Inflation and the Origins of Structure

Chris Smeenk

1 Introduction

The initial motivations for a theory are sometimes rendered dubious or superfluous by later work. The epistemic load borne by motivating ideas in the first stage of theoretical construction is shifted onto other ideas as work proceeds, leaving the original arguments with a largely ornamental rather than structural role. For example, Einstein described one of the three foundational ideas of general relativity (GR) as Mach's principle, roughly speaking the claim that spacetime geometry should be fully determined by the distribution of matter without appeal to "background structures." This principle was one of Einstein's guiding ideas in the discovery of GR, but the status of Mach's principle in the finished theory is quite controversial. When contemporary physicists make the case for accepting GR as our best theory of gravitation, Mach's Principle does not play a central role. It is fairly easy to produce other examples of this sort, in which the contemporary path to justification does not recapitulate the path to discovery.

The central idea of inflationary cosmology is that the early universe passed through a phase of exponential expansion driven by a scalar field displaced from the true minimum of its potential energy. Guth (1981) provided a rationale for this idea that proved to be quite persuasive: inflation nearly eliminates the need for special initial conditions required by the standard model of cosmology. It was soon discovered that inflation also suggested a solution to a long-standing problem in relativistic cosmology: what is the origin of the seeds for the formation of structure in the universe? A recent textbook draws a distinction between the original rationale for inflation, as a "theory of initial conditions," and a rationale based on predictions for the seeds of structure, as a "theory of the origins of structure":

> ...[T]hese problems [related to initial conditions] can no longer be regarded as the strongest motivation for inflationary cosmology because it is not at all clear that they could ever be used to falsify inflation. [...] 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 observed structures, and observations can discriminate strongly between them. ... *Inflation as the origin of structure is therefore very much a proper science of prediction and observation.* (Liddle and Lyth 2000, p. 5; my emphasis)

Liddle and Lyth clearly regard the initial motivations for inflation as not sufficiently empirical, unlike the case for inflation that can be made given its connection with structure formation. Below I will argue, in agreement with Liddle and Lyth, that there is an important contrast between the historical motivations for inflationary cosmology and the strongest case that can now be made in its favor. But I will suggest a different way of characterizing the contrast, based on how informative different bodies of data are regarding inflation.

The opening move made by inflationary cosmologists is to treat various properties of the early universe as the consequences of the dynamical evolution of a scalar field (or fields) in the early universe. In his persuasive initial case for inflation, Guth (1981) emphasized that such evolution could lead to a uniform, flat universe for a large range of initial conditions. Shortly after Guth's paper appeared, several groups of cosmologists formulated accounts of the creation of seeds for structure formation during inflation. The mechanism for generating density perturbations is the most fruitful consequence of the inflationary gambit, in two different senses. First, the problems Guth emphasized in presenting the theory were regarded as "enigmas" of the standard model of cosmology, when they were discussed at all. By way of contrast, the status of initial "seed" fluctuations was a major problem facing an appealing account of the origin of structure. Given that gravity should be the dominant force at large length scales, it is natural to suppose that structures such as galaxies evolved by the growth of small perturbations to an almost uniform initial distribution of matter. As described below in §2, while this gravitational instability picture was appealing it seemed to require an extremely implausible initial distribution of matter. Inflation countered this objection and provided theorists with a way of calculating the density perturbations as a consequence of a stage of exponential expansion. §3 recounts the historical route by which cosmologists developed this account, and contrasts the case of structure formation with the initial motivations for inflation. §4 turns to the second sense in which this aspect of inflation has been particularly fruitful, namely that it provides the grounds for a detailed comparison with competing theories based on observations of temperature fluctuations in the cosmic microwave background radiation (CMBR). The final concluding section attempts to move beyond the way in which questions regarding the empirical status of inflation have been couched in the physics literature, in terms of Popperian "falsifiability."

