effects".[24] Without some restriction on what counts as a common cause (or effect), we will count very complicated properties as common causes; and this appears to violate the spirit if not the letter of the PCC. For example, in a system of gas molecules governed by Newtonian mechanics, the screener off for a collision between two individual gas molecules at a specific time will generally be a complicated fact involving the positions and momenta of a large number of other molecules. Arntzenius (1999) discusses and ultimately rejects two ways of refining the notion of common cause so that some version of the PCC holds:

[24] Uffi nk (1999) points out this conclusion only follows if we take the PCC to *require* the absence of a screening off common effect; if we drop this additional requirement, Arntzenius's argument shows that a temporally symmetric version of the PCC trivially holds in a determinisitic universe. Here "deterministic" means that the complete instantaneous state of the universe at any time uniquely fi xes the complete instantaneous state at any other time.

first, to require that they correspond to macroscopic quantities or, second, to local quantities. The first cannot hold if macroscopic quantities have microscopic common causes. The second option is more appealing (aside from conflicts with quantum mechanics), but there is a general class of counterexamples: equilibrium correlations. Systems in thermodynamic equilibrium generally exhibit correlated quantities (such as temperature in different regions) which are not directly causally related, and are also not determined by some local quantity at an earlier time.

Third, modern physics incorporates several laws which restrict quantities in relatively spacelike regions (cf. Arntzenius 1999). Two of Maxwell's equations are hyperbolic differential equations governing the evolution of the fields, but the other two are elliptic constraint equations ($\nabla \cdot \mathbf{B} = 0$ and $\nabla \cdot \mathbf{E} = 4\pi\rho$, where ρ is the charge density) which must be satisfied by initial data specified on a spacelike Cauchy surface. The second of these differential equations can be integrated over a finite region to give Gauss's law, which fixes the E field on a surface in terms of the charge enclosed by the surface—evaluated "instantaneously" on a spacelike slice. These "laws of co-existence" do not conflict with relativity, but since they lead to correlated quantities without prior screener-offs they violate the PCC.

A very different objection to applying the PCC in the relativistic context arises from its exclusive focus on the common causal past of two regions (cf. Earman 1995). Consider two relatively spacelike regions S, S' with overlapping causal pasts (I will call this overlap $O = J^-(S) \cap J^-(S') \neq \emptyset$). The PCC would have us look to this region O to find the cause of any correlations between the two regions. However, in general conditions *outside* of O will have an influence on the states in S, S'. Consider the intersection of O with a Cauchy surface Σ (call this U), and define V to be the intersection of the *union* of the causal pasts of S, S' with Σ. In general S and S' are not subsets of the future domain of dependence of U, $D^+(U)$. $D^+(U)$ is the set

of points p such that all past-directed, inextendible timelike or null curves through p intersect U; more intuitively, all of the causal influences acting on the region $D^+(U)$ are reflected in the initial conditions set on the surface U, since by definition no causal curve can reach this region without intersecting U. Perhaps in some situations one could vindicate the PCC by showing that holding conditions fixed on $V - U$ while varying those on U does indeed yield the desired correlations, but in general, since S and S' are subsets of $D^+(V)$, one would expect that the conditions on V should be taken into account in considering the states in S, S'.

Penrose and Percival (1962) formulated a principle called the "law of conditional independence" (LCI) that avoids many of the shortcomings of Reichenbach's PCC.[25] The LCI is meant to capture the idea that a completely isolated system should exhibit no correlations with other regions of the universe. It is analogous to the Sommerfeld radiation condition, which requires that the contribution of source-free radiation to the surface integral in the Kirchoff integral representation of a solution to Maxwell's field equations vanishes as the region of integration is extended to infinity.[26] Imposing this condition eliminates "source-free" radiation traced to correlations at infinity rather than the motion of charged particles. Similarly, the LCI rules out correlations due to interactions coming in from infinity, and like the Sommerfeld condition the LCI is explicitly time-asymmetric. Define C to be any region that divides the union of the causal pasts of S and S' (i.e., $J^-(S) \cup J^-(S')$) into two pieces, one containing S and the other S' (see Figure 6.1 for an example of such a region).[27] Penrose and Percival (1962) then require

[25] Penrose and Percival (1962) cite Reichenbach's PCC along with his analysis of branch systems as one of the motivations for their work. This interesting proposal has been overlooked by philosophers of science until very recently; Uffi nk (1999) corrects this trend by clearly emphasizing the advantages of the LCI over other formulations of causal principles (cf. Arntzenius 1999).

[26] See, in particular, Ellis and Sciama (1972) for a clear discussion of the Sommerfeld radiation condition in the general context of a scalar wave equation in GTR.

[27] Recall that any region S is a subset of $J^-(S)$.

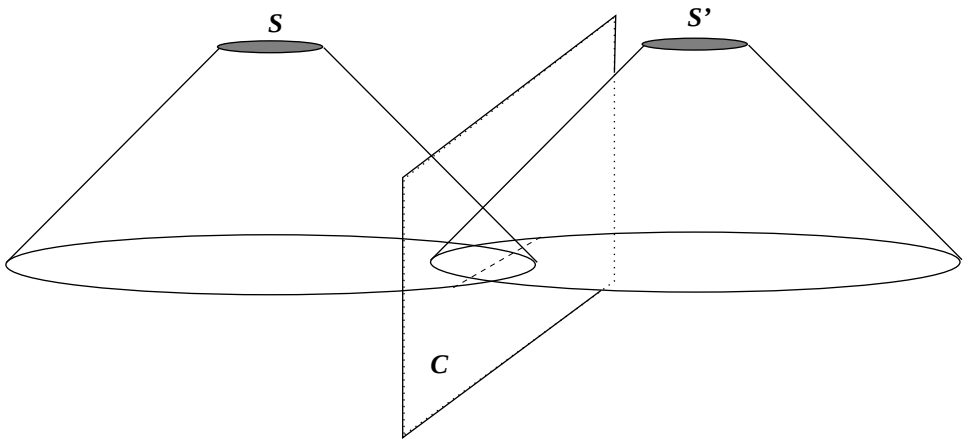

Fig. 6.1 The LCI requires that all correlations in the regions S, S' are screened off by completely specifying the state on a region such as C above, which divides $J^-(S) \cup J^-(S')$ into two pieces, one including S and the other S'.

that completely specifying the physical state in this region (labeling the states with lower-case letters, so the state in C is c) screens off all correlations between the physical states s, s' in S, S', i.e. $Pr(s \wedge s'|c) = Pr(s|c)Pr(s'|c)$. Influences coming in "from infinity" are presumed to be uncorrelated, so any correlations due to events in the common past of S, S' must be registered in the region C.

The LCI is much more modest than Reichenbach's PCC. As Penrose and Percival (1962) acknowledge, this principle forsakes any attempt to locate specific common causes for the correlations between the states in these two regions, and in this sense the LCI is weaker than Reichenbach's PCC. But by focusing on the physical state of the entire region C (one of many such regions), the LCI avoids difficulties with specifying which partitions of the state space count as legitimate common causes and there is also no risk of incompletely specifying causal factors. In the next section I will use the LCI as a diagnostic in assessing the horizon problem.

Before turning to that task, however, I should bring out one of the further consequences of the line of thought above. I have briefly reviewed cases above in which the PCC and similar causal principles do not hold, in order to show that these principles should not be taken as normative methodological principles of general scope. Approaching the subject from the opposite direction, under what conditions does the PCC or a variant of it hold? Although I will not pursue the idea further here, putting the question in this way suggests that causation as characterized by the PCC is a feature of systems sharing specific statistical regularities. So, for example, correlations that arise among the components of a system with an initially "chaotic" state and uncorrelated external causal influences would be required to have a common cause. If something along these lines can be made precise (see Arntzenius 1999, and references therein to the extensive causal modeling literature), then the question of the applicability of causal principles reduces to that of whether the appropriate statistical features hold for a given system.

6.2.2 Causality in the Early Universe

The most widely discussed fine-tuning problem in contemporary cosmology, the horizon problem, arises due to an apparent conflict between the presence of horizons in standard cosmological models and the striking uniformity of the microwave background radiation. The problem can be put succinctly in terms of two different horizon distances (see Appendix A.3 for more detail). In relativistic cosmology horizons measure the maximum distance travelled by a light signal during a fixed time period. The visual horizon d_{vh} measures the distance to the farthest objects visible to us by light emitted after t_d, the decoupling time when the universe became transparent. The primeval particle horizon d_{pph} measures the maximum distance from which light emitted "at the singularity" (in the limit as $t \to 0$) can reach a point by the time t_d. For the

standard FLRW models, the integrals defining these distances converge, yielding finite quantities obeying the following inequality:

$$d_{pph} << d_{vh} \tag{6.6}$$

As a consequence, if the FLRW models accurately model the early universe the snapshot of the CMBR encompasses several regions which lie beyond each other's primeval particle horizons. In particular, the CMBR emitted from points in the sky with an angular separation of more than 60° comes from two regions whose past light cones do not overlap (see 2.1). Antipodal points in the sky are separated by a distance of roughly $75\ d_{pph}$ at t_d (Blau and Guth 1987, pp. 534-535). Without a trans-horizon smoothing mechanism, the observed uniformity of the CMBR requires that all of these causally disconnected regions mysteriously have approximately the same state. The horizon problem is often more aptly called the "smoothness" or "uniformity" problem: the difficulty is not that particle horizons exist, but that their existence seems to be incompatible with observed uniformity.

This incompatibility is typically characterized in probabilistic terms: the standard cosmological models can accomodate uniformity, but it must be *stipulated* to hold as an incredibly improbable, contingent initial condition. Assume for the moment that the FLRW models accurately describe the early universe, so that there are regions $S, S' \subset \Sigma_{t_d}$ (where Σ_{t_d} is the surface of last scattering) whose past light cones do not overlap. We saw in the last section that the LCI enforces the intuitive requirement that "influences from infinity" should be uncorrelated; Penrose and Percival (1962) further note that the "LCI can be derived from the axiom that all spatially separated regions become uncorrelated in the limit $t \rightarrow +0$" (p. 614) in a cosmological model with a singularity at $t = 0$. For regions beyond each other's particle horizons, taking

the limit is unnecessary: the physical states in such regions must be uncorrelated for the LCI to hold. The LCI is based on the idea that local physical laws do not place constraints on the physical states associated with relatively spacelike regions. One might further require that any correlations among the field quantities in spatially separated regions should be contingent, in the sense that the field quantities in region S can take a variety of values even if those in S' remain fixed. Formally, call correlations between two relatively spacelike, open regions S, S' contingent if and only if given two cosmological models $\mathcal{M}= (M, g_{ab}, T_{ab})$ and $\mathcal{M}'= (M', g'_{ab}, T'_{ab})$ differing in the two regions ($\mathcal{M}|_S \neq \mathcal{M}'|_S$ and $\mathcal{M}|_{S'} \neq \mathcal{M}'|_{S'}$), one may always construct a third model $\mathcal{M}''= (M', g'_{ab}, T'_{ab})$ such that $\mathcal{M}''|_S = \mathcal{M}|_S$ and $\mathcal{M}''|_{S'} = \mathcal{M}'|_{S'}$. For contingent correlations the fact that S and S' have the same features in a particular model depends upon the initial conditions rather than upon the laws of the theory. On the other hand, if no model $\mathcal{M}''$ exists then the spacelike correlations are lawlike rather than contingent, since "wiggling" the field quantities in one region is incompatible with keeping the features of the other region fixed. The LCI demands that there are no lawlike correlations in this sense. Without such lawlike correlations, the uniformity of the early universe apparently results from initial conditions rather than subsequent dynamical evolution. As we saw in Chapters 3 and 4 above, cosmologists have not been satisfied with this need for special (uniform) initial conditions since they are apparently incredibly improbable.

Whatever the intuitive force of the LCI, a number of physical theories *do* incorporate lawlike constraints on relatively spacelike regions.[28] We have already seen one example of

[28] Earman (1987a) argues that the notion of locality introduced in the previous paragraph—equivalent to his (L9)—has remarkably little connection with the other ten definitions of locality he offers. Maxwell's equations satisfy every other notion of locality, and yet the constraint equations run afoul of (L9) and the PCC. I agree with Earman that this reflects a shortcoming of the PCC, in that it mistakenly diagnoses Maxwell's equations as exhibiting a form of non-locality.

this in the previous section—Maxwell's equations; general relativity also includes initial value constraints in close analogy with the electromagnetic case (Ellis and Sciama 1972; Wald 1984). Wald regards the neglect of possible non-local correlations as a serious oversight, and concludes that "it would be rather surprising if the existence of correlations beyond the horizon did not play an important role in accounting for some basic phenomena occuring in the early universe" (Wald 1993, p. 224). Wald's argument focuses on lawlike correlations present in relativistic QFT rather than classical physics: as a consquence of the Reeh-Schlieder theorem, in the vacuum state there are correlated elements A_1, A_2 of the local algebras of observables $\mathcal{A}_1, \mathcal{A}_2$ associated with the regions O_1, O_2 no matter how widely separated these regions are.[29] Assuming that the Reeh-Schlieder theorem generalizes to QFT on curved spacetimes, the question is then whether such spacelike correlations have any impact on calculations, such as estimates of monopole production, that typically neglect them.

Although these cases illustrate that the LCI should be modified to allow for lawlike constraints, a wide variety of initial data is still compatible with the constraint equations of the coupled Einstein-Maxwell field equations—including spacetimes which are not as smooth as the observed universe. Wald admits that the small correlation he calculates in a specific case would probably have a negligible effect on processes in the early universe. Constraints on relatively spacelike correlations based on well-established physical theories probably do not provide a strong enough restriction on the initial data to produce a uniform initial state. However, several speculative proposals for Planck scale physics, including Penrose's Weyl Curvature Hypothesis, quantum cosmology, and the "ekpyrotic" scenario (all discussed briefly in Section 5.4 above)

[29] A_1 and A_2 are correlated in a state ξ if $\langle\xi|A_1A_2|\xi\rangle \neq \langle\xi|A_1|\xi\rangle\langle\xi|A_2|\xi\rangle$. See Streater and Wightman (1964, Chapter 4) for a discussion and proof of the Reeh-Schlieder theorem.

provide such constraints. Imposing the LCI outright is in essence a blanket rejection of all three lines of research.

Returning to the question of the probability of the initial state, cosmologists from Misner onward have argued that nearly exact uniformity is intuitively improbable: even though the space of solutions of Einstein's field equations is not well understood, FLRW models can be singled out due to their high degree of symmetry. I have discussed the difficulties with bolstering these intuitions with precise measure theoretic results above 5.4.1, but here I want to briefly focus on the connection with horizons. Most cosmologists have laid the blame for the improbability of the required initial state on the presence of horizons; consequently, dynamical solutions to the horizon or uniformity problem have focused on radically altering the causal structure of the early universe so that the sources of the CMBR can be in causal contact. Earman (1995, pp. 144-145) has emphasized an important point usually glossed over in this argument: in order to justify laying the blame on the presence of horizons, one would need to show that among the cosmological models which start with a big bang and reach a uniform state quickly enough there is no open subset of models with particle horizons. A result along these lines would justify linking the improbability of a uniform initial state with the presence of particle horizons; otherwise, one may treat the improbability of the initial state as a *general* problem with the big bang model having nothing to do with horizons (as Penrose 1979 does). Granting the presumed linkage between improbability and the presence of horizons, there are several different ways to modify the FLRW models to completely get rid of particle horizons: horizons are absent in compact cosmological models with topological identifications, and modifications of Planck

scale physics can also alter the horizon structure.[30] But as we saw in Part I, the dominant research program in early universe cosmology, inflation, does not completely *eliminate* horizons; instead, sufficient inflation makes causal interactions between sources of the CMBR *possible* due to an overlap of their past light cones.

While inflation clearly satisfies a *necessary* condition for a dynamical explanation of uniformity, it is far less clear that inflation is *sufficient* to produce uniformity. The first is the question of whether causal horizons accurately delimit the range of local causal interactions. There are two reasons to suspect that the horizon does not accurately measure the limits of causal interactions: as noted by Ellis and Stoeger (1988) and others, the particle horizon is simply the wrong horizon, and the appropriate horizon distance is an upper limit which may not be reached in realistic physical models. Regarding the first point, eqn. (A.16) gives the maximal proper distance over which light propagates from $r_0 = 0$ in a time interval $\Delta t = t_0 - t_e$. Suppose, for example, that on this definition a light signal emitted from particle A at t_e reaches particle B within Δt. It does not follow that B can return a signal to A—i.e., in some solutions (such as the exponentially expanding de Sitter solution during inflation) a return light signal emitted by B at t_0 cannot reach A.[31] Thus, the fact that A and B lie within each other's horizons does not guarantee the possibility of "interaction" in the sense of signals being exchanged between A and

[30] The various proposals are discussed in more detail in Chapters 3 and 4 above. Note that the integral defining the particle horizon (A.17) diverges if $a(t) \propto t^n$ with $n \geq 1$; several of the proposals insure that the integral diverges (and thereby eliminate the horizons) by suggesting new physics which would lead to evolution of the scale factor of this type.

[31] This point motivated Patzelt (1990)'s introduction of the *interaction horizon* (cf. Ellis and Stoeger (1988), Lightman and Press (1989)), defined as the maximal coordinate distance between two world lines such that A may receive a return signal emitted at t_0:

$$d_{ih} = \max_{t_e} a(t_0) \int_{t_e}^{t_0} \frac{dt}{a(t)} \tag{6.7}$$

As Patzelt (1990) shows, during an inflationary stage the particle horizon is an upper bound on the interaction horizon.

B. Returning to the second point, it is not clear that local interactions in the early universe travel at the speed of light: as Ellis and Stoeger (1988) point out, even for massless particles frequent scattering interactions limit the effective speed of propagation.

6.2.3 Robustness

Sober has argued that the PCC derives whatever methodological force it legitimately has from the more general "likelihood principle." In cases in which the PCC applies, according to Sober, the theory postulating a common cause simply renders the evidence more probable than a separate cause theory does. Although inflation does not offer a causal explanation of the uniformity of the early universe, it does appear to satisfy the likelihood principle in that it dramatically enlarges the range of initial conditions compatible with observational constraints. In this section I will formulate a criterion of robustness for dynamical explanations; one advantage of this formulation is that in cases without a well-defined probability measure one can still retreat to a general "dominance" argument. The preference for a theory providing robust explanations is then a consequence of the likelihood principle: the probability of a set of observations O is greater according to a more robust theory.

This comparative notion of robustness should be contrasted with a more general complaint that *any* explanation based on special initial conditions is somehow flawed. Earman (1995) has argued convincingly that there is little support for such a strong robustness demand: in some cases good dynamical explanations *do* depend on special initial conditions. Consider the probability of certain features of our solar system (e.g., planets moving in elliptical orbits with low eccentricity lying in roughly the same plane, smaller interior planets composed of heavy elements, etc.) given a background dynamical theory. Several of these features, for example the

low eccentricity of the planetary orbits, seem intuitively improbable. Assuming that the current theory of planetary formation does not produce nearly circular orbits for a large range of initial conditions, then this theory does not offer a "robust" explanation of nearly circular planetary orbits in this general sense. A general demand for robustness surpasses a more modest complaint: perhaps we could gather evidence that in its early stages our solar system was *not* in the special state the current dynamical theory requires. But without the evidential basis for this modest complaint, there seems to be no reason to accept the more general demand—our planetary system may have started in a "special" initial state for all we know. In addition, robustness characterized as *complete* independence from initial conditions overshoots the mark; as I argued above, a dynamical theory may enlarge the set of initial data compatible with observations but it does not completely eliminate the dependence on initial conditions.

A comparative notion of robustness can be formulated directly if one can sensibly define a probability measure over the space of models of the theory. Suppose that there are two theories T_1 and T_2 such that the set of models $\mathcal{M}_1^i$ which matches observations is generic in the space of models of T_1, whereas the corresponding set of models $\mathcal{M}_2^i$ of T_2 is of small or zero measure (leaving aside for the moment the question of exactly how to assign this measure). In this case the theory T_1 provides more robust explanations of the observations than T_2, in that the set of models compatible with observations has much larger measure. For the sake of concreteness, note that for well-behaved generally relativistic spacetimes the space of models of the theory can be identified with sets of initial data. For globally hyperbolic spacetimes, the state of the universe at a given cosmic time may be represented by an initial data set specified on a Cauchy surface Σ. These initial data sets correspond one-to-one with cosmological models (up to diffeomorphism invariance), allowing us to assign a measure over cosmological models in terms of the initial

data specified on a Cauchy surface Σ. Unfortunately attempts to define a measure over the space of FLRW models with a massive scalar field in order to allow this type of comparison have led to the "unambiguously ambiguous" results (Donald Page, as quoted in Lightman and Brawer 1990) briefly described in 5.4.1 above.

Consider the following "dominance" criteria as a replacement for measure theoretic results: a theory T_1 is more robust than T_2 if the set of initial data compatible with observations given T_2 (call this set IC_2) is a proper subset of the set of observationally allowed initial data for T_1 (IC_1). The idea is that whatever measure is assigned over the initial data, since $IC_2 \subset IC_1$ we can still say that $Pr(IC_2) \leq Pr(IC_1)$. The flatness problem provides the clearest case of this type of "dominance": the set of values of $\Omega(t_p)$ compatible with the observational constraints is a superset of those compatible without inflation. In this sense inflation provides a more robust explanation of the value of Ω according to the dominance criterion.

There are two general difficulties with this dominance criterion. A substantially different cosmological theory such as one incorporating Penrose's Weyl curvature hypothesis could undercut the advantage of dominance by introducing a strong constraint on relatively spacelike initial data. This would have the effect of ruling out a large part of the space of classical cosmological models, and it is at least conceivable that the resulting small set of allowed models would be only those compatible with observations. Given that Penrose's hypothesis is explicitly formulated as a phenomenological description of the observed universe, it is not surprising that it singles out uniform initial states. The more interesting question is whether an independently motivated theory of quantum gravity will incorporate a similar constraint. The second difficulty is that the intuitive dominance arguments presented in the literature often involve an interplay between adjusting the parameters of a theory and eliminating fine-tuning of the initial conditions.

If the potential of the inflaton field is treated as a free function, then observational constraints can be used to fix parameters of the theory so that fine-tuning of the initial conditions is eliminated. I agree with Ellis (1991)'s qualms about this type of argument; it is not clear how to make rigorous probability assessments or how to apply a dominance criterion when one considers a trade-off between fine-tuning the initial conditions and the parameters of the theory.