2 Structure Formation in The Standard Model

By the early 70s two aspects of what Weinberg (1972) dubbed the Standard Model of Cosmology were well understood theoretically, supported by observational evidence, and accepted as a starting point for further research by most cosmologists.[1] First, the expanding universe models of general relativity, the Friedmann-Lemaître-Robertson-Walker (FLRW) models, were taken to provide an approximate description of the overall structure and evolution of the universe at some suitably large length scale. The properties of the FLRW models were explored in the early days of relativistic cosmology. Due to their symmetry, the dynamics of general relativity is reduced to simple equations relating the scale factor $R(t)$ to the matter - energy distribution. By 1970 almost all cosmologists had accepted the FLRW models as a useful approximation and had turned to the more specific task of measuring the expansion with sufficient accuracy to choose the best model (see, e.g., Sandage 1970).

Second, the theory accounted for two striking features of the universe as relics of the "primeval fireball." Nuclear reactions in the early universe governed by the rate of expansion leave a telling trace — a helium abundance of about $26-28\%$ according to a calculation by Peebles (1966), in agreement with observations. Further development of the theory of big bang nucleosynthesis clarified the dependence of the primordial element abundances

[1]The steady state theory was no longer a serious rival to the standard "big bang" model by this time, although a small group of proponents (including Hoyle, Narlikar, and others) continued to explore the idea and to challenge the empirical underpinnings of the big bang model. See Kragh (1996) for a thorough discussion of earlier stages of the debate and the resolution of the controversy in favor of the big bang model.

on various cosmological parameters. The cosmic microwave background radiation (CMBR) observed in 1965 was a second natural consequence of a hot big bang. In the early universe radiation and matter are coupled due to interactions, but as the temperature drops low enough for the existence of stable nuclei the universe becomes effectively transparent to photons.[2] The photons then cool adiabatically with the expansion of the universe while maintaining a black-body spectrum, and they carry a tremendous amount of information regarding the universe at the time of recombination. Since 1965 a series of increasingly sophisticated observational missions have succeeded in extracting more and more of this information. Although subsequent research has enriched both ideas considerably, the fundamentals were in place by the early 70s and are presented in the influential texts Weinberg (1972) and Peebles (1971).

By contrast with these successes, the Standard Model lacked a compelling account of structure formation. Weinberg (1972) prefaced his discussion with the caveat that:

> ...[w]e still do not have even a tentative quantitative theory of the formation of galaxies, anywhere near so complete and plausible as our theories of the origin of the cosmic abundance of helium or the microwave background (p. 562).

Providing such an account has remained one of the primary goals of cosmology. Unlike the successful aspects of the Standard Model, in the case of structure formation it has been much more difficult to link tractable pieces of theory to observations. This reflects the intrinsic difficulty of the subject, which requires integrating a much broader array of physical ideas than required for the study of nucleosynthesis or the FLRW models. This section will give a brief overview of the development of the field up to 1980, focusing on the status of initial conditions for structure formation.

The ideas Weinberg (1972) described as a speculative part of the Standard Model were first explored by Lemaître. Newtonian gravity enhances clumping of a nearly uniform distribution of matter, as matter is attracted more strongly to regions of greater density. In the early stages of clumping, fluctuations in density will be small enough that they can be treated as first-order perturbations to a background cosmological model. This will be the case if the density contrast $\Delta =: \frac{\delta\rho}{\rho}$ is less than 1, where $\delta\rho$ is the density enhancement over the background density ρ. A theory of the evolution of small fluctuations must be supplemented on both ends, so to speak. The theory assumes as given an initial spectrum of small fluctuations that are then enhanced via dynamical evolution. An appealing possibility is that the dynamics is unstable, leading to exponential growth of small fluctuations. Then, like the onset of turbulence in fluid mechanics, details regarding the initial state would be relatively unimportant. On the other end, the theory extends up to the point when the fluctuations "freeze out" from the cosmological expansion, and begin to collapse into structures with much higher density contrasts (such as $\Delta \approx 10^6$ for a typical galaxy). Developing a theory governing this later stage of structure formation poses enormous challenges: perturbation theory does not apply, and the non-gravitational interactions of the constituents of the collapsing region can no longer be ignored.[3] Despite these limitations,

[2]Recombination refers to the process by which nuclei capture free electrons and form neutral hydrogen and helium; although in the Standard Model, there was no earlier time at which stable nuclei existed, the historical term with the misleading "re-" has been retained. During recombination, photons decouple from matter as the cross section for Thomson scattering drops to zero.