6.3 Conclusion

My argument above has focused on the alleged explanatory advantages of inflationary cosmology. One of the primary appeals of inflation after its initial introduction was that it appeared to solve a number of outstanding problems in cosmology without grappling with Planck-scale physics and the nature of the initial singularity. The main thrust of the argument above is that the case for inflation does in fact depend upon very strong assumptions about Planck scale physics. The "overlapping domains" argument stated in §1 establishes that a complete theory of early universe cosmology should incorporate both QFT and GTR, but I agrued that there are three obstacles to satisfactory unification. After 20 years the status of the "inflaton" field with respect to underlying particle physics is still unsettled. In addition, the cosmological constant problem indicates that there are still substantial uncertainties in how to incorporate quantum fields as sources in semi-classical quantum gravity; resolutions of the problem may rob inflation of its power source, namely vacuum energy. Finally, a general "cosmic no hair" theorem would justify the neglect of the subtleties involved in the onset of inflation, but I argued that these theorems basically assume the applicability of flat space QFT during the "chaotic" start of inflation. In §2 I argued that the explanatory advantages of inflation related to robustness also require strong assumptions regarding Planck scale physics; namely, that the universe emerged

from the initial singularity in a state without lawlike correlations between relatively spacelike regions. Such “chaotic” initial states are incompatible with the regular and uniform state revealed by the CMBR. A stage of inflationary expansion insures that a much larger set of these allowed models are compatible with observational results, and thus inflation offers a more robust explanation of the universe’s uniformity and flatness. However, a radically different conception of the initial singularity incorporating constraints on relatively spacelike regions would render such dynamical mechanisms superfluous.

Chapter 7

Confirming Inflation

The preceding chapters considered the invocation of metaphysical principles and criteria of explanatory adequacy to motivate research programs in cosmology. But perhaps my criticisms of these approaches are irrelevant in the light of a more straightforward empirical case in favor of a particular program, such as inflationary cosmology. Ongoing attempts to decipher the "Cosmic Rosetta Stone"—i.e., the careful measurement of temperature anisotropies in the CMBR—have reached new levels of precision in the past several years, especially with the release of the initial WMAP (Wilkinson Microwave Anisotropy Probe) data.[1] With the advent of various other observational tools (such as the use of type IA supernovae as standard candles) and further satellite missions, frequent talk of a "golden age" in observational cosmology is not exaggerated.

Contemporary presentations of inflation often emphasize that these new results can be used to make such a strong empirical case for inflation that questions of explanatory adequacy can be set aside. In other words, the account of structure formation and its comparison with CMBR observations are taken to replace the original rationale for inflation, its ability to solve the fine-tuning problems of big bang cosmology. For example, Liddle and Lyth (2000) comment that

[1] WMAP is optimized to observe small angle temperature variations in the CMBR with a precision much greater than the earlier (1992) COBE satellite, which first detected temperature anisotropies. See, for example, Bennett et al. (2003); Spergel et al. (2003) for the analysis of the first year of WMAP data. Numerous earlier rocket and balloon based experiments such as BOOMERANG and DASI also measured these temperature anisotropies, and the European Planck satellite (scheduled for launch in 2007) is designed to measure polarization.

> By contrast to inflation as a theory of initial conditions, the model of inflation as a possible origin of structure in the Universe is a powerfully predictive one. Different inflation models typically lead to different predictions for the observed structures, and observations can discriminate strongly between them. Future observations certainly will exclude most of the models currently under discussion, and they are also capable of ruling out all of them. Inflation as the origin of structure is therefore very much a proper science of prediction and observation, meriting detailed examination. (Liddle and Lyth 2000, p. 5, cf. Barrow and Liddle 1997)

I take Liddle and Lyth to be arguing that inflation offers a fruitful reinterpretation of the spectrum of density perturbations subject to rich empirical tests. Within the context of inflationary theory, small departures from the exact symmetry of the FLRW models provide measures of the potential of the field driving inflation. According to the theory, these perturbations are the product of quantum fluctuations stretched and imprinted during the inflationary stage.

The compatability of inflation with the observations of the CMBR is clearly an important success of the theory. The COBE observations served to rule out inflation's only major competitor in the early 90s, the topological defect theory of structure formation. Yet there is a natural concern whether this success represents anything more than the malleability of the inflationary paradigm. Let me briefly illustrate this concern with an extreme example of a similar problem. Advocates of intelligent design claim that some empirical features of the world—say, the characteristics of cheetahs—can be explained as the product of a Designer's plans (this example is borrowed from Sober 1999). The notorious and obvious difficulty with assessing such a claim is that it neatly divides into two conjuncts: (1) the cheetah is the product of intelligent design, and (2) if a Designer were to make cheetahs, they would have the the following properties: "_____" . Given what we know of cheetahs, it is easy to fill in the "_____" appropriately to guarantee that the theory of intelligent design reproduces the observed characteristics of cheetahs. But this clean division poses a problem for the advocate of the hypothesis (1), since

the flexibility of (2) shields it from being independently tested. To answer this objection the advocate of intelligent design requires a detailed and independently motivated account of how to fill in the blank. Any account of (2) must answer the skeptic's immediate rebuttal, namely that the account of the Designer has been carefully designed to produce the desired results. (And presumably this account should differentiate between cheetahs and other animals, such as grey-hounds, with similar characteristics that are clearly *not* due to the Designer.) Thus the difficulty with intelligent design is not that it fails to make predictions, but that the predictions it does make do not provide grounds for assessing the claim of interest rather than the conjunct.

The analogy with inflationary cosmology runs as follows: the claim that various features of the early universe can be explained as the product of an inflationary stage also divides into two conjuncts, namely (1) an inflationary stage occurred, and (2) if inflation occured, then the universe has the following properties: "_____". Do we have grounds for filling in the blank that are independent of our knowledge of the observed properties of the early universe? The combination of "chaotic" variations in the initial conditions and form of the inflaton potential with an appeal to the anthropic principle seems to guarantee a negative answer in some versions of chaotic inflation. However, in general the prospects do seem much better than in the case of intelligent design. Suppose that the properties of the scalar field driving inflation are fully specified by considerations from particle physics. Then we would (*modulo* concerns about the various approximations involved and the computational tractability of the model) be able to calculate the "output" of an inflationary stage, for some given initial conditions, and compare it to observations of the CMBR. Obviously passing such a test would provide a confirmatory boost to inflation; Peebles, for example, counts "deduction of the inflaton and its potential from fundamental physics" as a classical test of inflation ("one that follows the old rule of validation

by the successful outcome of tests of the predictions of a theory"). But he also characterizes the status of this "deduction" as nothing but "a wonderful dream" (Peebles 1999). This worry about the independent assessment of the "output" of an inflationary stage has more force following the shift described in Chapter 4 towards treating the "inflaton" as a free parameter. Critics of inflation such as Neil Turok have frequently cited inflation's malleability as a fundamental obstacle to any meaningful test of the theory (Turok 2002): "I can't think of a conceivable test which would decisively prove inflation wrong. Therefore I don't think it's a testable model."

In the following, I will focus on two aspects of the difficulties with testing inflation. First, there are two related difficulties regarding the trio of inflationary predictions emphasized by Guth (1981): are these predictions "robust," in the sense that all "natural" inflationary models give the same predictions, and are they also distinctive, in that they differ from predictions made by alternative theories of the early universe? My sympathies are with the inflationary skeptics, who answer with a qualified "no" to both questions. The skeptic's qualms are due to our ignorance of the space of alternative theories, and they have offered arguments that several of the predictions of inflation are to be expected based on other background assumptions, and thus deliver a small confirmatory boost. By contrast, turning to the second aspect, inflationary predictions regarding the spectrum of density perturbations are not subject to the same criticisms.[2] Furthermore, in some presentations of inflation the account of structure formation is characterized as a "novel success" of the theory: inflation was not "designed" to give an account of structure formation, and thus its success should be given extra weight. But the invocation of design cuts both ways, since the inflaton potential needs to be tuned carefully to produce a workable account of structure

[2]Within the past three years, however, the "ekpyrotic scenario" has developed a viable alternative mechanism for generating density perturbations with many of the same features of inflation. In this chapter, I will unfortunately not take up a more detailed comparison of the two theories.

formation. Below I will argue that the fundamental issue is not one of novelty, but rather a question of what sources of information are used as a "diagnostic," that is, used to set various parameters occurring in the theory — in this case in the inflaton potential. In well understood theories multiple independent sources of observational or experimental data can be used to set parameters appearing in the theory, and the theory generates predictions of greater scope than the "diagnostics" used in fixing these parameters. My tentative suggestion is that cases of "novelty" are often associated with a theory's ability to generate predictions for a range of phenomena that were not invoked as diagnostics in developing the theory.

7.1 Testing Inflation I

Proponents of new theories often advocate reinterpreting a set of known regularities using their novel theoretical machinery. Inflationary cosmology replaces details regarding the universe's initial state with the effective potential $V(\phi)$ of a new fundamental scalar field, the inflaton field ϕ. Various features of the universe are then interpreted as the results of the evolution of ϕ in the early universe. One of the difficulties with assessing the evidential support for inflationary cosmology is that inflation is a "paradigm without a theory" (in Michael Turner's phrase). Models of inflation share the general feature that the early universe goes through a stage of rapid expansion (characterized by $\ddot{a} > 0$) driven by ϕ trapped in a false vacuum state. But the wide variety of models differ in the form of the effective potential $V(\phi)$ appearing in the equations of motion for ϕ (or for multiple fields) and in the initial conditions of the field(s). These differences lead to a number of variations—some subtle, but others quite dramatic—in the state of the early universe following inflation.

Rather than focusing on a particular model plucked from the inflationary zoo, I will review the frequently cited "robust" predictions of inflation; by robust I mean the predictions shared by the class of "natural" inflationary models. Qualms about whether these predictions are truly robust have two sources: the first is simply that theorist's opinions as to what constitutes a "natural" inflationary model—and hence whether a given prediction is truly robust—differ; the second is that the calculations may actually depend on strong assumptions regarding Planck-scale effects (to be discussed in section 7.5). These qualms notwithstanding, one may hope to find decisive empirical tests for a large class of inflationary models by specifying the fingerprint of a period of exponential expansion. The second difficulty, which I will address in Bayesian terms, is that the weight given to finding this fingerprint of inflation depends on whether alternative theories have similar fingerprints.

The following trio of inflationary "predictions" emphasized in Guth (1981) are trotted out in any introduction to inflation:

Massive Relics

> GUTs are expected to produce magnetic monopoles in a symmetry-breaking phase transition at an energy of roughly $10^{14}GeV$. Monopoles are produced with a density of roughly one monopole per horizon volume at the time the phase transition occurs; given that the observable universe (sans inflation) encompasses roughly 10^{80} such horizon volumes, it should contain roughly 10^{80} such monopoles. At the lightest, these monopoles are expected to have masses of about $10^{16}GeV$—so at this abundance, the monopoles alone would contribute an energy density roughly 11 orders of magnitude greater than the critical density.[3] Incoporating inflation leads to an expected monopole abundance in the observable universe of one, since the universe expands from a single horizon volume, and thus eases the disastrous conflict with observation.

Spatial Flatness

[3] See, e.g., Blau and Guth (1987), pp. 530-32 for this calculation of the energy density contributed by monopoles.

> Inflation drives the density paramater Ω rapidly towards one during the stage of exponential expansion, but later stages of FLRW evolution magnify any small differences from $\Omega = 1$ (see Appendix A.2). If Ω takes a value on the order of 1 initially, then if the scale factor increases by a factor f during inflation, it follws that at the end of the inflationary stage $\Omega - 1$ is on the order of f^{-2}. A sufficiently long inflationary stage ($\ln(f) \geq 60$, usually called the "number of e-foldings") insures that $|\Omega_0 - 1| \ll 1$. Inflationary models satisfying this requirement predict that $\Omega_0 \approx 1$.

Large-scale Uniformity

> On large scales (from angular scales of about $10''$ to 180°) various observations reveal that the CMBR temperature is incredibly uniform, to roughly one part in 10^5. Given the same number of "efoldings" as above, inflation traces this uniformity back to the uniform evolution of the inflaton field in the single horizon volume which expanded to encompass the observable universe.[4]

The development of an ever wider variety of inflationary models has cast some doubt on whether this trio qualify as "robust" predictions. Inflation models yielding a wide range of monopole densities have been produced, weakening Guth's original result that inflation necessarily results in a negligible monopole density (see Yokoyama 1989).

The "robustness" of the flatness prediction is undermined by two important points. First, "open" inflationary models have been designed to produce $\Omega_0 < 1$ (see, for example Bucher et al. 1995; Ratra and Peebles 1995). For those who accept open models, decisive observational evidence that $\Omega_0 \neq 1$ would rule out a large class of inflationary models but not the very idea of inflation. Yet many theorists describe open models as unacceptably ugly:

> I think open inflation is an example where you're trying to turn a model against itself. Inflation ... has the word "flat" in it to remind you that it's about flattening the universe. And, in particular, what's spectacular about inflation is that with exponential efficiency it flattens the universe. So it's like a bulldozer running through the universe and flattening it, and to make open inflation is like saying, "I'm going to have this bulldozer running at top speed and then I'm going to stop it at a dime after 50-52 efolds of inflation, because I need to have not quite enough inflation to make the universe flat." (Steinhardt 2002, p. 45)

[4] Here I am setting aside the difficulties with showing that inflation occurs in generic conditions in the early universe and that it produces uniformity (i.e. the cosmic "no hair" theorems) discussed in the previous chapter.

Ruling out ugly models requires placing constraints on $V(\phi)$ and/or barring monsters such as hybrid inflation models with multiple fields or multiple stages of inflation; without some degree of agreement regarding the background particle physics theory in which $V(\phi)$ is defined it is hard to see how to justify such constraints aside from subjective appeals to simplicity. Proponents of the open models certainly did not see their proposals as requiring an inherently "ugly" choice for the effective potential or other detailed aspects of their models. Debates on this issue were not so much settled as pushed to one side as a number of observational results supported a value of $\Omega_0 \approx 1$, contrary to earlier work that had consistently indicated a lower value of Ω_0; inflationary cosmologists who had insisted all along on excluding the monstrous open models took this as a final vindication.[5] The situation is similar to the tests of early GUTs via proton decay experiments: while the failure to detect proton decay on the appropriate time scale did rule out the standard $SU(5)$ GUT, it also did not lead theorists to abandon the project of unification.[6]

The second point undermining the robustness claim is that the "prediction" that $\Omega_0 = 1$ only holds given some assumptions regarding the pre-inflationary value of Ω (see, e.g, Madsen et al. 1992). This is simply Stewart's objection to Misner reiterated: pick any particular value of Ω_0 (even one far away from 1), and one can always trace it backwards through the evolution of $a(t)$—whether there is an inflationary stage or not—to find an initial value of Ω. This objection

[5] See Coles and Ellis (1997) for a review of the debate circa 1997 that makes a strong case for $\Omega_0 \approx 0.2$; within the last eight years several new observational results have persuaded most of the community to accept a large Ω_Λ, leading to a total $\Omega_0 = 1$ (see, e.g., Bahcall et al. 1999).

[6] Albrecht (1997), among others, emphasizes this analogy. Albrecht also points out that the failure of the standard $SU(5)$ GUT lead to a variety of different approaches to GUT-scale physics, and he expects that the failure of "natural" inflationary models would force theorists to work with a variety of fairly contrived models. One important disanalogy between the two cases is that no fundamental principles guide inflationary model-building in the same sense that gauge invariance and renormalizability guide the unification program.

is much weaker. The typical response is to appeal to the idea of dominance (discussed in §6.2.3 above): inflation renders a much larger range of initial values of Ω compatible with $\Omega_0 = 1$.

The more general issue is the tremendous flexibility of the inflationary paradigm if it is treated simply as scalar field dynamics. Since the false vacuum state dominates over all other sources of energy density during the inflationary stage, the effective potential $V(\phi)$ determines the behavior of the scale factor $a(t)$. The potential is usually taken to be the "input" fixed by particle physics with the behavior of $a(t)$ treated as the "output". But turning this around, Ellis and Madsen (1991) have shown that the form of the potential $V(\phi)$ (treated as a free function) can be derived for various different behaviors of the scale factor $a(t)$. This "recipe" for generating the form of the potential is a simple case of a more general trick: take a solution of EFE as given and simply define the matter content by the appropriate T_{ab}. In this case Ellis and Madsen (1991) consider exact FLRW models with a stress energy tensor including radiation and a scalar field, with no interaction between the two components. The EFE yield two equations relating the scale factor $a(t)$ to the energy density of the radiation and the energy density and pressure of the scalar field, along with a conservation equation for the total energy density and pressure. For a given monotonic function $a(t)$ these equations can be solved for the potential $V(\phi)$. Thus in the absence of constraints on $V(\phi)$, inflation can produce any desired behavior of the scale factor $a(t)$.

Aside from robustness concerns, the assessment of these predictions depends on the likelihood assigned to a particular evidence claim were inflation to be false. Arguments about the probable state of the universe *sans* inflation fall back on speculation about the space of alternative theories, and critics of inflation have often argued that observed features of the universe may ultimately be explained by a quantum theory of gravity, rendering inflation superfluous. Such

expectations leave critics less impressed by inflation's ability to predict, for example, a uniform early universe. This can be cashed out in Bayesian terms as follows. In Bayesian confirmation theory, the confirmatory power of an evidence statment E for a given hypothesis H is given by $C(H, E, K) = Pr(H|E \wedge K) - Pr(H|K)$.[7] In light of evidence E, the prior probability accorded to hypotheses H is updated according to Bayes's theorem:

$$Pr(H|E \wedge K) = \frac{Pr(H|K) \times Pr(E|H \wedge K)}{Pr(E|K)}. \tag{7.1}$$

We can rearrange this equation slightly, utilizing the principle of total probability and assuming that $\{H, K\} \models E$, so that $Pr(E|H \wedge K) = 1$:[8]

$$Pr(H|E \wedge K) = \frac{1}{1 + Pr(E|\neg H \wedge K) \times \frac{Pr(\neg H|K)}{Pr(H|K)}}. \tag{7.2}$$

An increase in the conditional probability $Pr(E|\neg H \wedge K)$ (assuming that $\neg H$ is given a non-negligible prior) leads to a decrease in $Pr(H|E \wedge K)$, and hence a decrease in the confirmational boost H receives in light of E.[9] The expectation that one of the hypotheses in the *terra incognita* of $\neg H$ entails E leads to a lower value for the confirmatory power of E.

---

[7] See, e.g., Howson and Urbach (1989) for a comprehensive introduction to Bayesian confi rmation theory. $Pr(H|K)$ is the prior probability accorded to the hypothesis H given background knowledge K, whereas $Pr(H|E \wedge K)$ is the posterior probability, clculated according to Bayes's Theorem. According to the subjectivist interpretation of Bayesianism, these probabilites are interpreted as rational degrees of belief.

[8] The principle of total probability states that $Pr(E|K) = Pr(H|K)Pr(E|H \wedge K) + Pr(\neg H|K)Pr(E|\neg H \wedge K)$. The equation above follows by expanding the denominator, dividing by $Pr(H|K)$ and then using $Pr(E|H \wedge K) = 1$.

[9] $Pr(E|\neg H \wedge K)$ and $Pr(E|H \wedge K)$ are called likelihoods. In cases of statistical hypotheses, past data regarding frequencies or explicit assumptions about chance processes built into H justify reading the likelihoods as objective propensities (see Salmon 1990, §6 for discussion). I will set aside for the present discussion the problem of how probabilities are to be interpreted in cosmology (and whether an objective propensity interpretation would apply here), and read the likelihood as a degree of belief.

As a concrete example, consider how this Bayesian machinery clarifies the difference between inflation's predictions of uniformity and monopole abundance. During discussion at the Seven Pines Symposium (May 10-14, 2000), Robert Wald argued that inflation's prediction of large-scale uniformity (E_U) does not merit a large confirmatory boost, due to the possibility that an as yet undreamt of theory may entail that the early universe is remarkably uniform. In the terms above, Wald would assign a high value to the likelihood $Pr(E_U|\neg H \wedge K)$, and a correspondingly low value of $C(H, E_U, K)$. By way of contrast, Wald acknowledged that the observed negligible abundance of monopoles (E_M) would provide convincing evidence of inflation if it could be shown that GUTs unambiguously predict a high monopole abundance which cannot be reduced by annihilation in the early universe. Unlike the assessment of $Pr(E_U|\neg H \wedge K)$, which in Wald's view depends on an educated guess regarding future theory at the Planck scale, monopole abundance calculations rely on a (somewhat) more modest semi-classical approximation. The space of hypotheses relevant to assessing $Pr(E_M|\neg H \wedge K)$ includes GUTs that do not produce appreciable monopole abundances and/or those that reduce abundances by means other than dilution during inflationary expansion. If no such theories exist—that is, if GUTs necessarily produce high monopole abundances, then the value of $C(H, E_M, K)$ would be very high. In both cases, Wald's assessment emphasizes the relevance of the (only partially understood) space of alternative theories.

This example illustrates that the evaluation of $Pr(E|\neg H)$ (dropping explicit conditionalization on K) relies on plausibility arguments that reflect an ignorance of the space of alternative theories. In a case where there are several viable alternative theories, $\neg H$ can be broken down further into a disjunction of alternatives to H (say, $H_1, H_2, ...$) and the "catchall" hypothesis H_C representing the undiscovered country (added to insure that the list of hypotheses is exhaustive).

For the enumerated alternatives, a degree of subjectivity in the evaluation of $Pr(E|H_i)$ persists in cases where it is not known whether H_i conjoined with K entails the evidence, and the hypothesis is not related to a statistical model. But the inflationary skeptic is in the more difficult position of arguing that we should take seriously the idea that for one of the unknown theories contained in the catchall (say, H_α) $\{H_\alpha, K\} \models E$, where E includes the evidence usually cited in favor of inflation. Even if such an H_α exists, it does not follow that $Pr(E|H_C)$ is close to unity (the likelihood of E for one of the disjuncts of H_C does not imply anything regarding the likelihood for H_C itself). Thus the analysis of $Pr(E|H_C)$ is particularly intractable, as Salmon (1990) has emphasized. In light of this difficulty Salmon (1990) suggests focusing on comparisons of existing alternative theories, so that we need only calculate the ratio of probabilities $\frac{Pr(H_1|E\wedge K)}{Pr(H_2|E\wedge K)}$ (and the problematic term $Pr(E|H_C)$ cancels).[10] Although this may be an appealing move in cases of theory choice between two (or more) well developed competing theories, in the present case, the inflationary skeptics' dissatisfaction stems from the lack of effort devoted to exploring alternatives to inflation.

7.2 Testing Inflation II: Structure Formation

A proponent of inflation could circumvent both difficulties by showing that inflation delivers a robust prediction of an otherwise unexpected result (i.e., one such that $Pr(E|\neg H \wedge K)$ is low). The leading candidate for such a prediction is the spectrum of density perturbations produced during inflation. These variations in density leave an observable imprint on the CMBR

[10] Earman (1992) argues that the cost of avoiding likelihoods (Salmon also eschews calculations of $Pr(E|K)$) is too high: one can neither compare the differential confirmation value of different evidence claims with respect to the same theory, nor assign an absolute low probability to a "theory" such as scientific creationism.

as temperature fluctuations. Inflation produces inhomogeneities in the early universe by amplifying variations from small length scales to cosmological scales.[11] On the assumption that the only pre-inflation inhomogeneities present are the vacuum fluctuations of the inflaton field ϕ, one can determine the spectrum of density perturbations following the completion of inflation by studying the evolution of these fluctuations during the inflationary stage. Briefly, one treats the vacuum fluctuations as a small perturbation ($\delta\phi$) to the field ϕ. If the so-called slow-roll approximation applies, the $\frac{dV}{d\phi}$ term in the equations of motion for ϕ can be neglected, and one obtains the following equations of motion for the perturbation:

$$(\ddot{\delta\phi}) + 3H(\dot{\delta\phi}) + \left(\frac{k}{a}\right)^2 (\delta\phi) = 0, \tag{7.3}$$

where k is the wave number for a given Fourier component. Expanding $\delta\phi$ in terms of creation and annihilation operators, and plugging back into the equations of motion, the solution for $\omega_k(t)$ is given by (neglecting time dependence of H):[12]

$$\omega_k(t) = L^{-3/2} \frac{H}{(2k^3)^{1/2}} \left(i + \frac{k}{aH}\right) exp\left(\frac{ik}{aH}\right). \tag{7.4}$$

[11] For a similar treatment that I draw on here, see Liddle and Lyth (2000), Chapter 7, or for a more detailed discussion of cosmological perturbations in general as well as the production of density perturbations in an inflationary stage, see Mukhanov et al. (1992).