[3]Modern studies of the non-linear regime employ numerical simulations, although there are a number of analytic techniques that were developed to study non-linear evolution during this time (e.g., Press and Schechter 1974). See, e.g., Chapter 17 of Peacock (1999) for an introduction.

the theory of structure formation via gravitational enhancement of non-uniformities covers a large dynamical range. If successful, it would provide a link between the physical processes in the very early universe responsible for the initial fluctuations and the observationally accessible imprints of perturbations at later times.

Lifshitz (1946) was the first to treat the evolution of linear perturbations to a background model in general relativity, only to reject gravitational instability as a viable account of structure formation. He showed that in an FLRW model, the density contrast as a function of time grows very slowly.[4] This result is surprising given the contrast with the classical account of instability due to Jeans (1902). Jeans derived an equation governing the evolution of small perturbations of a fluid including Newtonian gravity, and showed that the behavior of different modes depends on how their wavelength compares to a a critical wavelength, the Jeans length λ_J.[5] For modes with $\lambda = \lambda_J$ there is a balance between the pressure of the fluid, resisting collapse, and the gravitational force; perturbation modes with $\lambda < \lambda_J$ exhibit oscillatory behavior, whereas those with $\lambda > \lambda_J$ are unstable and grow exponentially. Physically, in the final case the matter density is sufficient to trigger gravitational collapse, leading to exponential growth of the amplitude of the fluctuation. If such rapid growth occurred in an expanding background as well, it would be possible for galaxies to form via gravitational enhancement of thermal fluctuations in the matter density, which Lifshitz (and many others) took to be a reasonable posit for the initial conditions. In this case the fluctuations away from uniformity would be given by the Poisson distribution, $\Delta \propto N^{-1/2}$ for N particles; for a galaxy-scale lump of particles, say 10^{68} particles, thermal fluctuations would give a low density contrast $\Delta_i \propto 10^{-34}$. However, Lifshitz showed that cosmological expansion works against gravitational instability, with the density contrast growing slowly ($\Delta(t) \propto t^{2/3}$) during the matter-dominated era in an expanding model. (Pressure prevents growth of the density contrast during the earlier radiation-dominated era.)[6] Suppose that the initial fluctuation spectrum is imprinted at, say, $t_i = 1$ second (following Bonnor 1956). There is then not nearly enough time for the initial fluctuations to grow into galaxies — the growth would be on the order of 10^{12} rather than the 10^{40} that is needed. Lifshitz concluded that gravitational instability fails to account for the formation of galaxies. Subsequent work on linear perturbation theory corrected and augmented Lifshitz's analysis in several significant respects, but with little impact on this line of argument.[7]

In the 50s and early 60s many theorists found this criticism so compelling that they pursued alternative accounts of structure formation. Gamow, for example, turned to developing a theory based on primeval turbulence.[8] However, Lifshitz's line of argument requires

[4]Lemaître (1933) and Tolman (1934) studied the evolution of a spherical region of higher density in a background FLRW model prior to Lifshitz's work; although it was not the focus of their work, their results also show that the density contrast grows very slowly.

[5]See, e.g., Longair (2008), Chapter 11, or Weinberg (2008), Chapter 5, for modern introductions to linear perturbation theory, which discuss the Jeans length and the derivation of the results sketched here.

[6]Lifshitz (1946) analyzed the behavior of small perturbations for two different equations of state, corresponding to radiation-dominated expansion, i.e. $p = \rho/3$, where p is the pressure and ρ the energy density, and matter-dominated expansion with $p = 0$. See, e.g., Longair (2008) for a modern treatment.

[7]See Peebles (1980), pp. 20 - 25, and Longair (2006), chapter 15 for historical overviews of this work.