[12] The field is defined in terms of creation and annihilation operators ($a_k, a_k^\dagger$) as follows: $\delta\phi(t) = \omega_k(t)a_k + \omega_k^*(t)a_k^\dagger$. The solution is found by stipulating that for wavelengths much smaller than the horizon size ($k \gg aH$), we have the familiar flat-space field theory result, written in co-moving coordinates as $\omega_k = a^{-3/2}\left(\frac{2k}{a}\right)^{-1/2} e^{-\frac{ikt}{a}}$.

When the horizon size is much greater than the wavelength of these fluctuations, $aH \gg k$ and the solution of $\omega_k(t)$ simplifies, yielding

$$\langle 0||\delta\phi_k|^2|0\rangle = |\omega_k|^2 = \frac{H^2}{2k^3}. \tag{7.5}$$

This indicates that for a given mode the perturbation in $\delta\phi$ becomes "frozen out" at this fixed value (it no longer depends explicitly on time, since H is treated as a constant). Thus the quantum field no longer fluctuates; instead, the modes stretched to super-horizon scales are imprinted as classical perturbations in the field value, which are then linked to metric perturbations via Einstein's equations. These curvature perturbations return to sub-horizon scales following the end of inflation (when the Hubble length becomes greater than the comoving length scale). Finally, one can express the spectrum of density perturbations in terms of equation 7.5, which in turn can be expressed in terms of the potential $V(\phi)$ and its derivatives with respect to ϕ; this establishes the desired link between observed density perturbations and the properties of the inflaton and its potential.[13]

Recent research on inflation has emphasized the difference between the inflationary predictions for density perturbations and the earlier predictions regarding initial conditions. Liddle and Lyth (2000) praise inflation as a theory of structure formation for making the following generic and robust predictions:

Near Scale Invariance

For a scale invariant spectrum of density perturbations, the density contrast $\delta M/M$ is constant (it is the same at all scales). Inflationary models generally predict a nearly scale

[13] In section 7.5 below, I will discuss recent criticisms that this calculation relies on a number of questionable assumptions.

invariant spectrum; however, this can be altered if H is not constant during the era when the density perturbations cross the Hubble radius.

Passivity

Inflationary perturbations are imprinted quite early, but after the inflationary period the perturbations evolve according to linear equations of motion (that is, they evolve passively). This linear evolution produces the characteristic "Sakharov oscillations."

Gaussianity

The distinct Fourier modes of the density perturbations produced during inflation are uncorrelated.

Consistency equation

Inflation produces tensor perturbations (also called relic gravitational waves) in addition to scalar perturbations. The vacuum fluctuations of the inflaton give rise to both types of perturbation, and both are determined by the continuous function $V(\phi)$. One can eliminate $V(\phi)$ in expressions for the spectra of scalar and tensor perturbations, thereby obtaining the so-called consistency equation $r = -2\pi n_G$, where r is the ratio of contributions from the tensor and scalar perturbations and n_G is the spectral index of the gravitational waves. This consistency equation holds for any $V(\phi)$ in a single field model, as long as the slow-roll approximation holds.

The inflationary skeptic is thus deprived of the two criticisms discussed above: in this case, inflation does appear to make specific, generic predictions, and (until very recently) there were no arguments to the effect that $Pr(E|\neg H \wedge K)$ should be high. However, the skeptic might change her strategy: as we saw in Chapter 4, current candidates for the inflaton field have all been chosen with observational constraints on the amplitude of the resulting density perturbations in mind. Should the use of this evidence in the construction of the theory weaken the confirmatory boost it renders to the completed theory?

7.3 Use-Novelty and Independence

Quarrels about whether *novel* predictions pack any extra confirmatory punch go back to (at least) the Whewell - Mill disputes, and have continued to the present. Those who attribute extra importance to novelty distinguish between two situations. In the first, a theory predicts

new phenomena that were not considered during the development of the theory. In the second, a theorist well aware of a set of observational results ingeniously designs a theory that accomodates all of these results. Surely, the argument goes, in the first situation the theory has passed an important test and our confidence in it should improve, whereas a theory is not similarly tested in the second situation. John Worrall has recently defended a view along these lines, and I will use his discussions of *use novelty* (in Worrall 1985, 1989) as a starting point for the discussion below.[14] The opposing side in the debate sees incorporating genealogy as a fundamental mistake. How *could* a particular scientist's (or even the scientific community's) path to the discovery of a theory have any impact on the *truth* of a theory? Surely psychological details, such as a scientist's awareness of a particular result (or failure to take it into account), are irrelevant to how the theory fares before the tribunal of experience. On this view confirmation theory consists of understanding the content of a theory and how it fits with available evidence, and the distinction between novel predictions and accomodation is nothing but history. Below I will discuss use novelty briefly, leading into Leplin's analysis of novelty in terms of the "independence condition." I will argue that the important conception underlying these discussions is that a theory should be informative in the sense of generating a number of entailments that are independent from the results used as diagnostics.

First it will be useful to state the main claims advanced by Worrall and others more precisely. Musgrave (1974) draws a useful contrast between "historical" and "logical" accounts of confirmation. According to the former, a theory's provenance has some role in assessing the confirmatory support delivered by the evidence, and this may allow one to evade the paradoxes

[14] Other advocates of this line of thought include Whewell, Duhem, Giere, Zahar, Leplin and Maher. Worrall calls his own view the "heuristic account" of confirmation, but in more recent discussions "use novelty" has become the standard term. For more comprehensive discussions, see Leplin (1997), Chapter 2, and Mayo (1996), Chapter 8.

of a Hempelian approach to confirmation theory by incorporating background knowledge. The difficulty lies in incorporating background knowledge without thereby making confirmation an entirely subjective notion. The "logical" account focuses entirely on the formal relations between theory and evidence, on the grounds that these are all that matters to confirmation theory. Worrall states the guiding idea of his historical account as follows (Worrall 1985, p. 301):

> ...in order to decide whether a particular empirical result supports or confirms or corroborates a particular theory the way in which that theory was developed or constructed needs to be known—more especially, it has to be checked whether or not that empirical result was itself involved in the construction of the theory.

We can render this claim more precisely in two different ways (following Earman 1992, pp. 114-16), leaving the phrase "used in the construction of" ambiguous for now:

UN: Given two theories T, T' that both entail the same evidence claim E, T receives more support than T' from E if E was used in the construction of T' but not T.

UN*: Given two evidence claims E_1, E_2 both entailed by T, T receives more support from E_1 than from E_2 if E_2 is used in the construction of T whereas E_1 is not used.

Furthermore, we can distinguish between a modest and stronger version of the use novelty thesis depending on the extent of our background knowledge regarding the content of a theory and its evidential support. According to the modest version of UN, assessments of novelty are important only when we are mostly ignorant of the content of the theories T, T'.[15] Modest UN states that one should prefer T in this case. Strong UN, on the other hand, holds that use novelty is relevant even when we have rich background knowledge regarding the theory and other evidential support for it. For example, according to the strong UN thesis, the extent to which the observed value of Mercury's anomalous perihelion motion supports GTR depends upon whether this fact was used

[15] Modest and strong versions of UN* can be formulated along the same lines.

in the construction of GTR, even in light of a detailed understanding of the theory's content and independent evidential support in its favor.

Accepting either of these two criteria would represent a significant departure from a Bayesian approach to confirmation theory. Bayesian updating is insensitive to whether E was used in the construction of T; any difference in confirmational support E delivers to T, T' depends upon differences in the priors and the likelihoods. Explicitly, since T entails E, $C(T, E, K) = (Pr(T|K)) \left[\frac{1}{Pr(E|K)} - 1\right]$ and likewise for $C(T', E, K)$. Only the priors and likelihood appear in this equation, and thus the Bayesian can agree with Worrall's intuition only to the extent that use-novelty can be smuggled into assignments of the priors and/or likelihoods. UN holds that $C(T, E, K) > C(T', E, K)$ in cases of use construction, and the Bayesian can concur only if $Pr(T|K) > Pr(T'|K)$; similarly, a Bayesian agent can agree with UN* only in cases where $Pr(E_1|K) < Pr(E_2|K)$. Regarding the first case, it seems plausible that assignment of $Pr(T|K)$ should depend upon how the theory was constructed. However, I see no reason to expect that evidence used in theory construction would always have lower likelihoods than other evidence used in assessing the theory.[16] Even if the Bayesian can juggle priors in order to save UN, she will not be able to account for the difference in incremental confirmation provided by E_1 and E_2.

Staunch Bayesians argue that the intuition underlying UN and UN* is simply misguided; Howson and Urbach (1989), for example, argue that the claim "that a hypothesis designed to fit some piece of data is not supported by it to as great an extent as one which also fits the data accidentally," though "certainly inconsistent" with Bayesianism, is also false (p. 276 ff.). There is also historical evidence that scientists may not agree with UN*: Brush (1989) argues that

[16] Here I am setting aside the problem of "old evidence."

scientists put more weight on general relativity's ability to predict Mercury's perihelion motion than on the confirmation of light bending in the sun's gravitational field, despite the novelty of the latter claim. Clearly Mercury's anomalous perihelion motion fails a temporal novelty requirement, since the result was known before Einstein began working on GTR.

The main line of defense in favor of the strong UN thesis (see, e.g. Worrall 1989) is typically presented as follows. Suppose that we have two theories T and T' that both entail a particular evidence claim E, which happened to be used in the construction of T' but not T. The success of T must have occurred either by chance or due to its accurate representation of the phenomena in question. The former would require an unlikely coincidence, a lucky break in the theory's favor. The truth of T accounts for its ability to produce accurate predictions, so the success of T gives further evidence of its truth. On the other hand, the success of T' can be readily explained in terms of the flexibility of the theory and the ingenuity of the theorists who exploit this flexibility to arrive at the right result. A slight variation on this argument emphasizes that in the latter case the experiment that produced the result E may have shown T to be incorrect, but it could not have falsified T'.

While this argument does have some plausibility, it does not fare well in response to the following objection: in what sense is the truth or use-construction of a theory relevant to the claim that T entails a particular result? The fact that a particular entailment holds for the theory T is not explained by T's truth; the structure of the theory is given and holds necessarily regardless of its truth or falsity. Similarly the truth of T does not have explanatory bearing on the historical contingencies that lead to its introduction: unless we're willing to postulate an extra faculty of judgement to, say, Newton, the truth or falsity of Newtonian theory doesn't have any

bearing on whether Newton did or did not "design" a lunar theory in order to get the proper motion of the lunar apsides.

Any account of novelty should clarify when a particular empirical result qualifies as a novel prediction of a theory or fails to do so. In the passage quoted above Worrall leaves the question of whether a result is "used in the construction of a theory" to historians. Early discussions of novelty focused on temporal novelty: a result unknown at the time a theory is formulated certainly could not have influenced theorists. But the important issue is the *extent* to which a theory depends on a particular result, and Worrall plausibly suggests using the historical development of the theory as a guide. For cases such as blatant parameter fixing, historians may well reach agreement about whether a result is used in the construction of a theory. However, knowledge of observational results shapes theorists' endeavors in a variety of ways, and it is not immediately clear when to discount the evidential support from a particular piece of evidence due to its use. Returning to the shopworn example above, Einstein was aware of the anomalous perihelion motion of Mercury during the development of general relativity. Although that knowledge did serve to guide theory development, Einstein did not "design" the field equations specifically to resolve the anomaly. In intermediate cases where E serves as a guide to theory development—perhaps by ruling out a number of otherwise reasonable lines of inquiry—should the theory still receive an extra confirmatory boost from E, or does E fail to count as "novel" evidence? On Worrall's account, all of these questions should be answered by a thorough examination of the provenance of the theory.[17]

[17] Worrall takes the emphasis on provenance to be an advantage of his view over Zahar's account, which focusd on the novelty of a given result with respect to the set of problems important to the scientist developing the theory: the provenance can be established based on a variety of historical records, whereas the psychological details important to Zahar's account would be almost impossible to determine.

Introducing the theory's provenance as a component of confirmation theory has an unappealling consequence. Suppose, for example, that a scientist "Alice" uses E to set a parameter of a theory T, without recognizing that other theoretical commitments already fix the parameter. Discovering Alice's error might prove quite difficult, yet on Worrall's account the assessment of E's support for T crucially depends on it. Even in cases where historical records allow a rich reconstruction of a theory's development, this reconstruction may fail to identify the role various evidential claims play in the structure of a theory. Alice's error would turn a successful novel prediction into a clear case of using evidence in the construction of the theory. As a result, Alice's theory would receive less support from E than that of another scientist "Bob," even if Bob's theory differed only in that Bob realized that parameter-fixing was unneccessary. One cannot ask whether E supports T without further specifying the provenance, thereby relativizing the confirmatory status of the evidence to the details of a theory's development. For advocates of a logical approach to confirmation, examples like this one merely illustrate the disastrous consequences of incorporating a historical component in confirmation theory: new historical details regarding a theory's development would lead to complete re-evaluations of a theory's evidential support. While historians would perhaps rejoice in their new status as the final appellate court in the tribunal of experience, this conclusion has the ring of a *reductio ad absurdum*.

Rather than abandoning a historical approach entirely in light of such examples, there is an important insight worth retaining that need not depend on detailed reconstructions of a theory's provenance. Whether a given result is novel with respect to a theory depends solely on the structure of the theory. Theorists may draw on an incredibly rich background of experimental results and theoretical commitments in order to produce a theory which, in the end, is only related to a small subset of this background knowledge. Even though a theory's provenance is

often a clear guide to the relation a theory bears to a particular evidential claim, it need not be. Recently Leplin has developed an account of novelty (Leplin 1997, Chapter 3), part of which I will adopt here. On Leplin's analysis, an evidence claim E is novel with respect to a theory T if the following two conditions hold:[18]

> *Independence Condition:* There is a minimally adequate reconstruction of the reasoning leading to T that does not cite any qualitative generalization of E.
>
> *Uniqueness Condition:* There is some qualitative generalization of E that T explains and predicts, and of which, at the time that T first does so, no alternative theory provides a viable reason to expect instances.

A qualitative generalization of E is the general type of effect or phenomenon of which E is a token; for example, various predictions for the degree of light bending around the sun would all count as members of the qualitative generalization of the general relativistic prediction. The independence condition is meant to distinguish those results actually used in constructing the theory from those which may have been known, but play no role in motivating the theory. Briefly, a minimally adequate reconstruction clarifies the essential steps in the reasoning leading to the basic hypotheses of a theory T.[19] Leplin argues that such a reconstruction takes the form of an argument leading to the theory T from premisses taken from among the following three types: specific empirical results or generalizations of them, pragmatic appeals, and higher level methodological constraints. Among the second type, Leplin has in mind primarily simplicity constraints, and he gives symmetry principles in particle physics as an example of the third type. A reconstruction is adequate if it accurately conveys the reasons for proposing and considering

[18] I have made minor modifications of Leplin's notation to accord with mine (see Leplin 1997, p. 77, for the original formulation).

[19] The basic hypotheses are those which constitute a theory, in the sense that abandoning them is incompatible with retaining T—for example, abandoning Einstein's field equations would mean abandoning general relativity, although one could certainly abandon various standard energy conditions in GTR without likewise abandoning the theory.

the theory. To further qualify as minimally adequate the reconstruction must satisfy two additional requirements: (1) none of the empirical results used in the reconstruction can be replaced by a logically weaker proposition, and (2) the conjuction of the premisses cannot be logically simplified. Thus, an evidence claim which satisfies the independence condition plays no role in the construction of T in the sense that the theory can be independently motivated without using a generalization of E.

Although Leplin further imposes the uniqueness condition, I take this to be a mistake—and it is even inconsistent with Leplin's other commitments. The uniqueness condition relativizes the notion of novelty to the class of competing theories, since a result only counts as novel if T alone predicts its occurence. Leplin elsewhere asserts that the novelty should be a binary relation between a theory and evidence, yet here his own account fails this requirement. I agree with Leplin's insistence that whether a result is novel for a particular theory should depend solely upon the evidence claim and the theory, and so I will drop the uniqueness condition. This account is still "quasi-historical" in two senses. First, the minimally adequate reconstructions may shift substantially over time due to changes in basic hypotheses as well as higher order methodological constraints. For example, new theoretical developments may dictate the values of parameters previously treated as free parameters constrained by qualitative generalizations of E. Second, assessments of independence may also be subject to revision based on earlier incomplete or inaccurate knowledge of a theory's content and entailments.

In conclusion, in developing this modified version of Leplin's account I have argued for a conception of independence rather than novelty. To go further in this direction I should establish that in historical cases used to support the importance of novelty (such as those discussed by Worrall 1989), independence is really at issue. In addition, I have not supplied a convincing

epistemic rationale for independence. But rather than taking up these issues here, I will turn to the implications of this discussion for the assessment of inflationary cosmology.

7.4 Inflation and Independence

Applying Leplin's analysis to the case of inflation helps to clarify the importance of the link between fundamental particle physics and inflation. The spectra of density perturbations have often been cited as a novel success for inflation: the theory was developed with no consideration given to large-scale structure formation. In a minimal reconstruction of Guth's development of the original theory of inflation, qualitative generalizations of results regarding structure formation or density perturbations have no place. Thus the theory provides a mechanism for producing density perturbations "for free," without any tinkering. Despite this apparent success, the same prediction has often been cited as a blatant example of the theory's malleability—without tinkering, the theory yielded density perturbations with an amplitude several orders of magnitude too large. The changes made to bring the prediction within observational constraints led to the introduction of the inflaton field, whose potential was adjusted to insure density perturbations with the correct amplitude.

Consider a purely phenomenological approach to inflation: the properties of the potential $V(\phi)$ are to be inferred from observation, with little or no effort devoted to finding a place for the inflaton in a particular particle physics model. Suppose for the sake of argument that the basic features of inflation are motivated by the horizon and flatness problems. In the minimally adequate reconstruction, these will be invoked to justify the introduction of a stage of inflaton-driven exponential expansion. In order to constrain the model, a minimally adequate reconstruction must include qualitative generalizations of the crucial CMBR observations (fixing,

for example, the overall amplitude of the density perturbations). These features would not be independent with respect to this set of data; however, features independent of the specific form of the inflaton potential (such as the consistency equation) would be. Incorporating constraints from particle physics would allow one to increase the number of independent predictions, since the qualitative generalization from CMBR measurements could be replaced by arguments from fundamental physics in the reconstruction of inflation.

7.5 Robustness Revisited

Above I listed several allegedly robust predictions of inflation related to the spectrum of density perturbations. In doing so I neglected recent research which undermines the robustness of even these predictions. The mechanism by which inflation produces density fluctuations in the early universe has been studied in detail from the early 80s onward. Calculations of the spectra of density perturbations produced during inflation rely on a number of assumptions; recently some of these assumptions have come under fire.[20] In particular, recent research by Brandenberger and Martin shows that the standard inflationary mechanism for producing density perturbations is sensitive to hidden assumptions about super-Planck scale physics. Another line of research suggests that the inflaton field's evolution during "preheating" may substantially alter the spectrum of density perturbations left over from earlier stages of inflation due to relativistic effects. In this section, I will briefly review this work and its implications for testing inflation.

One of the main appeals of inflation is its apparent independence from Planck-scale physics: almost all of the standard calculations use a scalar field weakly coupled to gravity, or in

[20]This brief review is by no means exhaustive—these are simply two of the criticisms of the standard calculations that have been brought to my attention.

a fixed background spacetime. The inflaton field is assumed to be in a vacuum state prior to inflation, and by evolving the field forward using the linearized equations of motion one computes the spectrum of fluctuations remaining after the inflationary stage.[21] Exponential expansion during inflation stretches pre-inflationary length scales dramatically, and for some models of inflation the fluctuations on cosmological length scales after inflation started out as fluctuations with a physical length shorter than the Planck length ($l_p = 8.10 \times 10^{-33} cm$). As Brandenberger and Martin (Martin and Brandenberger 2001; Brandenberger and Martin 2001) point out, one expects quantum field theory on a flat background (or even the semi-classical approximation) to break down at these length scales. To estimate the possible impact of new fundamental physics, they alter the dispersion relations for the initial vacuum state of the inflaton field above some momentum cutoff. Although some choices of modified dispersion relations do not alter the usual results, others do; Brandenberger and Martin conclude that:

> Our work indicates that the prediction of a scale-invariant spectrum in inflationary cosmology depends sensitively on hidden assumptions about super-Planck scale physics. (Martin and Brandenberger 2001, p. 2)

If inflation depends sensitively on super-Planck scale physics, as these results suggest, then the goal of establishing the robust predictions of inflation cannot be attained without significant advances in fundamental physics.

The assumptions underlying the standard calculations may also break down during a stage of the inflaton's evolution called "preheating." Early inflationary models assumed that following the inflationary stage, the inflaton energy density would be converted into normal, thermalized matter. The temperature drops during inflation as energy density is diluted, but the

[21] See, e.g., Liddle and Lyth (2000), Chapter 4 for a discussion and justification of the assumptions which go into this calculation.

inflaton decay “reheats” the universe as the inflaton transfers energy to other fields. Subsequent research regarding the decay of the inflaton has shown that reheating may be preceded by a “preheating” stage, during which non-equilibrium processes lead to much more violent and rapid energy transfer to other fields (associated with parametric resonance). This process takes place well after the spectrum of density perturbations has been imprinted on the early universe, but recently Bassett and his collaborators (Bassett et al. 1999, 2000) have argued that various effects produced during preheating could alter the spectrum. In particular, rather than assuming a fixed background spacetime throughout preheating, they assess the effects of metric perturbations coupled (via Einstein’s equations) to the perturbations of the inflaton field. Although the effects of these metric perturbations are highly model-dependent, in some cases they dramatically alter the spectra of density perturbations, leading to a spectra resembling that predicted by topological defect theory rather than the standard inflationary prediction. This research indicates that the usual neglect of relativistic effects is unjustified during the period of preheating, but whether these effects will wash out the standard spectra of density perturbations for a large class of inflationary models remains to be seen.

Both lines of research highlight the difficulty of finding robust predictions for inflation as long as the connection between inflation and fundamental physics is up for grabs. For some inflationary models, the effects of super-Planck scale physics and metric perturbations during preheating do not alter the standard predictions, but in others the implicit assumptions in the standard calculations neglect a variety of interesting effects. Thus, Liddle and Lyth (2000)’s optimistic claim that inflation’s prediction for the spectrum of density perturbations does not share the weakness of earlier predictions (namely, their malleability and model-dependence) may be premature.

Appendix A

Aspects of Relativistic Cosmology

A.1 A Primer on General Relativity

In this appendix I will briefly introduce the general theory of relativity, after filling in a few pieces of the mathematical and technical background needed to formulate the theory. I have tried to make this review relatively concise and self-contained, but the reader should consult any of a number of detailed treatments of general relativity (such as Wald 1984), and more comprehensive introductions to the mathematical ideas (such as Geroch 1985).

GTR is typically presented using the formalism of tensor analysis on differentiable manifolds. Roughly, a differentiable manifold can be thought of as a topological space that locally looks like $\mathbb{R}^n$. More precisely, a C^∞ differentiable manifold consists of a topological space M along with a maximal C^∞ atlas. A topological space is a set M with a family $\{O\}$ of subsets (the open sets) satisfying the following properties:

(i) $\{\emptyset, M\}$ belong to $\{O\}$

(ii) $\{O\}$ is closed under unions of an arbitrary collection of members of $\{O\}$

(iii) $\{O\}$ is closed under the intersection of a finite number of members of $\{O\}$.

A map between topological spaces $f : M \rightarrow N$ is continuous if the inverse image $f^{-1}(O)$ of every open set $O \subset N$ is an open set in M. A continuous map that is also one-to-one and onto, with a continuous inverse, is called a *homeomorphism*. A *coordinate chart* consists of a pair

S, ϕ where S is an open subset of M, ϕ is a homeomorphism, and $\phi(S)$ is an open subset of $\mathbb{R}^n$. A C^∞ *atlas* consists of a collection of coordinate charts covering M that are compatible. Coordinate charts covering overlapping subsets $S_1 \cap S_2 \neq \emptyset$ of M are said to be C^∞ compatible if the composition $\phi_2 \circ \phi_1^{-1} : \phi_1(S_1 \cap S_2) \to \phi_2(S_1 \cap S_2)$ and its inverse are both C^∞. Since this map and its inverse are both maps from $\mathbb{R}^n$ to $\mathbb{R}^n$, we are using the standard definition of C^∞: a mapping is C^∞ if all higher order derivatives exist and are continuous. Finally, maximality requires including all of the compatible coordinate charts in the atlas; otherwise, different choices of charts would lead to "different" manifolds. Hopefully this string of definitions hasn't obscured the underlying picture: one can think of a differentiable manifold as "pieces of $\mathbb{R}^n$" (the subsets mapped into $\mathbb{R}^n$ by coordinate charts) pasted together consistently (as required by the compatibility conditions between charts). Riemann was the first to recognize that (using modern terminology) the extra flexibility of requiring only local similarity to $\mathbb{R}^n$ allows one to treat a wide variety of spaces with very different global features.

Vector and tensor fields are defined on a differentiable manifold by introducing the tangent space and dual space at each point, along with every conceivable combination of these spaces. The typical characterization of a vector in $\mathbb{R}^n$ as an n-tuple of components relative to some chosen coordinates does not generalize to manifolds, which usually lack a global chart. However, the idea of a *directional derivative operator* can be generalized easily. In $\mathbb{R}^n$ these operators are in one-to-one correspondence to vectors. Denote the collection of C^∞ functions from open subsets of $\mathbb{R}^n$ ($O \subset \mathbb{R}$) containing p to $\mathbb{R}$ by S_p. The directional derivative of $f \in S_p$ in the direction of a vector $u = (u^1, ...u^n)$ is given by:

$$u(f) = u^1 \frac{\partial f}{\partial x^1}\bigg|_p + ... + u^n \frac{\partial f}{\partial x^n}\bigg|_p \tag{A.1}$$

Directional derivative operators are linear and obey the Leibniz rule (i.e., $u(fg) = f(p)u(g) + u(f)g(p)$ for $f, g \in S_p$) as a consequence of the properties of partial derivatives. This suggests an immediate generalization: define a vector at a point p in an arbitrary differentiable manifold M to be a map $u : S_p \to \mathbb{R}$ sharing precisely these properties, where S_p is the collection of C^∞ functions from neighborhoods O of a point p to $\mathbb{R}$. (Alternatively, one can introduce tangent vectors geometrically as tangents to smooth curves passing through the point p.) The collection of vectors at a given point naturally has the structure of a vector space, and it is called the tangent space, T_p. T_p has the same dimension as the manifold. In addition, the *dual space* T_p^* is a vector space (with the same dimension as T_p) composed of linear functionals that map elements of T_p into $\mathbb{R}$. Elements of the tangent space T_p are often called "contravariant vectors," symbolized by "upstairs indices" v^a (in the so-called "abstract index notation"), whereas elements of the dual space are called "covaraint vectors" (aka one-forms or functionals), and are written with downstairs indices, ω_a. A contravariant vector field (resp., covariant) is a map that assigns an element of T_p (T_p^*) to each point in M. A vector field v is said to be smooth (C^∞) if for all smooth functions $f \in S_p$, the map $v(f) : O \to \mathbb{R}$ defined by $v(f)(p) = v|_p f$ is smooth. (Non-smooth vector fields are almost never used – sometimes smoothness is even included as part of the definition of a vector field.) Finally, a tensor of type (r, s) is a multilinear map:

$$\underbrace{T_p \otimes \cdots \otimes T_p}_{r\ times} \otimes \underbrace{T_p^* \otimes \cdots \otimes T_p^*}_{s\ times} \to \mathbb{R}, \tag{A.2}$$

where *multilinear* means that the map is linear in each variable treated separately with the other variables fixed. An (r, s) tensor is written as $T^{ab...}_{mn...}$, with r upstairs indices $a, b, ...$ and s downstairs indices $m, n, ...$ These maps naturally have the structure of a vector space, and a tensor

field assigns an element of the appropriate tensor space constructed out of T_p and T_p^* to each point.

With this brief tour of the mathematical formalism we now turn to general relativity. GTR models spacetime as a four-dimensional manifold M equipped with a metric field g_{ab}. The manifold consists of the set of spacetime points, and the fact that 4 numbers are needed to uniquely specify a spacetime point fits nicely with the requirement of "local" correspondence with $\mathbb{R}^4$. In addition, the continuity of the manifold matches the continuity of spacetime at the classical level. The manifold is usually assumed to be Hausdorff, connected, paracompact, and without boundary. A topological space is *Hausdorff* if there are always disjoint open sets O_p, O_q containing any distinct points $p, q \in M$. A *connected* topological space is not the union of any two disjoint open sets. The closure of a set O (written $\bar{O}$) is the intersection of all closed sets containing O, and the interior of the set O is the union of all open sets contained in O. The *boundary* of a given set O (written $\dot{O}$) is the set of points in $\bar{O}$ but not in the interior of the set O. (It follows directly from these definitions that open sets do not contain their boundary points, whereas closed sets contain all of their boundary points.) The definition of paracompactness is more involved (see, e.g., Wald 1984, pp. 426-427), so here I will only note one important consequence: any paracompact manifold is homeomorphic to a metric space.

The introduction of a metric tensor field adds geometrical structure to the manifold topology. Intuitively, the metric g_{ab} supplies a measure of distance for small displacements on the mainfold. Acting on a single vector, the metric is an invertible, one-to-one and onto map from T_p into T_p^*. This map corresponds to "raising and lowering" of indices as follows: $g_{ab}u^a = u_b$ and $g^{ab}u_b = u^a$, where the inverse of the metric g^{ab} is defined by $g_{ab}g^{bc} = \delta_a^c$ (and δ_a^c is the

identity map). This gives us a natural inner product between covariant and contravariant vectors. But the metric can also be characterized as a map from $T_p \times T_p \rightarrow \mathbb{R}$, satisfying two additional requirements. First, it is *symmetric*, which means that the order of vectors does not matter: $g_{ab}u^a v^b = g_{ab}v^b u^a$, or more succinctly $g_{ab} = g_{ba}$. In addition, g_{ab} is required to be *non-degenerate*, which holds if the determinant of the metric is non-zero. In sum, the metric field is defined to be a symmetric, non-degenerate rank two tensor field defined everywhere on M. For a given coordinate chart, the metric can be written as $ds^2 = g_{ij}dx^i dx^j$ with summation over i, j understood, and g_{ij} are the components of the metric tensor in this chart. In the tangent space, one can always find a coordinate basis that "diagonalizes" the metric, i.e. for basis vectors $e^1 ... e^n$, $g_{ij}e^i e^j = 0$ for $i \neq j$, and $= \pm 1$ for $i = j$. The *signature* of the metric specifies the sign of the diagonal elements; GTR uses a *Lorentzian* metric with signature written as either $(-+++)$ or $(+---)$, where I will take the former convention throughout. This signature gives the tangent space the familiar light cone structure of Minkowski space; tangent vectors $u^a \in T_p$ are classified as follows: timelike, $g_{ab}u^a u^b < 0$; spacelike, $g_{ab}u^a u^b > 0$; and null, $g_{ab}u^a u^b = 0$.

Next we need to introduce the structure used to compare vectors from tangent spaces at different points: the *affine connection*. Roughly, the connection generalizes the notion of parallelism from Euclidean space to curved spaces. For a vector $u^a \in T_p$ and a curve in M connecting p to q, the connection can be used to define a vector u'^a that results from "moving u^a along the curve, keeping it parallel to itself" (parallel transport). In Euclidean space, parallely transporting a vector around a closed loop produces the same vector, but for a curved manifold this is generally not the case. The sphere is a standard example: carry a tangent vector on the surface of a sphere around a triangle consisting of a quarter of the equator and an excursion up

to the north pole and back; this yields a vector rotated 90° with respect to the original. The most natural way to define the connection is in terms of the fiber bundle formalism, which I will not introduce here.[1] But I will note that there is a unique connection, called the Levi-Civita connection, satisfying the following three requirements: (i) the "output" u'^a depends linearly on u^a, (ii) the connection is compatible with the metric, and (iii) there is no torsion. The second condition holds if a parallely transported tangent vector does not change length; formally, $\nabla_a g_{ab} = 0$. Finally, (iii) can be characterized as follows. Consider parallely transporting a vector (in $\mathbb{R}^n$) ϵu^a in the v^b direction by an infinitesimal amount ϵ, and similarly transport ϵv^b in the direction u^a. If the resulting figure is a closed parallelogram (to order ϵ^2), the connection is torsion free, and the vectors have not "rotated" under parallel transport. The connection is often first introduced in the guise of the *covariant derivative* operator, denoted ∇_a. Taking the derivative of a vector field in $\mathbb{R}^n$ requires comparing the value of the field at a point with the value at neighboring points; the connection makes it possible to give a path-dependent comparison of vectors in neighboring tangent spaces. In addition, it is linear and satisfies the Leibniz rule characteristic of derivatives. For given coordinates, the covariant derivative of a vector is given by $\nabla_a v^b = \partial_a v^b + \Gamma^b_{ac} v^c$, where Γ^b_{ac} are the coordinate components of the connection. (This formula generalizes in a straightforward way to covariant derivatives of tensor fields.)

We will need two more concepts before formulating the field equations. First, in a curved geometry the straightest possible lines are those that remain parallel to themselves. A *geodesic* is a curve such that its tangent vector remains parallel to itself as it is parallely propagated along

[1] In some senses the fundamental insights of GTR can be introduced more clearly using the connection and the fiber bundle formalism (and it is clearer how GTR generalizes Newtonian gravitation); see Stachel (2001) for a counterfactual fable in which "Newstein" formulates GTR using the affine connection. See Baez and Muniain (1994) for a very readable introduction to the fiber bundle formalism, which I draw on here.

the curve; symbolically, $v^a \nabla_a v^b = 0$. (The right hand side of this equation is 0 only if the curve is affinely parameterized, otherwise a non-zero term function of the parameter appears on the right.) Second, the path dependence of parallel transport depends upon the curvature of the surface, which is fully characterized by the *Riemann curvature tensor* $R_{abc}{}^{d}$. The connection with parallel transport can be understood as follows. Imagine parallely propagating a vector u^a around a parallelogram whose sides are given by ϵv^a and ϵw^a, where ϵ is a small number. To order ϵ^2, the result of parallel transport, u'^a, is given by

$$u'^a = u^a - \epsilon^2 R_{bcd}{}^{a} v^b w^c u^d. \tag{A.3}$$

The second term on the RHS characterizes the departure from flat space. The Riemann curvature tensor can be expressed directly in terms of the metric tensor and its first and second derivatives, although I will not provide the details here.

Curvature is manifested directly by geodesic deviation. Suppose we have a family of "infinitesimally neighboring" timelike geodesics, such that the tangent to the worldlines is u^a, and ξ^a is the "deviation vector" between neighboring curves (i.e., a spacelike vector orthogonal to u^a). Here it may be useful to visualize a swarm of particles: changes in ξ^a correspond to the swarm converging or diverging. The "relative acceleration" of neighboring geodesics is then given by (see, e.g., Wald 1984, pp. 46-47):

$$a^c = u^a \nabla_a (u^b \nabla_b \xi^c) = -R_{abd}{}^{c} \xi^b u^a u^d \tag{A.4}$$

This equation contains a great deal of information regarding how the congruence of geodesics responds to curvature, but very roughly, for positive curvature the geodesics converge, whereas negative curvature leads to divergence. Two tensors derived from $R_{abc}{}^{d}$ by contraction appear in EFE. The contraction of a rank n tensor is an $n-2$ tensor obtained by summing over repeated indices. The *Ricci tensor* is defined by $R_{ac} = R_{abc}{}^{b}$. Further contracting the Ricci tensor, we have the *Ricci scalar* $R = R^{a}_{a}$.

Now we have introduced two of the quantities appearing on the left hand side of EFE:

$$R_{ab} - \frac{1}{2}Rg_{ab} + \Lambda g_{ab} = \kappa T_{ab}. \tag{A.5}$$

The right hand side is based on one of the hard-won insights of early relativistic mechanics: the appropriate mathematical object for representing momentum and energy density is the stress-energy tensor, a symmetric rank two tensor that obeys the covariant conservation law $\nabla^{a} T_{ab} = 0$ (more on this in a moment). The left hand side is the most general symmetric tensor that can be constructed from the metric and its first and second derivatives. The cosmological constant Λ can be treated as a geometric term in the EFE (appearing on the LHS), but it can also be introduced on the RHS as a component of T_{ab} — although usually a field only "mimics" a true Λ term.

The stress-energy tensor fully characterizes the energy and momentum of a system; roughly, it keeps track of energy density and momentum flow in different directions. Contracting with two vectors at a point gives $T_{ab}u^{a}v^{b}$, which represents the flux of four-momentum u^{a} flowing through the point in the v^{b} direction. Cosmologists often use "perfect fluids," where perfect means that the fluid completely lacks pressure anisotropies, viscosity, and shear stresses. As a result the only non-zero components lie along the diagonal if T_{ab} is written out as a matrix;

the T_{tt} component represents the energy density ρ, and the only nonzero spatial components T_{ii} (where $i = x, y, z$) give the pressure p. The stress-energy tensor for a perfect fluid is given by

$$T_{ab} = (\rho + p)u_a u_b + p g_{ab}, \tag{A.6}$$

where u^a represents the velocity of the perfect fluid. More generally, the stress-energy tensor for a given field is defined in terms of a variational derivative of its Lagrangian (see, e.g., Appendix E in Wald 1984).

A diffeomorphism is a one-to-one, onto map $\phi : M \rightarrow M'$ that is C^∞, with a C^∞ inverse; this preserves all manifold structure. See, for example, Wald (1984, Appendix C) for a discussion of how diffeomorphisms act on vector, dual vector, and tensor fields. An isometry is a diffeomorphism ϕ that maps the metric into itself, i.e. $\phi * g_{ab} = g_{ab}$.

A.2 FLRW models

Here I will briefly review the ubiquitous FLRW models and highlight the aspect of FLRW dynamics dubbed the "flatness problem." These models are based on the fundamental assumptions of homogeneity and isotropy. Homogeneity and isotropy together entail that the models are topologically $\Sigma \times \mathbb{R}$, where the three-dimensional surfaces Σ are orthogonal to the wordlines of fundamental observers. The spatial geometry induced on the surfaces Σ is such that there is an isometry carrying any $p \in \Sigma$ to any other point lying in the same surface (homogeneity), and at any p the three spatial directions are isometric (isotropy).

Any line element compatible with these symmetry requirements can be written in the following form:

$$ds^2 = -dt^2 + a(t)^2 \left(\frac{dr^2}{1 - kr^2} + f(r)^2 (d\theta^2 + \sin^2 \theta \, d\phi^2) \right), \tag{A.7}$$

where k classifies the three distinct possibilities for the curvature of the Σ surfaces: positive, $k = +1$ corresponding to spherical space, with $f(r) = sin\, r$; zero for flat space, $f(r) = r$; and negative curvature ($k = -1$) corresponding to hyperbolic space, $f(r) = sinh\, r$. Furthermore, imposing isotropy and homogeneity implies that the source term is the stress-energy tensor for a perfect fluid, eqn. (A.6). Both p and ρ are functions only of the cosmic time t, with no spatial variation. (Alternatively one can stipulate that the source term is given by this equation and derive the line element above.)

In general, the EFE reduce to a set of 10 coupled non-linear equations, but for isotropic and homogeneous models we have a much simpler situation: two independent equations. Take v^a to be the normalized four velocity of the perfect fluid, and choose the coordinates of the line element above, in which $v_t = 1$ and $v_i = 0$ where i ranges over the spatial coordinates (r, θ, ϕ). Symmetry dictates that the Einstein tensor in these coordinates, $G_{\mu\nu}$, is non-zero only for G_{tt} and G_{ii}.[2] Rewriting the field equations for G_{tt} in terms of the scale factor a, we have the "Friedmann equation":

$$\left(\frac{\dot{a}}{a} \right)^2 + \frac{k}{a^2} = \left(\frac{8\pi}{3} \right) \rho + \frac{\Lambda}{3}. \tag{A.8}$$

[2] A nonzero component G_{it} would provide a geometrically preferred spatial direction, in violation of isotropy, and isotropy further implies that the metric of the homogeneous surfaces is diagonal in these coordinates. See, in particular, Weinberg (1972), Chapter 15, and Wald (1984), Chapter 5 for clear derivations of the FLRW dynamics.

The G_{ii} terms all yield the same equation, namely

$$2\frac{\ddot{a}}{a} + \left(\frac{\dot{a}}{a}\right)^2 + \frac{k}{a^2} = -8\pi p + \Lambda. \tag{A.9}$$

The difference of these two equations yields the following equation for the evolution of the scale factor:

$$\frac{\ddot{a}}{a} = -\frac{4\pi}{3}(\rho + 3p) + \frac{\Lambda}{3}. \tag{A.10}$$

Finally, the following equation represents stress-energy conservation:

$$\dot{\rho} = -3(\rho + p)\frac{\dot{a}}{a}. \tag{A.11}$$

Derivations of FLRW dynamics often use the equations (A.8, A.11) as their starting point, but in fact the three equations (A.8, A.9, A.11) are *not* independent due to the Bianchi identities.

Rewriting the Friedmann equation (A.8) in terms of the density parameter, Ω, which represents the total energy density and is defined as the ratio $\Omega \equiv \frac{\rho}{\rho_{crit}}$, brings the flatness problem into focus. The critical density is the value of ρ such that $k = 0$ in the Friedmann equation, namely

$$\rho_{crit} = \frac{3}{8\pi}\left(H^2 - \frac{\Lambda}{3}\right), \tag{A.12}$$

where I have introduced the inappropriately named (time dependent) Hubble "constant," $H = \frac{\dot{a}}{a}$. Now plugging all of these new quantities into the Friedmann equation, after some algebra we have:

$$\frac{|\Omega - 1|}{\Omega} = \frac{3|k|}{8\pi\rho a^2}. \tag{A.13}$$

If the rest-mass energy density is negligible, the energy density is diluted with expansion, $\rho \propto a^{-3\gamma}(t)$ (with $\gamma = 4/3$ for radiation and $\gamma = 1$ for pressureless dust, for example). Thus, the time dependence of the density parameter is related to the scale factor as follows:

$$\frac{|\Omega - 1|}{\Omega} \propto a^{3\gamma-2}(t) \tag{A.14}$$

As the scale factor grows with the expanding universe, if the value of Ω differs from 1 it evolves rapidly away from 1; in other words, the value $\Omega = 1$ is an unstable fixed point for dynamical evolution governed by the Friedman equation (with $\gamma > 2/3$). Despite this instability, observations indicate that the present value of the density parameter lies in the range $0.1 \leq \Omega_0 \leq 1.5$.[3] In order to fit these observations, the value of the density parameter at the Planck time ($t_p = 10^{-43}s$) must have been extraordinarily close to one, $|\Omega(t_p) - 1| \leq 10^{-59}$.[4] Thus the early universe must have been incredibly close to the "flat" FLRW model, $\Omega = 1, k = 0$. This straightforward calculation turns into the flatness problem when one adds the judgement that such a finely tuned value of $\Omega(t_p)$ is highly unlikely.

Two points follow immediately from the equations above. First, it is clear how to alter the Friedmann equation to insure that dynamical evolution drives $\Omega(t)$ towards rather than away

[3] This is a very conservative estimate of Ω_0. Briefly, evidence places different constraints on the different components of Ω. Coles and Ellis (1997), pp. 199-203, give a best estimate of $\Omega_0 \approx 0.2$, with $\Omega_{baryon} < 0.10$ based on element abundances and the total matter contribution (including non-baryonic dark matter) $\Omega_M \approx 0.20$ based primarily on estimates of the masses of galaxy clusters and large scale motions. The observational constraints on Ω_Λ are much weaker than those on the matter density, since the overall effect of a small non-zero Λ on large scale motions is quite small. Coles and Ellis (1997) favor $\Omega_0 = \Omega_M \approx 0.2$, but models with $\Omega_0 \approx 1$—with $\Omega_\Lambda \approx 0.80$ to be accounted for by "dark energy" or "quintessence"—are compatible with observations. Observations of the magnitude-redshift relation for type Ia supernovae support a universe with a large Ω_Λ component. Following the report of these results in 1998 (and for a number of other reasons), a model with $\Omega_{total} \approx 1$, with $\Omega_\Lambda \approx 0.7$ and $\Omega_M \approx 0.3$ has become widely accepted.