[8]Gamow and Teller (1939) advocated an account of structure formation based on gravitational instability that is undermined by Lifshitz's results (as Lifschitz explicitly noted). Gamow (1952, 1954) are the original papers on the turbulence theory; see **?** for a critical review of Gamow's proposal and other similar ideas. Two other problems with the gravitational instability account were also important in motivating the search for alternatives. First, there is no preferred length or mass scale in general relativity (with the cosmological constant set to zero), so it is unclear how to introduce scales such as the mass of a typical galaxy (see Harrison 1967a,b for a detailed discussion of this point). Second, alternative accounts often claimed to give

an assessment of whether a given spectrum of density fluctuations at early times is plausible. Even a spectrum of thermal fluctuations is not immediately ruled out; Bonnor's argument shows that thermal fluctuations at $t_i = 1s$ will not undergo sufficient growth, but one can treat t_i as a free variable and simply impose the fluctuation spectrum at an earlier time. Such an initial fluctuation spectrum is still mysterious, as we will see in more detail shortly. But the enigmatic nature of the initial conditions was not a sufficient objection to cosmologists who explicitly adopted a more phenomenological approach to galaxy formation (see, e.g., Harrison 1968, Peebles 1968, Zel'dovich 1965). All the available cosmological theories required some specification of the initial conditions, and the gravitational instability account is not obviously more objectionable in this respect. Furthermore, projecting backwards to find the required initial conditions could provide insight into new physics relevant in the early universe. The discovery of the CMBR provided an important new constraint along with the potential to observationally establish the fluctuation spectrum at the time of decoupling. The phenomenological approach focused on giving a more precise characterization of the initial fluctuations that were required for gravitational instability along with a detailed account of their dynamical evolution over time.

Throughout the 70s theorists developed competing accounts of structure formation with the common aim of describing the evolution of the different physical degrees of freedom involved — radiation, baryonic matter, and the gravitational field. Solving the complete set of equations capturing all of the details of their interactions and dynamics, the coupled Boltzmann-Einstein equations, would have been computationally intractable. But given the background of an FLRW model, different physical effects are dominant at different stages of evolution. Initial matter and radiation perturbations would in general be a combination of two distinct modes:[9]

- *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.

One can then analyze the evolution of these distinct perturbation modes through different stages of the universe's history. Prior to recombination, radiation ionizes the baryons and the photons and free electrons are coupled via Thomson scattering. As a result, fluctuations in the baryonic matter and radiation move together like a single fluid (Peebles 1965); galactic-scale perturbations undergo acoustic oscillations during this phase. In the later matter-dominated era, radiation and matter decouple and the matter fluctuations can be treated in isolation along the lines of Lifshitz's analysis, and galactic-scale perturbations grow with $\Delta(t) \propto t^{2/3}$.

There were also debates regarding the appropriate initial spectrum and later stages of structure formation (see also Longair 2006). Two different schools of thought dominated the field: Zel'dovich's school focused on solutions in which large "blinis" (pancakes) formed

natural explanations of features of galaxies, such as their rotation and spiral structure.

[9]This terminology is due to Zel'dovich (1966). The factor of $\frac{4}{3}$ arises since the energy density of radiation is $\propto T^4$, compared to T^3 for matter (where T is the temperature). 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.

first from adiabatic perturbations, fragmenting into galaxies and structures much later due to non-gravitational processes. The other school of thought led by Peebles developed a "bottom-up" scenario, in which initial isothermal fluctuations developed into proto-galaxies with larger structures forming later by hierarchical clustering. Despite the stark differences between the account these theories gave of later stages of structure formation, they had similar implications for the epoch of recombination.