[4] See, e.g., Blau and Guth (1987), pp. 532-534 for this calculation. The flatness problem is often described in terms of the balance between the kinetic energy of expansion (the H^2 term) and the gravitational potential energy (the ρ term) in equation A.8 (Dicke and Peebles 1979).

from one. The value of γ in (A.14) depends upon the equation of state of the idealized fluid one takes to represent the matter-energy content of the universe. The equation of state of a perfect fluid may be written as $p = (\gamma - 1)\rho$, where p is the pressure, ρ the density, and the index γ is used to classify different types of fluids. From equation (A.14), for $\gamma < 2/3$, the density parameter evolves towards one. In the inflationary scenario, during the inflationary stage $\gamma = 0$ (for most models of inflation) and the density parameter is driven towards one. (In power-law inflation or "coasting" models, during the inflationary stage $\gamma = 2/3$, so the inflationary stage does not drive Ω towards one for these models.) Thus, an inflationary stage can serve to drive arbitrary values of $\Omega(t_{_p})$ closer to one by the end of the inflationary stage, enlarging the range of initial values compatible with observational constraints on $\Omega_{_0}$. For the usual models of inflation, $|\Omega_{_0} - 1| << 1$, although there are a number of "open models" that yield $\Omega_{_0} < 1$ (see, e.g., Lyth and Riotto 1999). Second, the usual rendition of the flatness problem neglects to mention that for *all* FLRW models, if $\gamma > 2/3$ then $\lim_{a\to 0} \Omega = 1$, regardless of the value of $\Omega_{_0}$. Due to this feature of Friedmann dynamics, extrapolating *any* value of $\Omega_{_0}$ back to the Planck era yields an initial value of $\Omega(t_{_p})$ very close to one. In other words, the early universe is "flat" regardless of the present energy density.

I will close with a brief comment comparing the kinematical quantities (the shear, vorticity, and volume expansion) in the FLRW models to a more general (non-symmetric) case. In general these quantities are local characteristics of the "flow" of fundamental observers, but in the FLRW models the volume expansion is directly related to a global quantity, the scale factor $a(t)$. One can treat the overall matter distribution as a fluid and use the average velocity of matter at each point to define a 3 + 1 split (projecting into the spacelike surface orthogonal to the velocity vector), and then define kinematical quantities characterizing the fluid flow. More

precisely, given a congruence of timelike geodesics v^a, normalized so that $v^a v_a = -1$, one can define a projection tensor h_{ab} into the subspace of the tangent space orthogonal to v^a as: $h_{ab} = g_{ab} + v_a v_b$. The fundamental kinematical quantities are then defined by:

$$\nabla_b v_a = \frac{1}{3}\theta h_{ab} + \sigma_{ab} + \omega_{ab}. \tag{A.15}$$

θ is a scalar quantity representing volume expansion along the flow, σ_{ab} measures volume-preserving shear, and ω_{ab} is the vorticity tensor, which fixes both an axis and speed of rotation. (See, for example, section 2 of Ellis and van Elst (1999) for a concise review of this approach.) For the "flow" defined by fundamental observers in an FLRW model, the shear and vorticity both vanish and the volume expansion is directly related to the scale factor, $\theta = 3(\dot{a}/a)$; for non-FLRW models one can take this equation as a *definition* of the scale factor, demoted to a local rather than global property.

A.3 Horizons

The cosmologists' lexicon includes a number of different entries under "horizon," but behind this sometimes bewildering variety lies the common idea that a horizon measures the maximum distance light travels during a fixed time period. Rindler (1956)'s classic paper defined an observer's particle horizon as a surface in a three-dimensional hypersurface of constant cosmic time t_0 dividing the fundamental particles which could have been observed by t_0 from those which could not.[5] More intuitively, the worldlines of particles inside this horizon intersect the observer's past light cone, whereas the worldlines of particles beyond the horizon do not.

[5] See Ellis and Rothman (1993) for a more recent (and remarkably clear) discussion of horizons.

Rindler's definition applies to fundamental observers/particles in an FLRW model, i.e. those moving along geodesics. Although it would be preferable to remove this restriction (as MacCallum (1971) pointed out, the focus on fundamental particles leads to difficulties in applying Rindler's definition to non-FLRW spacetimes), in the following I will stick with standard usage.

The coordinate distance traveled by a light signal emitted from an observer at r_e with emission time t_e reaching another observer at $r_0 = 0, t_0$ is given by:

$$u = \int_0^{r_e} \frac{dr}{\sqrt{1 - kr^2}} = \int_{t_e}^{t_0} \frac{dt}{a(t)}. \tag{A.16}$$

This calculation is quite simple due to the symmetry of the FLRW model: thanks to isotropy we can focus on radial null geodesics without loss of generality, setting $d\theta = d\phi = 0$. For null geodesics $ds^2 = 0$, leading to the following equality for radial null geodesics: $\frac{dr}{\sqrt{1-kr^2}} = \pm\frac{dt}{a(t)}$, where $+$ corresponds to inward-going geodesics, and $-$ to outward-going.. Although the behavior of $a(t)$ must be specified in order to calculate the integral, the integral converges as long as $a(t) \propto t^n$ with $n < 1$, which holds for matter or radiation dominated FLRW models. The physical distance from the observer at r_0 to the horizon, measured at t_0, is then $d = a(t_0)u$. 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).

By changing the limits of integration we can define three different horizon distances. Taking the time of emission to approach the initial singularity we have an expression for the particle horizon:

$$d_{ph} = \lim_{t \to 0} a(t_0) \int_t^{t_0} \frac{dt}{a(t)} \tag{A.17}$$

This integral converges in the FLRW models, yielding a finite value for d_{ph}. Light emitted from fundamental particles at a distance greater than d_{ph}, even the light emitted from arbitrarily early times, cannot reach an observer at $r = 0$ by the time t_0. The past light cones of points separated by a distance greater than d_{ph} do not overlap. The visual horizon and primeval particle horizon (following the terminology of Ellis and Stoeger 1988) are defined in terms of the decoupling time t_d, when photons decouple from matter (before t_d the early universe is opaque). The visual horizon measures the distance of the farthest visible fundamental particles (whose light emitted at or after the decoupling time t_d reaches r_0 by t_0):

$$d_{vh} = a(t_0) \int_{t_d}^{t_0} \frac{dt}{a(t)} \tag{A.18}$$

Finally, the primeval particle horizon is simply the particle horizon evaluated at t_d:

$$d_{pph} = \lim_{t \to 0} a(t_0) \int_{t}^{t_d} \frac{dt}{a(t)} \tag{A.19}$$

For the standard big bang model—a radiation-dominated phase followed by a matter-dominated phase—a straightforward calculation yields the inequality:

$$d_{pph} << d_{vh} \tag{A.20}$$

Points on the surface of last scattering separated by distances greater than d_{pph} are not in causal contact (see Fig. 2.1). If the FLRW models accurately describe the early universe, our observations of the surface of last scattering encompass several regions lying beyond each other's horizons.

The effect of an inflationary stage on the horizon problem can perhaps best be illustrated by comparing the horizon distances in two simple models: a model that is radiation dominated until t_d, and then matter dominated, compared to a model with an inflationary stage lasting from t_i to t_f, sandwiched between two radiation dominated phases (from t_0 to t_i, and t_f to t_d) and followed by a matter dominated phase. For the model without inflation the horizons are given by:[6]

$$d_{pph} = \frac{1}{H_0}\left(\frac{a_d}{a_0}\right)^{1/2} \tag{A.21}$$

$$d_{vh} = \frac{2}{H_0}\left(\frac{a_d}{a_0}\right)^{1/2}\left[\left(\frac{a_0}{a_d}\right)^{1/2} - 1\right], \tag{A.22}$$

where H_0 is the Hubble constant at t_0 and a_0, a_d are the values of the scale factor at t_0, t_d. The inequality above (A.20) follows since $a_0/a_d \approx 1000$. For a simple inflationary model d_{vh} still has the same value, whereas the primeval particle horizon is given by:

$$d_{pph} = \frac{1}{H_0}\left(\frac{a_d}{a_0}\right)^{1/2}\left(1 + 2\left(\frac{a_f}{a_d}\right)\left[\frac{a_f}{a_i} - 1\right]\right), \tag{A.23}$$

where a_f, a_i represent the scale factor at t_i, t_f. Typical inflationary models have a number of "e-foldings" Z, defined as $\frac{a_f}{a_i} = e^Z$, greater than 65. The factor $\frac{a_f}{a_d}$ is typically on the order of 10^{-24}. Thus the overall of effect of incorporating an inflationary stage is to multiply d_{pph} by a number $> 10^4$, reversing the inequality in (A.20).

[6] These equations are derived by integrating the contribution to the horizon distance at each stage, given the behavior of $a(t)$, and then requiring continuity of $a, \dot{a}$ at the boundaries between the stages; see Ellis and Stoeger (1988, Appendix).

A.4 Causal Structure

The study of global causal structure of relativistic spacetimes developed along with efforts to prove the singularity theorems, since these theorems required some degree of "control" over global aspects of spacetime. Below I will briefly discuss various causality conditions, which can roughly be thought of as characterizing the ways in which global causal structure resembles or departs from the "nice" behavior of Minkowski space. Then I will turn to boundary constructions.

The tangent space at any $p \in M$ has the familiar causal structure of Minkowski space. Study of the causal structure of spacetime focuses on the conformal geometry of spacetime since the light cone structure depends on the metric only up to a conformal factor. The causal structure can depart significantly from that of Minkowksi space locally and globally since the tangent spaces may be "tilted" with respect to each other.

The discussion below will focus on the causal sets $J^{\pm}(p)$ and $I^{\pm}(p)$. In order to even define these sets consistently, a spacetime must be *time orientable*, which requires the existence of a continuous, nowhere-vanishing vector field that makes a globally consistent designation of one half of the null cone as the "future lobe" possible. (*Choosing* which half should be so designated is part of the problem of the direction of time, but without this condition no consistent global specification is possible.) A piecewise smooth curve is timelike (resp., null) if its tangent vectors u^a are timelike (null) at each point of the curve; a timelike curve is future (resp., past) directed if its tangent vectors lie in the future (past) lobe of the light cone at every point. A point p *chronologically precedes* q (symbolically, $p << q$), if there is a future-directed timelike curve of non-zero length from p to q. Similarly, p *causally precedes* q ($p < q$), if there is a

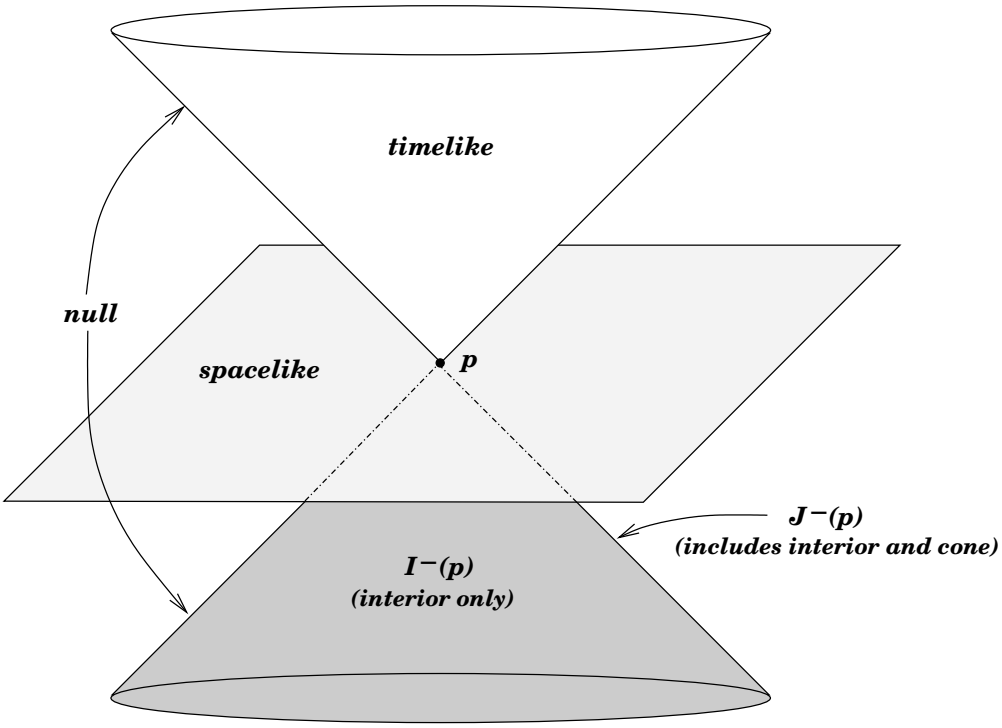

Fig. A.1 This figure illustrates the null cone structure of Minkowski spacetime, which also holds in the tangent space at any point in a general relativistic spacetime. Regions corresponding to the causal sets are identifed (note that these are sets of points in M and not in the tangent space).

future-directed curve with tangent vectors that are timelike or null at every point. Finally, the causal sets are defined in terms of these relations (as illustrated in figure A.1): $I^-(p) = \{q : q << p\},\ I^+(p) = \{q : p << q\}$, the *chronological past* and *future*, and $J^-(p) = \{q : q < p\},\ J^+(p) = \{q : p < q\}$, the *causal past* and *future*. These definitions generalize immediately to spacetime regions: for the region S, $I^+(S) = \cup_{p\in S} I^+(p)$.

The various causality conditions are reflected in the properties of the causal sets $J^\pm(S)$ and $I^\pm(S)$. In the standard manifold topology, $I^\pm(S)$ are always open sets, but $J^\pm(S)$ are generally neither open nor closed. $J^+(S)$ fails to be closed if, for example, $\exists q \in \dot{I}^+(S)$ such that q is not in $J^+(S)$. A spacetime is *causally simple* iff the $J^\pm(S)$ sets are always closed. Turning to the boundaries of these sets, it is intuitively clear that the boundary $\dot{I}^+(S)$ consists of

null surfaces, since some points on either side of a timelike or spacelike surface can be connected with timelike curves (similarly with $+$ replaced by $-$). More precisely, $\dot{I}^+(S)$ is called an *achronal* boundary since no two points in the boundary lie in each others' chronological pasts (see Hawking and Ellis 1973, Prop. 6.3.1 for further properties of this surface). Any given point $q \in \dot{I}^+(S)$ lies either in the closure of S or in a null geodesic generator of the boundary. In Minskowski spacetime, or locally in a general relativistic spacetime, the boundary is "ruled" by null geodesic generators with past endpoints on the closure of S; but in general the generators may encounter "missing points" before reaching S. A very strong causality condition, *global hyperbolicity*, insures that the generators of the boundary in a general spacetime have the same properties as in Minkowski spacetime. A globally hyperbolic spacetime possesses a Cauchy surface, a spacelike surface Σ without edge intersected exactly once by every inextendible null or timelike curve. (See Hawking and Ellis 1973; Wald 1984 for further discussion of the various equivalent definitions of global hyperbolicity.) In a globally hyperbolic spacetime, the EFE admit a well-posed initial value formulation: specifying appropriate initial data on a Cauchy surface Σ determines a unique solution to the field equations (up to diffeomorphism invariance). In addition, the topology of a globally hyperbolic spacetime is $\Sigma \times \mathbb{R}$, and the Cauchy surfaces correspond to the level surfaces of a global time function; however, the foliation is not unique.

A brief discussion of some formal machinery due to Penrose helps to illustrate the intuition behind these constructions: he defines a "TIP," a terminal indecomposable past set, to be the past of a future-directed curve γ (Penrose 1979). Unlike a "PIP," a proper indecomposable past set, which is the chronological past $I^-(p)$ for some $p \in M$, a TIP can be thought of roughly as the past of the "boundary point" reached by γ. Obviously a TIP will not pick out a unique curve; instead an equivalence class of curves approaching the same boundary point correspond to the

same TIP. Finally, a singular TIP corresponds to a curve γ with finite proper length. More abstractly, one defines a set of boundary points on the manifold by identifying equivalence classes of incomplete curves approaching the same boundary points. One also needs to add a topology connecting the boundary points and the interior of the original manifold M.

A.5 No-Hair Theorems

The cosmic no-hair conjecture, originally proposed by Gibbons and Hawking (1977), holds that for fairly general initial conditions a period of Λ dominated exponential expansion "smooths out the wrinkles" of an initially inhomogeneous model. This appendix briefly reviews the first component of a no-hair theorem, i.e., proofs that a transient effective Λ leads to a "locally de Sitter" solution. Several counterexamples, such as closed FLRW models which recollapse before reaching an inflationary stage or the generalized Schwarzschild solution with positive Λ, show that such a theorem cannot hold in full generality—hence the vague phrase "fairly general initial conditions." A locally de Sitter solution resembles the de Sitter solution in that it has nearly flat spatial sections and a constant volume expansion rate $\theta = (\Lambda/3)^{1/2}$ (the expansion rate for de Sitter), but the global structure may differ. The volume expansion rate is defined in terms of the spatial metric h_{ab}: for a smooth congruence of timelike geodesics, generated by the vector field ξ^a (normalized so that $\xi_a\xi^a = -1$), the metric can be decomposed as $g_{ab} = h_{ab} - \xi_a\xi_b$. The volume expansion rate is then given by $\theta = \nabla^b\xi^a h_{ab}$(see, e.g. Wald 1984, §9.2, for a much more detailed discussion).[7]

[7] For further discussion of the definition of a "locally de Sitter" spacetime, see Wald (1983); Goldwirth and Piran (1992).

Wald (1983) proved a no-hair theorem for anisotropic, homogeneous models with a positive Λ satisfying the constraints that ${}^{(3)}R \leq 0$, where ${}^{(3)}R$ is the curvature of the homogeneous spatial sections, and that the strong and dominant energy conditions hold for T_{ab} (excluding Λ).[8] The theorem proceeds by showing that the following equation follows from Raychaudhuri's equation, the field equations, and the energy conditions:

$$-\dot{\theta} \geq \theta^2 - \frac{\Lambda}{3} \geq 0. \tag{A.24}$$

In other words, given these assumptions it follow that the expansion rate is decreasing ($\dot{\theta}$ is negative). If the universe is initially expanding the expansion rate is bounded below by the de Sitter value; integrating eqn. (A.24) shows that the expansion rate rapidly approaches the de Sitter value. Jensen and Stein-Schabes (1987) generalized Wald's result slightly to inhomogeneous models satisfying the similar restriction that ${}^{(3)}R \leq 0$ everywhere. In Wald's case the homogeneous spatial sections can be defined naturally, but in the latter proof the three curvature is defined in terms of the spatial sections according to a synchronous coordinate system.[9] There are several reasons for dissatisfaction with these results: as Goldwirth and Piran (1992) point out, the synchronous coordinate system required by the proof probably does not cover the entire region of interest, and requiring ${}^{(3)}R \leq 0$ everywhere is a very strong restriction (it fails if

[8] The dominant energy condition requires that $-T^a_{\ b}\xi^b$ is future-directed timelike or null for all future-directed and timelike ξ^a; physically, this condition states that a comoving observer measures a non-negative local energy density and the vector representing energy flow is timelike or null. The strong energy condition requires that $(T_{ab} - \frac{1}{2}g_{ab}T)\xi^a\xi^b \geq 0$. It is important to note that these conditions must hold throughout the inflating region for the theorem to hold. Wald does not actually state ${}^{(3)}R \leq 0$ as a condition of the proof; instead he shows that it holds in all the Bianchi models except type IX.

[9] In a synchronous coordinate system the metric can be written as $g_{ab} = h_{ab} - \xi_a\xi_b$; the timelike vector fields ξ_a are orthogonal to the spatial hypersurfaces. Synchronous coordinate systems break down when the curves generated by ξ_a cross, and as a result these coordinate system do not cover large regions of spacetime in the presence of strong gravitational fields.

any region undergoes collapse); in addition, there are a number of "defeating conditions" for a completely general result, such as the presence of primordial magnetic fields.[10]

A.6 Conservation Laws

In general relativity the stress-energy tensor T_{ab} characterizes the sources of the gravitational field. Typical presentations of general relativity duly note that the stress-energy tensor obeys the following conservation law:[11]

$$\nabla^a T_{ab} = 0. \tag{A.25}$$

Here I will briefly review arguments that unlike the familiar conservation laws of classical mechanics, this covariant conservation law does not underwrite the claim that local interactions conserve total energy of the interacting systems.[12] As we will see, the accounting fails due to difficulties incorporating the energy of the gravitational field itself. This difficulty illustrates the global nature of energy conservation in GTR, since definitions of energy and momentum and the corresponding conservation laws can be recovered globally ("at infinity") for spacetimes representing an isolated system.

---

[10] See Rothman and Ellis (1986) for a more detailed criticism of Jensen and Stein-Schabes (1987) and other earlier results; although they criticize the earlier arguments, Rothman and Ellis ultimately support the same conclusions (modulo a few caveats).

[11] This holds as a consequence of Einstein's field equations and the Bianchi identities $\nabla_{[a} R_{bc]d}{}^{e} = 0$, which imply that $\nabla^b G_{ab} = 0$. Einstein's insistence that energy-momentum conservation holds was one of the crucial physical requirements in his search for the gravitational field equations (see Chapter 1).

[12] Energy conservation in general relativity has recently been discussed in the philosophy of physics literature by Hoefer (2000) and by Erik Curiel; I have also benefitted from discussions with Curiel on this subject.

In most familiar cases from classical and special relativistic physics, during interactions energy is transferred continuously from one system to another.[13] These continuity properties depend upon conservation laws, which in this case can be given equivalent differential or integral formulations. Roughly speaking, differential conservation laws guarantee that energy flows continuously through individual spacetime points. On the other hand, integral conservation laws govern the flow of energy into and out of finite regions of spacetime. But the difference between the two is only a matter of mathematics. The equation of motion for a perfect fluid in special relativity subject to no external forces has the form $\partial^a T_{ab} = 0$, which is a differential conservation law (the fluid has no sources or sinks). In special relativity, the energy-momentum density ascribed to a particular point depends upon the state of motion of the observer; technically, one obtains the "energy density," a scalar quantity, by contracting T_{ab} twice with timelike vectors corresponding to the observer's four velocity. A family of inertial observers with four-velocities u^a at rest with respect to each other (so $\partial_a u^b = 0$) will measure an energy-momentum current given by $J_a = -T_{ab}u^b$, and it follows immediately that $\partial^a J_a = 0$. Integrating J_a over a finite spacetime region V with a boundary S and then using Gauss's law, we have the integral conservation law:[14]

$$\int_V \partial^a J_a d^4V = \int_S J_a n^a d\eta = 0. \tag{A.26}$$

Imagine that the surface S consists of two spacelike "caps" Σ_1 and Σ_2 joined by a tube. The integral conservation law guarantees that the difference in energy-momentum-density flux between

[13] The difficulties with giving a local definition of gravitational energy in Newtonian gravitation foreshadow the problems in GTR.

[14] The unit normal to the boundary S is n^a, conventionally defined to be "outward pointing" for spacelike surfaces and "inward pointing" for timelike surfaces, and $d\eta$ is the natural volume element for this surface.

the two caps must be compensated by a flux through the sides of the tube; if the flux through the sides is zero, then the flux through the caps is conserved.

In general relativity, the differential and integral formulations are not two sides of the same coin. The fundamental problem is that we need to extract a suitable quantity from T_{ab} to integrate over spacetime regions, since there is no way to directly integrate tensor fields over a curved manifold. Hitting T_{ab} with a single timelike vector u^b yields J_a as mentioned above; hitting T_{ab} twice yields the energy density, a scalar quantity. In the previous paragraph we took advantage of a family of observers at rest with respect to each other to define a rest frame and the corresponding J_a. In Minkowski spacetime the worldlines of these observers are both geodesics and the integral curves of a timelike Killing vector field, defined as a vector field u^a such that $\nabla_a u_b + \nabla_b u_a = 0$ (this obviously holds for u^a above).[15] In the general relativistic case we would like to choose a vector field u^b such that $\nabla^a(-T_{ab}u^b) = 0$ follows from the covariant conservation law (A.25), and only a timelike Killing vector field will do.[16] But in general solutions to EFE there are no Killing vectors to be had; requiring a timelike Killing vector is equivalent to postulating a time-translation symmetry, in that an observer moving along the orbit of the Killing vector field sees no change in the geometrical properties of spacetime. Thus in general there is no way to turn the covariant conservation law into an integral conservation law. This reflects the difficulty in accounting for the energy which a system gains or loses through interactions with the gravitational field. Textbook presentations of general relativity (such as Weinberg 1972) sometimes introduce a pseudo-tensor t_{ab} representing the stress-energy of the gravitational field

[15] The fact that these worldlines are geodesics is responsible for the zero on the RHS of $\partial^a T_{ab} = 0$; for non-geodesic motion there is a non-zero term representing an external force.

[16] For a Killing vector field u^b, $\nabla^a(T_{ab}u^b) = (\nabla^a T_{ab})u^b + T_{ab}\nabla^a u^b = 0$. The first term vanishes by eqn. (A.25), and the second term vanishes since $T_{ab}\nabla^a u^b = T_{(ab)}\nabla^{[a}u^{b]} = 0$.

itself. It is only a "pseudo" tensor because it does not have the transformation properties of a tensor— as a consequence of the equivalence principle, one can always choose coordinates locally such that t_{ab} vanishes. Gravitational energy differs from the energy-momentum carried by matter fields since it can always be locally "transformed away."

Despite the difficulties with defining the energy and momentum of the gravitational field locally, these quantities and their conservation laws can be formulated *globally* for asymptotically flat spacetimes. Since the 60s powerful techniques have been developed to analyze isolated gravitational systems in terms of the behavior of various quantities "at infinity" in a coordinate-independent manner. To study these quantities one constructs the conformal completion of the given spacetime, which roughly corresponds to "adding in" points at infinity. Although one can construct conformal completions for a wide variety of spacetimes, for asymptotically flat spacetimes the conformal completion resembles the conformal infinity of Minkwoski spacetime.[17] In this context one can define energetic quantities and prove conservation theorems regarding them.

[17] There are a number of conditions required for the construction of the conformal completion. Roughly, one requires the existence of a spacetime $\tilde{M}, \tilde{g_{ab}}$, which consists of the union of the physical spacetime (M, g_{ab}) and the asymptotic region $\mathcal{I}$, and a conformal isometry such that $\tilde{g_{ab}} = \Omega^2 g_{ab}$ on M. A spacetime is asymptotically flat if the conformal completion satisfies a number of additional conditions imposed on the conformal factor Ω and the structure of the asymptotic region, which enforce similarity to conformal infinity in Minkowski spacetime (see, e.g., Wald 1984, pp. 276 f.).

Appendix B

Topics in Quantum Field Theory

B.1 Spontaneous Symmetry Breaking

For a rough idea of the meaning of symmetry breaking in QFT, begin with the phase space Γ for some simple classical system. The dynamics of the system are given by specifying a Hamiltonian function, H, a real-valued function on Γ that gives the energy of the state. Dynamical trajectories are the integral curves of a vector field obtained from H by introducing further geometric structure (namely, the symplectic form). A symmetry of the theory corresponds to the action of some Lie group G on Γ such that the Hamiltonian is invariant; the action of such a group maps dynamical trajectories into dynamical trajectories. Suppose that we are dealing with a system whose quantum description, in terms of an operator algebra $\mathcal{A}$ and a Hilbert space of states, can be constructed directly from Γ (e.g., by canonical quantization). The action of G can be naturally extended to the operator algebra $\mathcal{A}$, resulting in a group of automorphisms on the algebra. SSB reflects a mismatch between symmetries defined at this algebraic level and symmetries on the Hilbert space of states. A symmetry $g \in G$ is spontaneously broken when the corresponding automorphism cannot be defined as a unitary transformation on the Hilbert space (i.e., it cannot be "unitarily implemented"). Since picking out a particular vacuum state determines the Hilbert space representation of $\mathcal{A}$, we can translate this into a statement regarding vacuum expectation values. In particular, a symmetry g is spontaneously broken if the vacuum

expectation value of some operator is non-invariant under the action of the associated automorphism on $\mathcal{A}$.

Rather than filling out this characterization, I will shift gears and focus on the connection between symmetry breaking and conserved currents in order to clarify the claim that a broken symmetry is not unitarily implementable.[1] Nöther's seminal theorems link symmetries of the Lagrangian to conserved currents. Nöther's first theorem establishes that there are n conservation laws of the form $\partial_\mu j^\mu = 0$ for a Lagrangian invariant under an n-parameter "global" Lie group.[2] Thus for a Lagrangian invariant under a one parameter global internal symmetry, we have a single conserved current j^μ. The hermitian "charge" operator is defined as the spatial integral of the time component of the current, $Q = \int j^0 d^3x$. The Fabri-Picasso theorem (Fabri and Picasso 1966) shows that the charge operator can be exponentiated to define a unitary operator, $U(\xi) = e^{i\xi Q}$ only if Q annihilates the vacuum. To prove the theorem, begin with the product of the current and the charge operator in the vacuum state,

$$\langle 0|j_0(\mathbf{x})Q|0\rangle = \langle 0|e^{-iP\cdot x} j_0(0) e^{iP\cdot x} Q|0\rangle, \tag{B.1}$$

where the equality follows from translation invariance. But since the charge operator commutes with P^μ, and P^μ annihilates the vacuum, we have further that

$$= \langle 0|e^{-iP\cdot x} j_0(0) Q e^{iP\cdot x}|0\rangle = \langle 0|j_0(0)Q|0\rangle. \tag{B.2}$$

[1] Here I am following Aitchison (1982); for a more careful discussion see Guralnik et al. (1968).

[2] Martin (2002), §6.2 points out that "global" in this sense should not be confused with global in the context of spacetime theories. Roughly, so-called "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*.

Therefore the norm of Q in the vacuum state,

$$\langle 0|QQ|0\rangle = \int d^3x \langle 0|j_0(x)Q|0\rangle = \int d^3x \langle 0|j_0(0)Q|0\rangle \tag{B.3}$$

diverges unless $Q|0\rangle = 0$. Since the norm of Q is infinite if it does not annihilate the vacuum, it does not exist as a bounded operator acting on the Hilbert space.

B.2 Vacuum Energy

In QFT the vacuum state is defined as the lowest energy state of the assemblage of fields, and here I will briefly review the calculation of vacuum energy for a scalar field. One approach to quantizing the Klein-Gordon field is to treat each Fourier mode of the field as an independent harmonic oscillator. To find the spectrum of the Hamiltonian, one introduces the creation and annihilation operators ($a^\dagger$ and a, respectively) and proceeds by analogy with the treatment of the non-relativistic simple harmonic oscillator. In this case the commutation relation between $a, a^\dagger$ is

$$[a_{\mathbf{p}}, a^\dagger_{\mathbf{p}'}] = (2\pi)^3 \delta^{(3)}(\mathbf{p} - \mathbf{p}'), \tag{B.4}$$

where $\mathbf{p}, \mathbf{p}'$ are three-space momentum vectors. Writing the Hamiltonian of the Klein-Gordon field in terms of these operators yields:[3]

$$H = \int \frac{d^3p}{(2\pi)^3} \omega_{\mathbf{p}} (a^\dagger_{\mathbf{p}} a_{\mathbf{p}} + \frac{1}{2}[a_{\mathbf{p}}, a^\dagger_{\mathbf{p}}]) \quad \text{where} \quad \omega_{\mathbf{p}} = \sqrt{|\mathbf{p}|^2 + m^2}. \tag{B.5}$$

[3] Starting with the usual expression for the Klein-Gordon Hamiltonian, namely $H = \int d^3x [\frac{1}{2}\pi^2 + \frac{1}{2}(\nabla\phi)^2 + \frac{1}{2}m^2\phi^2]$, eqn. (B.5) follows by rewriting ϕ, π in terms of $a, a^\dagger$. See, e.g., Peskin and Schroeder (1995), Chapter 2.

The commutator term is analogous to the non-zero ground state energy of the non-relativistic oscillator, but in this case it contributes an energy of $\frac{\omega_{\mathrm{P}}}{2}$ per mode in the sum over all modes. The delta function can be handled by "box regularization" (i.e. imposing boundary conditions on the field as if it were in a finite-volume box), but even so the second term is still divergent. Imposing a finite frequency cut-off ω_{max} renders the integral convergent, and restoring the $\hbar$'s and c's, we have the following estimate for the vacuum energy density of a scalar field:[4]

$$\langle \rho_{vac} \rangle = \frac{\hbar}{8\pi^2 c^3} \omega_{max}^4 . \tag{B.6}$$

See, e.g., Miloni (1994); Rugh and Zinkernagel (2001) for more detailed discussions and derivations of these results. A similar calculation leads to the same result for the vacuum energy of the free electromagnetic field in QED. The cut-off ω_{max} is usually interpreted as the upper limit of the applicability of the theory; choosing roughly $100\ GeV$ for QED, this yields an energy density of $\langle \rho_{vac} \rangle \approx 10^{46}\ erg/cm^3$.

One can ignore this incredible energy density almost as soon as one has calculated it: S-matrix calculations are entirely unaffected by the presence of vacuum energy. The calculated energy density has all the characteristics of an unrenormalized quantity: it is cut-off dependent, and diverges in the limit as $\omega_{max} \to \infty$. As long as gravitational effects are neglected, the vacuum energy can be "rescaled" by simply dropping the zero point contribution to the Hamiltonian. For some free field theories a general prescription known as normal ordering (a.k.a. Wick ordering) can be used to accomplish this: simply move all of the a's to the right of the $a^\dagger$'s. In a normal ordered product of fields there are no commutator terms (the second term in eqn.(B.5)),

[4] The notation $\langle \rho \rangle$ is shorthand for $\langle 0|\rho|0\rangle$ where $|0\rangle$ is the vacuum state.

so normal ordering eliminates the vacuum energy by getting rid of the zero point energy term. Unfortunately normal ordering does not carry over to interacting field theories or gauge field theories (one cannot simply rearrange the operators while respecting the gauge symmetry). To my knowledge there is no more general formal prescription for eliminating zero point energy in interacting gauge theories. The position of most physicists on this subject seems to be the following: the formal consistency of field theory requires the introduction of the vacuum field with zero point energy, and (as discussed in the text) various effects, such as the Casimir effect, are observable consequences of the zero point energy.[5]

[5]Regarding the second point Miloni (1994) is much more careful than most physicists: unlike Weinberg (1989), who argues that the Casimir effect unambiguously establishes the reality of zero point energy, Miloni acknowledges that the Casimir effect can be treated as a vacuum effect or in terms of source fi elds. He calls the difference between the two approaches a "matter of taste," and argues instead that a consistent formulation of QFT requires the introduction of vacuum fi elds with zero point energies.

Bibliography

Agazzi, E. and Cordero, A., editors (1991). *Philosophy and the Origin and Evolution of the Universe*, volume 217 of *Studies in epistemology, logic, methodology, and philosophy of science*, Boston. Kluwer.

Aitchison, I. J. R. (1982). *An informal introduction to gauge field theories*. Cambridge University Press, Cambridge.

Albert, D. Z. (2000). *Time and Chance*. Harvard University Press, Cambridge.

Albrecht, A. (1997). How to falsify scenarios with primordial fluctuations from inflation. In Turok, N., editor, *Critical Dialogues in Cosmology*, pages 265–273, Singapore. World Scientific.

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

Alpher, R. A., Bethe, H., and Gamow, G. (1948). The origin of chemical elements. *Physical Review*, 73:803–804.

Alpher, R. A., Follin, J. W., and Herman, R. C. (1953). Physical Conditions in the Initial Stages of the Expanding Universe. *Physical Review*, 92:1347–1361.

Alpher, R. A. and Herman, R. C. (1949). Remarks on the evolution of the expanding universe. *Physical Review*, 75:1089–1099.

Altshuler, B. L., Bolotovsky, B. M., Dremin, I. M., Fainberg, V. Y., and Keldysh, L. V., editors (1991). *Andrei Sakaharov: Facets of a life*. Editions Frontières, Gif-sur-Yvette Cedex.

Anderson, J. (1967). *Principles of Relativity Physics*. Academic Press, New York.

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

Anderson, P. W. (1963). Plasmons, gauge invariance, and mass. *Physical Review*, 130(1):439–442.

Arageorgis, A. (1995). *Fields, Particles, and Curvature: Foundations and Philosophical Aspects of Quantum Field Theory in Curved Spacetime*. PhD thesis, University of Pittsburgh.

Arntzenius, F. (1990). Physics and common causes. *Synthese*, 82:77–96.

Arntzenius, F. (1993). The common cause principle. In Hull, D. and Okruhlik, K., editors, *PSA 1992*, pages 227–237. Philosophy of Science Association, East Lansing, MI.

Arntzenius, F. (1999). Reichenbach's common cause principle. Entry in Stanford Encyclopedia of philosophy, http://plato.stanford.edu/entries/physics-Rpcc.

Baez, J. and Muniain, J. (1994). *Gauge Fields, Knots and Gravity*. World Scientific, Singapore.

Bahcall, N. A., Ostriker, J. P., Perlmutter, S., and Steinhardt, P. J. (1999). The cosmic triangle: Revealing the state of the universe. *Science*, 284:1481–1488.

Bain, J. (1998). *Representations of Spacetime: Formalism and Ontological Commitment*. PhD thesis, University of Pittsburgh.

Bardeen, J., Cooper, L. N., and Schrieffer, J. R. (1957). Theory of superconductivity. *Physical Review*, 108:1175–1204.

Bardeen, J. M. (1980). Gauge invariant cosmological perturbations. *Phys. Rev.*, D22:1882–1905.

Bardeen, J. M., Steinhardt, P. J., and Turner, M. S. (1983). Spontaneous creation of almost scale - free density perturbations in an inflationary universe. *Physical Review D*, 28:679.

Barrow, J. (1978). Quiescient cosmology. *Nature*, 272:211–216.

Barrow, J. (2002). Interview with John Barrow conducted by Chris Smeenk. 36 pp. manuscript, to be deposited in the Oral History Archives at the American Institute of Physics.

Barrow, J. and Liddle, A. (1997). Can inflation be falsified? gr-qc/9705048.

Barrow, J. and Matzner, R. (1977). The homogeneity and isotropy of the universe. *Monthly Notices of the Royal Astronomical Society*, 181:719–727.

Barrow, J. D. (1980). Galaxy formation - The first million years. *Royal Society of London Philosophical Transactions Series A*, 296:273–288.

Barrow, J. D. (1995). Why the universe is not anisotropic. *Physical Review D*, 51:3113–3116.

Barrow, J. D. and Tipler, F. J. (1986). *The anthropic cosmological principle*. Oxford University Press, Oxford.

Barrow, J. D. and Turner, M. S. (1981). Inflation in the universe. *Nature*, 292:35–38.

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

Bassett, B. A., Gordon, C., Maartens, R., and Kaiser, D. I. (2000). Restoring the sting to metric preheating. *Physical Review D*, 61:061302.

Bassett, B. A., Tamburini, F., Kaiser, D. I., and Maartens, R. (1999). Metric preheating and limitations of linearized gravity. II. *Nuclear Physics*, B561:188–240.

Beatty, J. (1995). The evolutionary contingency thesis. In Wolters, G. and Lennox, J. G., editors, *Concepts, Theories, and Rationality in the Biological Sciences*. University of Pittsburgh Press, Pittsburgh.

Bekenstein, J. 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 Longair, M., editor, *Confrontation of Cosmological Theories with Observational Data*, number 63 in IAU Symposium, pages 261–275, Dodrecht. D. Reidel.

Belinsky, V. A., Ishihara, H., Khalatnikov, I. M., and Sato, H. (1988). On the degree of generality of inflation in Friedmann cosmological models with a massive scalar field. *Progress in Theoretical Physics*, 79:676–684.

Belinsky, V. A. and Khalatnikov, I. M. (1987). On the degree of generality of inflationary solutions in cosmological models with a scalar field. *Soviet Physics JETP*, 66:441–449.

Bennett, C. L. et al. (2003). First year Wilkinson microwave anisotropy probe (WMAP) observations: Preliminary maps and basic results. astro-ph/0302207.

Bernstein, J. and Feinberg, G., editors (1986). *Cosmological Constants: Papers in Modern Cosmology*. Columbia University Press, New York.

Bertola, F. and Curi, U., editors (1993). *The Anthropic Principle: Proceedings of the second Venice conference on cosmology and philosophy*. Cambridge University Press.

Birrell, N. C. and Davies, P. C. W. (1982). *Quantum fields in curved space*. Cambridge University Press, Cambridge.

Blau, S. K. and Guth, A. (1987). Inflationary cosmology. In Hawking, S. W. and Israel, W., editors, *300 years of gravitation*, pages 524–603. Cambridge University Press, Cambridge.

Blewitt, G. et al. (1985). Experimental limits on the free proton lifetime for two and three-body decay modes. *Physical Review Letters*, 55:2114–2117.

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

Bonnor, W. B. (1956). The formation of the nebulae. *Zeitschrift für Astrophysik*, 39:143–59.

Bonnor, W. B. (1957). Jeans' formula for gravitational instability. *Monthly Notices of the Royal Astronomical Society*, 117:104–117.

Born, M. (1956). Physics and relativity. In Mercier, A. and Kervaire, M., editors, *Fünzig Jahre Relativitätstheorie. Bern, 11-16 Juli 1955. Verhandlungen*, pages 244–260. Birkhäuser, Basel. *Helvetica Physica Acta* **4** (Supplement).

Börner, G. (1993). *The early universe: facts and fiction*. Springer-Verlag, New York, third edition.

Bostrom, N. (2002). *Anthropic Bias: Observation Selection Effects in Science and Philosophy*. Routledge, New York.

Brandenberger, R. H. and Martin, J. (2001). The robustness of inflation to changes in super-Planck- scale physics. *Modern Physics Letters*, A16:999–1006.

Brawer, R. (1996). Inflationary cosmology and the horizon and flatness problems: The mutual constitution of explanation and questions. Master's thesis, MIT, Physics.

Brout, R., Englert, F., and Gunzig, E. (1978). The creation of the universe as a quantum phenomenon. *Annals of Physics*, 115:78–106.

Brout, R., Englert, F., and Gunzig, E. (1979). The causal universe. *General Relativity and Gravitation*, 10:1–6.

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

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

Brush, S. (1989). Prediction and theory evaluation: the case of light bending. *Science*, 246:1124–29.

Brush, S. (1993). Prediction and theory evaluation: Cosmic microwaves and the revival of the big bang. *Perspectives on Science*, 1:565–602.

Bucher, M., Goldhaber, A. S., and Turok, N. (1995). An open universe from inflation. *Phys. Rev.*, D52:3314–3337.

Buckingham, M. J. (1957). A note on the energy gap model of superconductivity. *Nuovo Cimento*, 5:1763–65.

Burbidge, E. M., Burbidge, G. R., Fowler, W. A., and Hoyle, F. (1957). Synthesis of the elements in stars. *Reviews of Modern Physics*, 29:547–650.

Burbidge, E. M., Burbidge, G. R., and Hoyle, F. (1963). Condensations in the intergalactic medium. *Astrophysical Journal*, 138:873–888.

Callender, C. (2001). Taking thermodynamics too seriously. *Studies in the History and Philosophy of Modern Physics*, 32:539–553.

Callender, C. and Huggett, N., editors (2001). *Philosophy Meets Physics at the Planck Scale*. Cambridge University Press.

Carter, B. (1974). The anthropic principle and large number coincidences. In Longair, M., editor, *Confrontation of cosmological theories with observation*. D. Reidel, Dodrecht.

Cartwright, N. (1999). *The Dappled World: A Study of the Boundaries of Science*. Cambridge University Press, Cambridge.

Cheng, T.-P. and Li, L.-F. (1984). *Gauge theory of elementary particles*. Oxford University Press, Oxford.

Cho, H. T. and Kantowski, R. (1994). Measure on a subspace of FRW solutions and the flatness problem of standard cosmology. *Physical Review D*, 50(10):6144–6149.

Churchland, P. M. and Hooker, C. A., editors (1985). *Images of Science*. University of Chicago Press, Chicago.

Clarke, C. J. S. (1993). *The Analysis of Space-Time Singularities*. Number 1 in Cambridge Lecture Notes in Physics. Cambridge University Press, Cambridge.

Clarkson, C. and Barrett, R. (1999). Does the isotropy of the CMB imply a homogeneous universe? Some generalised EGS theorems. *Classical and Quantum Gravity*, 16:3781–3794.

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

Coleman, S. R. and Weinberg, E. (1973). Radiative corrections as the origin of spontaneous symmetry breaking. *Physical Review D*, 7:1888–1910.

Coles, P. and Ellis, G. F. R. (1997). *Is the universe open or closed?* Cambridge University Press, Cambridge.

Collins, C. B. and Hawking, S. W. (1973). Why is the universe isotropic? *Astrophysical Journal*, 180:317–334.

Collins, C. B. and Stewart, J. M. (1971). Qualitative cosmology. *Monthly Notices of the Royal Astronomical Society*, 153:419–434.

Coule, D. H. (1995). Canonical measure and the flatness of a FRW universe. *Classical and Quantum Gravity*, 12:445–469.

Criss, T., Matzner, R., Ryan, M., and Shepley, L. (1975). Modern theoretical and observational cosmology. In Shaviv and Rosen, editors, *General Relativity and Gravitation*, number 7, pages 33–108, New York. John Wiley & Sons.

Curiel, E. N. (1999). The analysis of singular spacetimes. *Philosophy of Science*, 66:S119–S145.

de Sitter, W. (1911). On the bearing of the principle of relativity on gravitational astronomy. *Monthly Notices of the Royal Astronomical Society*, 71:388–415.

de Sitter, W. (1916a). On Einstein's theory of gravitation and its astronomical consequences, I. *Monthly Notices of the Royal Astronomical Society*, 76:699–738.

de Sitter, W. (1916b). On Einstein's theory of gravitation and its astronomical consequences, II. *Monthly Notices of the Royal Astronomical Society*, 77:155–184.

de Sitter, W. (1917a). On Einstein's theory of gravitation and its astronomical consequences, III. *Monthly Notices of the Royal Astronomical Society*, 78:3–28.

de Sitter, W. (1917b). On the curvature of space. *Koninklijke Akademie von Wetenschappen. Section of Sciences. Proceedings.*, 20:229–243.

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

Descartes, R. (1985). *The philosophical writings of Descartes*, volume I. Cambridge University Press, Cambridge.

Dicke, R. (1961). Dirac's cosmology and Mach's principle. *Nature*, 192:440–441.

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

Dicke, R. H. (1969). *Gravitation and the Universe: Jayne Lectures for 1969*. American Philosophical Society, Philadelphia.

Dirac, P. A. M. (1937). The cosmological constants. *Nature*, 139:323.

Dolan, L. and Jackiw, R. (1974). Symmetry behavior at finite temperature. *Physical Review D*, 9(12):3320–3341.

Doroshkevich, A. G., Zel'dovich, Y. B., and Novikov, I. G. (1968). Neutrino viscosity and isotropization? *Soviet Physics JETP*, 48:408–412.

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

Earman, J. (1986). *A Primer on Determinism*. MIT Press, Cambridge.

Earman, J. (1987a). Locality, nonlocality, and action at a distance: A skeptical review of some philosophical dogmas. In Achinstein, P. and Kargon, R., editors, *Kelvin's Baltimore Lectures and modern theoretical physics: historical and philosophical perspectives*, pages 449–490. MIT Press, Cambridge.

Earman, J. (1987b). The SAP also rises: A critical examination of the anthropic principle. *American Philosophical Quarterly*, 24:307–317.

Earman, J. (1992). *Bayes or Bust?* MIT Press, Cambridge.

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

Earman, J. (1999). The Penrose-Hawking singularity theorems: History and implications. In Renn, J., Sauer, T., and Gönner, H., editors, *The Expanding Worlds of General Relativity*, number 7 in Einstein Studies, pages 235–270. Birkhäuser, Boston.

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

Earman, J. (2002). Thoroughly modern McTaggart: Or what McTaggart would have said if he had learned the general theory of relativity. *Philosopher's Imprint*, 2.

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

Earman, J. and Roberts, J. (1999). *Ceteris Paribus*, there is no problem of provisos. *Synthese*, 118:439–478.

Eddington, A. S. (1930). On the instability of Einstein's spherical world. *Monthly Notices of the Royal Astronomical Society*, 90:668–678.

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

Ehlers, J., Geren, P., and Sachs, R. K. (1968). Isotropic solutions of the Einstein–Liouville equations. *Journal of Mathematical Physics*, 9:1344–1349.

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

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

Einstein, A. (1907). Über das Relativitätsprinzip und die aus demselben gezogenen Folgerungun. *Jahrbuch der Radioaktivität und Elektronik*, 4:411–462.

Einstein, A. (1916). Die Grundlagen der allgemeinen Relativitätstheorie. *Annalen der Physik*, 49:769–822. Reprinted in translation in Lorentz et al. 1923.

Einstein, A. (1917). Kosmologische Betrachtungen zur allgemeinen Relativitätstheorie. *Preussische Akademie der Wissenschaften (Berlin). Sitzungsberichte*, pages 142–152. Reprinted in translation in Lorentz et al. 1923.

Einstein, A. (1918a). Kritischen zu einer von Herrn de Sitter gegebenen Lösung der Gravitationgleichungen. *Sitzungsberichte der Preussischen Akademie der Wissenschaften*, pages 270–272.

Einstein, A. (1918b). Prinzipielles zur allgemeinen Relativitätstheorie. *Annalen der Physik*, 55:241–244.

Einstein, A. (1922). Bemerkung zu der Arbeit von A. Friedmann, 'Über die Krümmung des Raumes'. *Zeitschrift für Physik*, 11:326.

Eisenstaedt, J. (1986). The low water mark of general relativity, 1925-1955. In Stachel, J. and Howard, D., editors, *Einstein and the History of General Relativity*, volume 1 of *Einstein Studies*, pages 277–292. Birkhäuser, Boston.

Elby, A. (1992). Should we explain the EPR correlations causally? *Philosophy of Science*, 59:16–25.

Ellis, G. F. R. (1980). Limits to verification in cosmology. In *9th Texas Symposium on Relativistic Astrophysics*, volume 336 of *Annals of the New York Academy of Sciences*, pages 130–160.

Ellis, G. F. R. (1989). The expanding universe: A history of cosmology from 1917 to 1960. In Howard, D. and Stachel, J., editors, *Einstein and the History of General Relativity*, volume 1 of *Einstein Studies*, pages 367–431. Birkhäuser, Boston.

Ellis, G. F. R. (1990). Innovation, resistance, and change: The transition to the expanding universe. In Bertotti, B., Balbinot, R., Bergia, S., and Messina, A., editors, *Modern Cosmology in Retrospect*, pages 98–113. Cambridge University Press, Cambridge.

Ellis, G. F. R. (1991). Standard and inflationary cosmologies. In Mann, R. and Masson, P., editors, *Gravitation. A Banff Summer institute*, pages 3–53, Singapore. World Scientific.

Ellis, G. F. R., Maartens, R., and Nel, S. D. (1978). The expansion of the universe. *Monthly Notices of the Royal Astronomical Society*, 184:439–465.

Ellis, G. F. R. and Madsen, M. S. (1991). Exact scalar field cosmologies. *Classical and Quantum Gravity*, 8:667–676.

Ellis, G. F. R. and Rothman (1993). Lost horizons. *American Journal of Physics*, 61(10):883–893.

Ellis, G. F. R. and Sciama, D. W. (1972). Global and non-global problems in cosmology. In O'Raifeartaigh, L., editor, *General Relativity: Papers in Honour of J. L. Synge*, pages 35–59, Oxford. Clarendon Press.

Ellis, G. F. R. and Stoeger, W. (1988). Horizons in inflationary universes. *Classical and Quantum Gravity*, 5:207–220.

Ellis, G. F. R. and van Elst, H. (1999). Cosmological models: Cargése lectures 1998. In Lachiéze-Rey, M., editor, *Theoretical and Observational Cosmology*, pages 1–116. Kluwer, Dordrecht. gr-qc/9812046.

Ellis, J., Gaillard, M. K., and Nanopoulos, D. V. (1980). The smoothness of the universe. *Physics Letters B*, 90:253–257.

Ellis, J., Nanopoulos, D. V., Olive, K. A., and Tamvakis, K. (1982). Cosmological inflation cries out for supersymmetry. *Physics Letters B*, 118:335–338.

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

Evrard, G. and Coles, P. (1995). Getting the measure of the flatness problem. *Classical and Quantum Gravity*, 12:L93–L97.

Fabri, E. and Picasso, L. E. (1966). Quantum field theory and approximate symmetries. *Physical Review Letters*, 16:409–410.

Farhi, E. and Jackiw, R., editors (1982). *Dynamical Gauge Symmetry Breaking: A collection of reprints*. World Scientific, Singapore.

Friedman, M. (1974). Explanation and scientific understanding. *Journal of Philosophy*, 71:5–19.

Friedman, M. (1983). *Foundations of Space-time theories*. Princeton University Press, Princeton.

Gale, G. (2002). Cosmology: Methodological debates in the 1930s and 1940s. Published online at plato.stanford.edu/entries/cosmology-30s/.

Gamow, G. (1948a). The evolution of the universe. *Nature*, 162:680–682.

Gamow, G. (1948b). The origin of elements and separation of galaxies. *Physical Review*, 74:505–506. Reprinted in Bernstein and Feinberg (1986).

Gamow, G. (1952). The role of turbulence in the evolution of the universe. *Physical Review*, 86:251.

Gamow, G. (1954). On the formation of protogalaxies in the turbulent primordial gas. *Proceedings of the National Academy of Science*, 40:480–84.

Geroch, R. (1967). Topology in general relativity. *Journal of Mathematical Physics*, 8:782–786.

Geroch, R. (1968). Spinor structure of space-times in general relativity I. *Journal of Mathematical Physics*, 9:1739–1744.

Geroch, R. (1985). *Mathematical Physics*. University of Chicago Press, Chicago.

Geroch, R., Can-bin, L., and Wald, R. (1982). Singular boundaries of space-times. *Journal of Mathematical Physics*, 23:432–35.

Gibbons, G. and Hawking, S. (1993). *Euclidean Quantum Gravity*. World Scientific, Singapore.

Gibbons, G. W. and Hawking, S. W. (1977). Cosmological event horizons, thermodynamics, and particle creation. *Physical Review D*, 15(10):2738–51.

Gibbons, G. W., Hawking, S. W., and Stewart, J. M. (1987). A natural measure on the set of all universes. *Nuclear Physics*, B281:736–751.

Gilbert, W. (1964). Broken symmetries and massless particles. *Physical Review Letters*, 12:713–14.

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

Gliner, E. B. (1970). The vacuum-like state of a medium and Friedmann cosmology. *Soviet Physics Doklady*, pages 559–561.

Gliner, E. B. and Dymnikova, I. G. (1975). A nonsingular Friedmann cosmology. *Soviet Astronomy Letters*, 1:93–94.

Glymour, C. (1972). Topology, cosmology, and convention. *Synthese*, 24:195–218.

Glymour, C. (1977). Indistinguishable space-times and the fundamental group. In Earman, J., Glymour, C., and Statchel, J., editors, *Foundations of Space-Time Theories*, volume VIII of *Minnesota Studies in the Philosophy of Science*, pages 50–60. University of Minnesota Press.

Glymour, C. (1980a). Explanation, tests, unity and necessity. *Noûs*, 14:31–50.

Glymour, C. (1980b). *Theory and Evidence*. Princeton University Press, Princeton.

Glymour, C. (1985). Explanation and realism. In Churchland and Hooker (1985).

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

Goldstone, J., Salam, A., and Weinberg, S. (1962). Broken symmetries. *Physical Review*, 127(3).

Goldwirth, D. S. and Piran, T. (1992). Initial conditions for inflation. *Physics Reports*, 214:223–292.

Goode, S. W., Coley, A. A., and Wainwright, J. (1992). The isotropic singularity in cosmology. *Classical and Quantum Gravity*, 9:445–455.

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

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

Guralnik, G. S., Hagen, C. R., and Kibble, T. W. B. (1968). Broken symmetries and the Goldstone theorem. In Cool, R. L. and Marshak, R. E., editors, *Advances in Particle Physics*, volume 2, pages 567–708.

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

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

Guth, A. (1997a). *The inflationary universe*. Addison-Wesley, Reading, MA.

Guth, A. (1997b). Thesis: Inflation provides a compelling explanation for why the universe is so large, so flat, and so old, as well as a (almost) predictive theory of density perturbations. In Turok, N., editor, *Critical Dialogues in Cosmology*, pages 233–248, New Jersey. World Scientific.

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

Guth, A. H. and Pi, S. Y. (1982). Fluctuations in the new inflationary universe. *Physical Review Letters*, 49:1110–1113.

Guth, A. H. and Weinberg, E. J. (1981). Cosmological consequences of a first order phase transition in the SU(5) grand unified model. *Physical Review D*, 23:876–885.

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

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

Harper, W. (1997). Isaac newton on empirical success and scientific method. In Earman, J. and Norton, J., editors, *Cosmos of Science: Essays of Exploration*, pages 55–86. University of Pittsburgh Press, Pittsburgh.

Harré, R. (1962). Philosophical aspects of cosmology. *British Journal for the Philosophy of Science*, 13:104–119.

Harré, R. (1986). *Varieties of Realism*. Basil Blackwell, Oxford.

Harrison, E. R. (1968). On the origin of galaxies. *Monthly Notices of the Royal Astronomical Society*, 141:397–407.

Harrison, E. R. (1970). Fluctuations at the threshold of classical cosmology. *Phys. Rev.*, D1:2726–2730.

Hartle, J. B. (1986). Initial conditions. In Kolb, E. W., Turner, M. S., Lindley, D., Olive, K., and Seckel, D., editors, *Inner Space / Outer Space*, pages 467–478, Chicago. University of Chicago Press.

Hartle, J. B. and Hawking, S. W. (1983). Wave function of the universe. *Physical Review*, D28:2960–2975.

Hawking, S. (1974). The anisotropy of the universe at large times. In Longair, M. S., editor, *Confrontation of cosmological theories with observational data*, pages 317–334. D. Reidel, Dordrecht.

Hawking, S. (1988). *A Brief History of Time*. Bantam Doubleday, New York.

Hawking, S. and Penrose, R. (1996). *The nature of space and time*. Isaac Newton Institute series of lectures. Princeton University Press, Princeton.

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

Hawking, S. W. (1982). The development of irregularities in a single bubble inflationary universe. *Physics Letters B*, 115:295–297.

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

Hawking, S. W. and Ellis, G. F. R. (1973). *The Large Scale Structure of Space-Time*. Cambridge University Press, Cambridge.

Hawking, S. W., Gibbons, G. W., and Siklos, S. T. C., editors (1983). *The very early universe*. Cambridge University Press, Cambridge.

Hawking, S. W. and Moss, I. G. (1982). Supercooled phase transitions in the very early universe. *Physics Letters B*, 110:35–38.

Hawking, S. W. and Page, D. (1988). How probable is inflation? *Nuclear Physics*, B298:789–809.

Hawking, S. W. and Tayler, R. (1966). Helium production in an anisotropic big-bang cosmology. *Nature*, 209:1278–1282.

Hempel, C. G. (1988). Provisoes: A problem concerning the inferential function of scientific theories. *Erkenntnis*, 28:147–164.

Henneaux, M. (1983). The Gibbs entropy production in general relativity. *Nuovo Cimento Lettere*, 38:609–614.

Higgs, P. (1997). *Spontaneous breaking of symmetry and gauge theories*, chapter Panel Session: Spontaneous Breaking of Symmetry. In Hoddeson et al. (1997).

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

Hoddeson, L., Brown, L., Riordan, M., and Dresden, M., editors (1997). *Rise of the Standard Model: Particle physics in the 1960s and 1970s*. Cambridge University Press, Cambridge.

Hoefer, C. (2000). Energy conservation in GTR. *Studies in the History and Philosophy of Modern Physics*, 31:171–186.

Hollands, S. and Wald, R. (2002a). An alternative to inflation. gr-qc/0205058.

Hollands, S. and Wald, R. (2002b). Comment on inflation and alternative cosmology. hep-th/0210001.

Howson, C. and Urbach, P. (1989). *Scientific Reasoning: the Bayesian approach*. Open Court, La Salle, Illinois.

Hu, B. L. (1986). Notes on cosmological phase transitions. In Kolb, E. W., Turner, M., Schramm, D., and Lindley, D., editors, *Inner Space/Outer Space*, pages 479–483, Chicago. University of Chicago Press.

Hu, B. L., Ryan, M. P., and Vishveshwara, C. V. (1993). *Directions in General Relativity*, volume 1. Cambridge University Press, Cambridge.

Isenberg, J. and Marsden, J. (1982). A slice theorem for the space of solutions of Einstein equations. *Physics Reports*, 89:180–222.

Isham, C. and Butterfield, J. (2000). On the emergence of time in quantum gravity. In Butterfield, J., editor, *The Arguments of Time*, pages 111–168. Oxford University Press, Oxford.

Janssen, M. (1997). Reconsidering a scientific revolution: the case of Einstein versus Lorentz. Unpublished manuscript.

Janssen, M. (2002). COI stories: Explanation and evidence in the history of science. *Perspectives on Science*, pages 457–522.

Jensen, L. G. and Stein-Schabes, J. A. (1987). Is inflation natural? *Physical Review D*, 35:1146–1150.

Kaiser, D. (1998). A Ψ is just a Ψ? Pedagogy, practice and the reconstitution of general relativity, 1942-1975. *Studies in the History and Philosophy of Modern Physics*, 29:321–338.

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

Kevles, D. J. (1977). *The physicists: the history of a scientific community in modern America*. Knopf, New York.

Kibble, T. W. B. (1966). Symmetry breaking in non-Abelian gauge theories. *Physical Review*, 155(5):1554–1561.

Kibble, T. W. B. (1976). Topology of cosmic domains and strings. *Journal of Physics*, A9:1387–97. Reprinted in Bernstein and Feinberg (1986).

Kibble, T. W. B. (1980). Some implications of a cosmological phase transition. *Physics Reports*, 67:183.

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

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

Kitcher, P. (1989). Explanatory unification and the causal structure of the world. In Kitcher, P. and Salmon, W., editors, *Scientific Explanation*, volume XIII of *Minnesota Studies in the Philosophy of Science*, pages 410–505. University of Minnesota Press, Minneapolis.

Kitcher, P. (1993). *The Advancement of Science*. Oxford University Press, Oxford.

Klein, A. and Lee, B. W. (1964). Does spontaneous breakdown of symmetry imply zero-mass particles? *Physical Review Letters*, 12:266–68.

Koertge, N. (1992). Explanation and its problems. *British Journal for the Philosophy of Science*, 43:85–98.

Kolb, E. W. (1994). Particle physics and cosmology. In Peach, K. J. and Vick, L. L. J., editors, *High Energy Phenomenology*, pages 361–416. St. Andrews Press, Fife.

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

Kolb, E. W. and Wolfram, S. (1980). Spontaneous symmetry breaking and the expansion rate of the early universe. *Astrophysical Journal*, 239:428–432.

Kragh, H. (1996). *Cosmology and Controversy*. Princeton University Press, Princeton.

Kristian, J. and Sachs, R. (1966). Observations in cosmology. *Astrophysical Journal*, 143:379–399.

Kubrin, D. (1967). Newton and the cyclical cosmos: Providence and the mechanical philosophy. *Journal of the History of Ideas*, 28:325–46.

Kuhn, T. (1970). *Structure of Scientific Revolutions*. University of Chicago Press, Chicago, 2nd edition.

Kuhn, T. (1977). *The Essential Tension*. University of Chicago Press, Chicago.

Kukla, A. (1995). Scientific realism and theoretical unification. *Analysis*, 55(4):230–238.

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), pages 213–230. Proceedings of the second Seminar on Quantum Gravity; Moscow, October 13-15, 1981.

Laudan, L. and Leplin, J. (1991). Empirical equivalence and underdetermination. *Journal of Philosophy*, 88:449–472.

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

Leplin, J. (1997). *A novel defense of scientific realism*. Oxford University Press, Oxford.

Leslie, J. (1989). *Universes*. Routledge, New York.

Lewontin, R. (1974). *The Genetic Basis of Evolutionary Change*. Columbia University Press, New York.

Liddle, A. and Lyth, D. (2000). *Cosmological Inflation and Large-Scale Structure*. Cambridge University Press, Cambridge.

Lifshitz, Y. M. (1946). On the gravitational stability of the expanding universe. *Journal of Physics USSR*, 10:116–129.

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

Lightman, A. and Press, W. (1989). Surfaces of constant redshift in an inflationary universe. *Astrophysical Journal*, 337:598–600.

Linde, A. (1974). Is the Lee constant a cosmological constant? *Soviet Physics JETP*, 19(5):183–184.

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

Linde, A. (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.

Linde, A. (1990). *Particle physics and inflationary cosmology*. Harwood Academic Publishers, Amsterdam.

Linde, A. (2002). Interview with Andrei Linde conducted by Chris Smeenk. 89 pp. manuscript, to be deposited in the Oral History Archives at the American Institute of Physics.

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

Lorentz, H. A., Einstein, A., Minkowski, H., and Weyl, H. (1952). *The Principle of Relativity*. Dover. First published by Methuen and Company in 1923. Translations by W. Perrett and G.B. Jeffrey.

Lukash, V. N. (1980). Production of phonons in an isotropic universe. *Soviet Physics JETP*, 52:807–814.

Lyth, D. and Riotto, A. (1999). Particle physics models of inflation and the cosmological density perturbation. *Physics Reports*, 314:1–146.

MacCallum, M. A. H. (1971). Problems of the mixmaster universe. *Nature*, 230:112–115.

MacCallum, M. A. H. (1979). Anisotropic and inhomogeneous relativisitic cosmologies. In Hawking, S. and Israel, W., editors, *General Relativity: An Einstein centenary survey*, pages 533–580. Cambridge University Press, Cambridge.

Madsen, M. S., Ellis, G. F. R., Mimoso, J. P., and Butcher, J. A. (1992). Evolution of the density parameter in inflationary cosmology reexamined. *Physical Review D*, 46:1399–1415.

Malament, D. (1977). Obervationally indistinguishable space-times. In Earman, J., Glymour, C., and Statchel, J., editors, *Foundations of Space-Time Theories*, volume VIII of *Minnesota Studies in the Philosophy of Science*, pages 61–80. University of Minnesota Press.

Mannheim, P. D. (2000). Attractive and repulsive gravity. *Foundations of Physics*, 30:709–746.

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

Martin, C. (2002). *Gauging Gauge: Remarks on the Conceptual Foundations of Gauge Symmetry*. PhD thesis, University of Pittsburgh.

Martin, J. and Brandenberger, R. H. (2001). The trans-Planckian problem of inflationary cosmology. *Physical Review D*, 63:123501.

Matravers, D. R., Ellis, G. F. R., and Stoeger, W. R. (1995). Complementary approaches to cosmology: Relating theory and observations. *Quarterly Journal of the Royal Astronomical Society*, 36:29–45.

Maudlin, T. (1996). On the unification of physics. *Journal of Philosophy*, 93(3):129–44.

Mayo, D. G. (1996). *Error and the growth of experimental knowledge*. University of Chicago Press, Chicago.

McMullin, E. (1993). Indifference principle and anthropic principle in cosmology. *Studies in the History and Philosophy of Science*, 24(3):359–389.

McVittie, G. C. (1956). *General Relativity and Cosmology*. University of Illinois Press, first edition.

McVittie, G. C. (1965). *General Relativity and Cosmology*. University of Illinois Press, second edition.

Miloni, P. W. (1994). *Quantum vacuum: an introduction to quantum electrodynamics*. Academic Press, New York.

Misner, C. (2001). Interview with Charles Misner conducted by Chris Smeenk. 34 pp. manuscript, to be deposited in the Oral History Archives at the American Institute of Physics.

Misner, C. W. (1963). The flatter regions of Newman, Unti, and Tamburino's generalized Schwarzschild space. *Journal of Mathematical Physics*, 4:924–937.

Misner, C. W. (1967). Transport processes in the primordial fireball. *Nature*, 214:40–41.

Misner, C. W. (1968). The isotropy of the universe. *Astrophysical Journal*, 151:431–457.

Misner, C. W. (1969a). Absolute zero of time. *Physical Review*, 186:1328–1333.

Misner, C. W. (1969b). Mixmaster universe. *Physical Review Letters*, 22:1071–1074.

Misner, C. W. (1969c). Quantum cosmology. I. *Physical Review*, 186:1319–1327.

Misner, C. W. (1994). The mixmaster cosmological metrics. In Hobil, Burd, and Coley, editors, *Deterministic Chaos in General Relativity*, pages 317–328, New York. Plenum Press.

Misner, C. W. and Matzner, R. (1972). Dissipative effects in the expansion of the universe. I. *Astrophysical Journal*, 171:415–432.

Misner, C. W., Thorne, K., and Wheeler, J. A. (1973). *Gravitation*. W. H. Freeman & Co., New York.

Mohapatra, R. N. (1980). Cosmological constant, grand unification and symmetry behavior in early universe. *Physical Review D*, 22:2380–2383.

Mohapatra, R. N. and Senjanovic, G. (1979a). Broken symmetries at high temperature. *Phys. Rev.*, D20:3390–3398.

Mohapatra, R. N. and Senjanovic, G. (1979b). Soft CP violation at high temperature. *Physical Review Letters*, 42:1651–1654.

Morrison, M. (2000). *Unifying Scientific Theories: Physical Concepts and Mathematical Structures*. Cambridge University Press, Cambridge.

Mukhanov, V. F. and Chibisov, G. V. (1981). Quantum fluctuations and a nonsingular universe. *JETP Letters*, 33:532–535.

Mukhanov, V. F., Feldman, H. A., and Brandenberger, R. H. (1992). Theory of cosmological perturbations. part 1. classical perturbations. part 2. quantum theory of perturbations. part 3. extensions. *Physics Reports*, 215:203–333.

Munitz, M. K. (1962). The logic of cosmology. *British Journal for the Philosophy of Science*, 13:34–50.

Musgrave, A. (1974). Logical versus historical theories of confirmation. *British Journal for the Philosophy of Science*, 25:1–23.

Nambu, Y. (1961). Quasi-particles and gauge invariance in the theory of superconductivity. *Physical Review*, 117(3):648–663.

Nambu, Y. and Jona-Lasinio, G. (1961a). A dynamical model of elementary particles based on analogy with superconductivity. I. *Physical Review*, 122:345–358.

Nambu, Y. and Jona-Lasinio, G. (1961b). A dynamical model of elementary particles based on analogy with superconductivity. II. *Physical Review*, 124:246–254.

Nanopoulos, D. V., Olive, K. A., and Srednicki, M. (1983). After primordial inflation. *Physics Letters B*, 127:30–34.

Newton, I. (1999). *The Principia: Mathematical Principles of Natural Philosophy*. University of California Press, Berkeley, California. Assisted by Julia Budenz. Includes Cohen's *Guide to Newton's Principia*.

North, J. D. (1965). *The Measure of the Universe*. Oxford University Press, Oxford.

Norton, J. (1984). How Einstein found his field equations, 1912-1915. *Historical Studies in the Physical Sciences*, 14:253–316. Reprinted in Howard and Stachel 1989.

Norton, J. (1985). What was Einstein's principle of equivalence? *Studies in the History and Philosophy of Science*, 16:203–246. Reprinted in Howard and Stachel 1989.

Norton, J. (1993). General covariance and the foundations of general relativity: Eight decades of dispute. *Reports on Progress in Physics*, 56:791–858.

Norton, J. (1994). Science and certainty. *Synthese*, 99:3–22.

Norton, J. (1999). The cosmological woes of Newtonian gravitation theory. volume 7 of *Einstein Studies*, pages 271–323, Boston. Birkhäuser.

Norton, J. (2000). 'Nature is the realisation of the simplest conceivable mathematical ideas': Einstein and the canon of mathematical simplicity. *Studies in the History and Philosophy of Modern Physics*, 31:135–170.

Novikov, I. D. and Zel'dovich, Y. B. (1973). Physical processes near cosmological singularities. *Annual Review of Astronomy and Astrophysics*, 11:387–412.

Olive, K. A. (1990). Inflation. *Physics Reports*, 190:307–403.

Omnes, R. (1971). On the origin of matter and galaxies. *Astronomy and Astrophysics*, 10:228–245.

Ostriker, J. (2002). Interview with Jeremiah Ostriker conducted by Chris Smeenk. 130 pp. manuscript, to be deposited in the Oral History Archives at the American Institute of Physics.

Parker, L. (1969). Quantized fields and particle creation in the expanding universes. I. *Physical Review*, 183(5):1057–1068.

Parker, L. (1970). Quantized fields and particle creation in the expanding universes. II. *Physical Review D*, 3(2):346–356.

Parker, L. and Fulling, S. A. (1973). Quantized matter fields and the avoidance of singularities in general relativity. *Physical Review D*, 7:2357–2374.

Patzelt, H. (1990). On horizons in homogeneous isotropic universes. *Classical and Quantum Gravity*, 7:2081–2087.

Peacock, J. R. (1999). *Cosmological Physics*. Cambridge University Press, Cambridge.

Peebles, P. J. E. (1971). *Physical Cosmology*. Princeton University Press, Princeton.

Peebles, P. J. E. (1972). Light out of darkness vs. order out of chaos. *Comments on Astrophysics and Space Science*, 4:53–58.

Peebles, P. J. E. (1980). *Large-scale Structure of the Universe*. Princeton University Press, Princeton.

Peebles, P. J. E. (1993). *Principles of Physical Cosmology*. Princeton University Press, Princeton.

Peebles, P. J. E. (1999). Summary: Inflation and traditions of research. To appear in Pritzker Symposium on the Status of Inflationary Cosmology, ed. by Michael Turner. astro-ph/9905390.

Peebles, P. J. E. (2002). Interview with James Peebles conducted by Chris Smeenk. 64 pp. manuscript, deposited at the Oral History Archive of the American Institute of Physics.

Peebles, P. J. E. and Yu, J. T. (1970). Primeval adiabatic perturbation in an expanding universe. *Astrophys. J.*, 162:815–836.

Penrose, O. and Percival, I. C. (1962). The direction of time. *Proceedings of the Physical Society*, 79:605–616.

Penrose, R. (1979). Singularities and time-asymmetry. In Hawking, S. and Israel, W., editors, *General Relativity: An Einstein centenary survey*, pages 581–638. Cambridge University Press, Cambridge.

Penrose, R. (1989). Difficulties with inflationary cosmology. *Annals of the New York Academy of Sciences*, 271:249–264.

Peskin, M. E. and Schroeder, D. V. (1995). *An Introduction to Quantum Field Theory*. Perseus Books, Cambridge.

Pickering, A. (1984). *Constructing Quarks: A sociological history of particle physics*. University of Chicago Press, Chicago.

Polyakov, A. M. (1974). Particle spectrum in quantum field theory. *JETP Letters*, 20:194–195.

Preskill, J. P. (1979). Cosmological production of superheavy magnetic monopoles. *Physical Review Letters*, 43:1365–8. Reprinted in Bernstein and Feinberg, pp. 292-298.

Press, W. H. (1980). Spontaneous production of the Zel'dovich spectrum of cosmological fluctuations. *Physica Scripta*, 21:702–702.

Press, W. H. and Vishniac, E. T. (1980). Tenacious myths about cosmological perturbations larger than the horizon size. *Astrophysical Journal*, 239:1–11.

Ratra, B. and Peebles, P. J. E. (1995). Inflation in an open universe. *Physical Review D*, 52:1837–1894.

Raychaudhuri, A. (1955). Relativistic cosmology I. *Physical Review*, 98:1123–1126.

Rees, M. (2002). Interview with Martin Rees conducted by Chris Smeenk. 42 pp. manuscript, to be deposited at the Oral History Archive of the American Institute of Physics.

Reichenbach, H. (1956). *The Direction of Time*. University of Chicago Press, Berkeley.

Reichenbach, H. (1958). *The Philosophy of Space and Time*. Dover, New York. English translation by M. Reichenbach and J. Freund.

Renn, J., Sauer, T., Janssen, M., Norton, J., and Stachel, J., editors (2003). *General Relativity in the Making: Einstein's Zurich Notebook*. Number 1 in Genesis of General Relativity: Documents and Interpretation. Kluwer, Dordrecht. Forthcoming.

Rindler, W. (1956). Visual horizons in world models. *Monthly Notices of the Royal Astronomical Society*, 116:662–677.

Robertson, H. P. (1929). On the foundations of relativistic cosmology. *Proceedings of the National Academy of Science*, 15:822–829.

Roseveare, N. T. (1982). *Mercury's perihelion, from Le Verrier to Einstein*. Oxford University Press, New York.

Rothman, T. and Ellis, G. (1986). Can inflation occur in anisotropic cosmologies? *Physics Letters B*, 180:19–24.

Roush, S. (1999). *Conditions of Knowledge: The Weak Anthropic Principle, Selection Effects, Transcendental Arguments and Provisoes*. PhD thesis, Harvard University, Cambridge, MA.

Ruetsche, L. (2002). Interpreting quantum field theory. *Philosophy of Science*, 69:348–78.

Rugh, S. E. and Zinkernagel, H. (2001). The quantum vacuum and the cosmological constant problem. To appear in *Studies in the History and Philosophy of Modern Physics*, hep-th/0012253.

Ryan, M. P. and Shepley, L. (1975). *Homogeneous Relativistic Cosmologies*. Princeton University Press, Princeton.

Ryder, L. H. (1996). *Quantum field theory*. Cambridge University Press, Cambridge, second edition.

Sachs, R. G. (1987). *The physics of time reversal*. University of Chicago Press, Chicago.

Sakharov, A. 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*.

Sakharov, A. D. (1967). Violation of CP invariance, C asymmetry, and baryon asymmetry of the universe. *Soviet Physics JETP*, 5:32–35.

Sakharov, A. D. (1970). A multisheet cosmological model. Preprint, Moscow Institute of Applied Mathematics. Translated in *Collected Scientific Works*.

Salmon, W. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton University Press, Princeton.

Salmon, W. (1989). Four decades of scientific explanation. In Kitcher, P. and Salmon, W., editors, *Scientific Explanation*, Minnesota Studies in the Philosophy of Science. University of Minnesota Press, Minneapolis.

Salmon, W. C. (1990). Rationality and objectivity in science *or* Tom Kuhn meets Tom Bayes. In Savage, C. W., editor, *Scientific Theories*, volume 14 of *Minnesota Studies in the Philosophy of Science*, pages 175–204. University of Minnesota Press, Minneapolis.

Salmon, W. C. (1998). *Causality and Explanation*. Oxford University Press, Oxford.

Sandage, A. (1961). Ability of the 200-inch telescope to discriminate between world models. *Astrophysical Journal*, 133:355–380.

Sandage, A. (1970). Cosmology: A search for two numbers. *Physics Today*, 23:33–43.

Sanders, R. H. and McGaugh, S. S. (2002). Modified newtonian dynamics as an alternative to dark matter. astro-ph/0204521.

Sato, K. (1981). First-order phase transition of a vacuum and the expansion of the universe. *Monthly Notices of the Royal Astronomical Society*, 195:467–479.

Schafroth, M. R. (1958). Remark on the Meissner effect. *Physical Review*, 111:72–74.

Schild, A. (1960). Equivalence principle and red-shift measurments. *American Journal of Physics*, 28:778–780.

Schrödinger, E. (1957). *Expanding Universes*. Cambridge University Press, Cambridge.

Schücking, E. and Heckmann, O. (1958). Models of the universe. In *Proceedings Solvay Conference*, pages 1–10. Institut International de Physique Solvay, Bruxelles.

Schulmann, R., Kox, A. J., Janssen, M., and Illy, J., editors (1998). *Collected Papers of Albert Einstein. The Berlin Years: Correspondence, 1914-1918*, volume 8. Princeton University Press.

Sciama, D. (1955). On the formation of galaxies in a steady state model. *Monthly Notices of the Royal Astronomical Society*, 115:3–14.

Sciama, D. (1959). *The Unity of the Universe*. Doubleday and Co., Garden City.

Sciama, D. (1971). *Modern Cosmology*. Cambridge University Press, Cambridge, first edition.

Scott, S. and Szekeres, P. (1994). The abstract boundary — a new approach to singularities of manifolds. *Journal of Geometry and Physics*, 13:223–53.

Shafi, Q. and Vilenkin, A. (1984). Inflation with SU(5). *Physical Review Letters*, 52:691–694.

Sklar, L. (1993). *Physics and Chance: Philosophical issues in the foundations of statistical mechanics*. Cambridge University Press, Cambridge.

Sklar, L. (2000). *Theory and Truth: Philosophical critique within foundational science*. Oxford University Press, Oxford.

Smith, G. E. (2002a). From the phenomenon of the ellipse to an inverse-square force: Why not? In Malament, D., editor, *Reading Natural Philosophy*, pages 31–70. Open Court, Chicago.

Smith, G. E. (2002b). The methodology of the *Principia*. In Cohen, I. B. and Smith, G. E., editors, *Cambridge Companion to Newton*, pages 138–173. Cambridge University Press, Cambridge.

Smith, S. (2002c). Violated laws, *ceteris paribus* clauses, and capacities. *Synthese*, 130:235–264.

Smolin, L. (2000). The present moment in quantum cosmology: Challenges to the arguments for the elimination of time. In Durie, R., editor, *Time and the Instant*, pages 112–143. Clinamen Press, Manchester.

Sober, E. (1994). The principle of the common cause. In *From a Biological Point of View*, pages 158–174. Cambridge University Press, Cambridge. Originally publised in *Probability and Causation*, J. Fetzer (ed.), 1987.

Sober, E. (1999). Testability. *Proceedings and Addresses of the APA*.

Sober, E. (2001). Venetian sea levels, British bread prices, and the principle of the common cause. *British Journal for the Philosophy of Science*, 52:331–346.

Spergel, D. N. et al. (2003). First year Wilkinson microwave anisotropy probe (WMAP) observations: Determination of cosmological parameters. astro-ph/0302209.

Stachel, J. (2001). The story of Newstein, or: is gravity just another pretty force? Unpublished manuscript.

Starobinsky, A. (1978). On a nonsingular isotropic cosmological model. *Soviet Astronomy Letters*, 4:82–84.

Starobinsky, A. (1979). Spectrum of relict gravitational radiation and the early state of the universe. *JETP Letters*, 30:682–685.

Starobinsky, A. (1980). A new type of isotropic cosmological models without singularity. *Physics Letters B*, 91:99–102.

Starobinsky, A. (1982). Dynamics of phase transitions in the new inflationary scenario and generation of perturbations. *Physics Letters B*, 117:175–178.

Starobinsky, A. (1983). The Perturbation Spectrum Evolving from a Nonsingular Initially De-Sitter Cosmology and the Microwave Background Anisotropy. *Soviet Astronomy Letters*, 9:302–304.

Starobinsky, A. (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.

Steel, D. (2002). Unifying power and explanatory asymmetries: Why unification needs causation. Submitted to *Philosophical Studies*.

Steigman, G. (1976). Observational tests of antimatter cosmologies. *Annual Review of Astronomy and Astrophysics*, 14:339–372.

Steinhardt, P. (2002). Interview with Paul Steinhardt conducted by Chris Smeenk. 100 pp. manuscript, to be deposited in the Oral History Archives at the American Institute of Physics.

Steinhardt, P. J. and Turner, M. S. (1984). A prescription for successful new inflation. *Physical Review D*, 29:2162–2171.

Stewart, J. M. (1968). Neutrino viscosity in cosmological models. *Astrophysical Letters*, 2:133–135.

Stoeger, W. R., Ellis, G. F. R., and Hellaby, C. (1987). The relationship between continuum homogeneity and statistical homogeneity in cosmology. *Monthly Notices of the Royal Astronomical Society*, 226:373–381.

Stoeger, W. R., Maartens, R., and Ellis, G. F. R. (1995). Proving almost-homogeneity of the universe: an almost Ehlers-Geren-Sachs theorem. *Astrophysical Journal*, 443:1–5.

Streater, R. F. and Wightman, A. S. (1964). *PCT, spin and statistics, and all that*. W. A. Benjamin, New York.

Sullivan, W. (1965). Signals imply a "Big Bang" universe. *New York Times*, pages A1, A18. May 21, 1965.

't Hooft, G. (1974). Magnetic monopoles in unified gauge theories. *Nuclear Physics*, B79:276–284.

Tariq, N. and Tupper, B. O. J. (1992). Conformal symmetry inheritance with cosmological constant. *Journal of Mathematical Physics*, 33:4002–4007.

Taub, A. (1951). Empty space-times admitting a three parameter group of motions. *Annals of Mathematics*, 53:472–490.

Thorne, K. (1967). Primordial element formation, primordial magnetic fields, and the isotropy of the universe. *Astrophysical Journal*, 148:51–68.

Tipler, F. J. (1980). General relativity and the eternal return. In Tipler, F. J., editor, *Essays in General Relativity*, pages 21–37. Academic Press, New York.

Tolman, R. C. (1934). *Relativity, Thermodynamics, and Cosmology*. Oxford University Press, Oxford. Reprinted by Dover.

Torretti, R. (1983). *Relativity and Geometry*. Pergamon Press, Oxford. Reprinted by Dover, 1996.

Torretti, R. (2000). Spacetime models of the world. *Studies in the History and Philosophy of Modern Physics*, 31(2):171–186.

Tryon, E. (1973). Is the universe a vacuum fluctuation? *Nature*, 246:396–397.

Turok, N. (2002). Interview with Neil Turok conducted by Chris Smeenk. 17 pp. manuscript, to be deposited in the Oral History Archives at the American Institute of Physics.

Uffink, J. (1999). The principle of the common cause faces the Bernstein paradox. *Philosophy of Science*, 66 (Proceedings):S512–S525.

Unruh, W. G. and Wald, R. M. (1989). Time and the interpretation of canonical quantum gravity. *Physical Review D*, 40:2598–2614.

Vachaspati, T. and Trodden, M. (2000). Causality and cosmic inflation. *Physical Review D*, 61:3502–06. gr-qc/9811037.

van Fraassen, B. C. (1980). *The Scientific Image*. Clarendon Press, Oxford.

van Fraassen, B. C. (1985). Empiricism in the philosophy of science. In Churchland and Hooker (1985), pages 245–308.

van Fraassen, B. C. (1989). *Laws and Symmetry*. Clarendon Press, Oxford.

Vanderburgh, W. (2001). *Dark Matters in Contemporary Astrophysics: A Case Study in Theory Choice and Evidential Reasoning*. PhD thesis, University of Western Ontario.

Veltman, M. J. G. (1974). Cosmology and the Higgs mechanism. Rockefeller University Preprint.

Veltman, M. J. G. (1975). Cosmology and the HIGGS mass. *Physical Review Letters*, 34:777.

Veltman, M. J. G. (2000). Nobel lecture: from weak interactions to gravitation. *Reviews of Modern Physics*, 72:341–349.

Vilenkin, A. (1983). The birth of inflationary universes. *Physical Review D*, 27:2848.

Vilenkin, A. (1998). The wave function discord. *Phys. Rev.*, D58:067301.

Wagoner, R. V., Fowler, W. A., and Hoyle, F. (1967). On the synthesis of light elements at very high temperatures. *Astrophysical Journal*, 148:3–49.

Wainwright, J. and Ellis, G. F. R. (1997). *Dynamical Systems in Cosmology*. Cambridge University Press, Cambridge.

Wald, R. (1983). Asymptotic behavior of homogeneous cosmological models in the presence of a positive cosmological constant. *Physical Review D*, 28:2118–2120.

Wald, R. (1984). *General Relativity*. University of Chicago Press, Chicago.

Wald, R. (1992). Correlations beyond the horizon. *General Relativity and Gravitation*, 24(11):1111–1116.

Wald, R. (1994). *Quantum field theory in curved spacetimes and black hole thermodynamics*. University of Chicago Press.

Wald, R. M. (1993). Correlations Beyond the Cosmological Horizon. In Gunzig, E. and Nardone, P., editors, *The Origin of Structure in the Universe. Proceedings of the NATO Advanced Research Workshop on the Origin of Structure in the Universe*, pages 217–226, Dordrecht. Kluwer.

Weinberg, S. (1967). A model of leptons. *Physical Review Letters*, 19:1264–66.

Weinberg, S. (1972). *Gravitation and Cosmology*. John Wiley & Sons, New York.

Weinberg, S. (1974a). Gauge and global symmetries at high temperature. *Physical Review D*, 9(12):3357–3378.

Weinberg, S. (1974b). Recent progress in gauge theories of the weak, electromagnetic, and strong interactions. *Reviews of Modern Physics*, 46(2):255–277.

Weinberg, S. (1977). *The First Three Minutes*. Basic Books, Inc., New York.

Weinberg, S. (1980). Conceptual foundations of the unified theory of weak and electromagnetic interactions. *Reviews of Modern Physics*, 52(3):515–523.

Weinberg, S. (1989). The cosmological constant problem. *Reviews of Modern Physics*, 61:1–23.

Weinberg, S. (1995). *The quantum theory of fields*, volume I: Foundations. Cambridge University Press.

Weinberg, S. (1996). *The quantum theory of fields*, volume II: Modern Applications. Cambridge University Press.

Weingard, R. (1990). Realism and the global topology of space-time. In Bhaskar, R., editor, *Harré and his Critics*, pages 112–121. Basil Blackwell.

Will, C. M. (1993). *Was Einstein right?: putting general relativity to the test*. Basic Books, New York, 2nd edition.

Will, C. M. (2001). The confrontation between general relativity and experiment. *Living Reviews in Relativity*, 4. http://www.livingreviews.org/Articles/Volume4/2001-4will/, cited on 15 Aug. 2001.

Worrall, J. (1985). Scientific discovery and theory-confirmation. In Pitt, J. C., editor, *Change and progress in modern science*, pages 301–331, Dordrecht. D. Reidel.

Worrall, J. (1989). Fresnel, Poisson and the white spot: The role of successful predictions in the acceptance of scientific theories. In Gooding, D., Pinch, T., and Schaffer, S., editors, *The uses of experiment: Studies in the natural sciences*, pages 135–57. Cambridge University Press, Cambridge.

Yokoyama, J. (1989). Relic magnetic monopoles in the inflationary universe. *Physics Letters B*, 231:49–52.

Zee, A. (1979). Broken-symmetric theory of gravity. *Physical Review Letters*, 42:417–421.

Zee, A. (1980). Horizon problem and broken-symmetric theory of gravity. *Physical Review Letters*, 44:703–706.

Zee, A. (1982). Calculating Newton's gravitational constant in infrared stable Yang-Mills theories. *Physical Review Letters*, 48:295–298.

Zel'dovich, Y. B. (1967). Cosmological constant and elementary particles. *JETP Letters*, 6:316–317.

Zel'dovich, Y. B. (1968). The cosmological constant and the theory of elementary particles. *Soviet Physics Uspekhi*, 11:381–393. Translated by J. G. Adashko.

Zel'dovich, Y. B. (1972). A hypothesis, unifying the structure and the entropy of the universe. *Mon. Not. Roy. Astron. Soc.*, 160:1–3.

Zel'dovich, Y. B. (1974). Creation of particles in cosmology. In *IAU Symp. 63: Confrontation of Cosmological Theories with Observational Data*, volume 63, pages 329–333.

Zel'dovich, Y. B. (1980). Cosmological fluctuations produced near a singularity. *Monthly Notices of the Royal Astronomical Society*, 192:663–667.

Zel'dovich, Y. B. (1981). Vacuum theory - A possible solution to the singularity problem of cosmology. *Soviet Physics Uspekhi*, 24:216–230.

Zel'dovich, Y. B. and Khlopov, M. Y. (1978). On the concentration of relic magnetic monopoles in the universe. *Physics Letters B*, 79:239–41.

Zel'dovich, Y. B., Kobzarev, I. Y., and 'Okun, L. B. (1975). Cosmological consequences of a spontaneous breakdown of a discrete symmetry. *Soviet Physics JETP*, 40:1–5.

Zel'dovich, Y. B. and Novikov, I. (1983). *Relativistic Astrophysics, Volume II: The Structure and Evolution of the Universe*. University of Chicago Press, Chicago. G. Steigman, ed. and L. Fishbone, trans.

Zel'dovich, Y. B. and Novikov, I. D. (1967). The Uniqueness of the Interpretation of Isotropic Cosmic Radiation with T = 3 K. *Soviet Astronomy*, 11:526–528.

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

Zinkernagel, H. (2002). Cosmology, particles, and the unity of science. *Studies in the History and Philosophy of Modern Physics*, 33:493–516.