Both schools of thought also needed to address gravitational perturbations, and there was a natural choice for the initial spectrum. Harrison (1970), Peebles and Yu (1970), and Zel'dovich (1972)[10] proposed that the gravitational perturbations are scale-invariant in the sense that $\Delta|_\lambda = constant$ when λ, the perturbations' wavelength, is equal to the Hubble radius, $\lambda = H^{-1}$.[11] The arguments for this proposal differed. Harrison (1970) placed upper and lower bounds on the spectrum, arguing that for $n > 1$ the large amplitude perturbations on small scales would have produced black holes in the early universe and for $n < 1$ the perturbations on large scales would be too small to act as seeds for galaxy formation. Zel'dovich appealed to earlier work (**?**) to argue that scale-independent perturbations with an amplitude of $\Delta|_\lambda \approx 10^{-4}$ are compatible with large-scale structure. This argument constrains the large scale perturbations that would become galaxies; Zel'dovich argues more obscurely that small-scale perturbations with the same features could account for the observed entropy per baryon. Peebles and Yu (1970), by contrast, treat the spectrum as the appropriate way of characterizing "cosmological white noise." In any case, one crucial feature of the HPZ spectrum is that it lacks any characteristic length scale. For different wavelengths the perturbation amplitude is fixed at different times: in an expanding universe, the wavelength λ increases with the scale factor $R(t)$ whereas the Hubble radius increases at a slower rate as the expansion slows.[12] 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. Estimates of the magnitude of density perturbations when length scales associated with galaxies crossed the Hubble radius fell within the range $\Delta \approx 10^{-3} - 10^{-4}$. In addition, the initial perturbations were often also assumed to be "random" in the sense that the mass found within a sphere of fixed radius has a Gaussian distribution (for different locations of the sphere).[13]

Two features of HPZ spectrum are particularly puzzling. The first puzzle arises from the causal structure of the FLRW models. Even though the distance between freely falling particles decreases as $t \to 0$, the decrease is not rapid enough to insure that sufficiently distant regions of the universe were in causal contact. More precisely, the FLRW models have particle horizons. Horizons in cosmology measure the maximum distance light travels

[10]Peebles and Yu clearly propose the same spectrum (see, in particular, pp. 829-830), but, for reasons unknown to me, it is sometimes called the Harrison-Zel'dovich spectrum. Although Zel'dovich (1972) cites Harrison's paper he is usually credited with independent discovery of the idea, which can be traced back to earlier work with Sunyaev.

[11]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-Peebles-Zel'dovich spectrum corresponds to a choice of $n = 1$ (given that the volume element in the inverse Fourier transform is $\frac{dk}{k}$; for the other conventional choice, $k^2 dk$, we then have $n = -3$). The Hubble radius has the appropriate dimension, length: restoring c, it is given by $\frac{c}{H}$, and the Hubble constant H has units of km per second per megaparsec.

[12]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.

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

within a given time period; the horizon delimits the spacetime region from which signals emitted at some time t_e traveling at or below the speed of light could reach a given point.[14] The "particle horizon" is defined as the limiting case $t_e \to 0$. The existence of particle horizons in the FLRW models indicates that distant regions are not in causal contact. Many discussions refer to a related quantity, the Hubble radius H^{-1}, as the "horizon."[15] A simple scaling argument shows that in standard FLRW expansion perturbation wavelengths "cross the horizon": the perturbation wavelengths simply scale with the expansion whereas H^{-1} scales as $H^{-1} \propto R^{1/n}$ for $R(t) \propto t^n$. For the length scale associated with a galaxy, horizon crossing occurs at around $t \approx 10^9$ seconds. What is puzzling is that the HPZ spectrum requires the perturbations to be coherent prior to this time, at a length scale larger than the Hubble radius. One response to this puzzle was to hope that new physics would lead to a different causal structure of the early universe. 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.[16]

But Bardeen's focus on particle horizons as a fundamental obstacle set him apart from others in the field; Peebles (1980), for example, 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.

The second puzzle regards the amplitude of the perturbations as they crossed the horizon. While this could be treated as a parameter to be fixed by observations, many theorists hoped for a further physical account of how this amplitude was fixed in the early universe. One can evolve backwards to determine the amplitude of the fluctuation spectrum at a given "initial" time t_i. For t_i on the order of the Planck time, for example, these fluctuations are much *smaller* than thermal fluctuations, which are taken to be physically plausible.[17] It seems inappropriate to treat t_i as a free variable, choosing when to "imprint" a spectrum of

[14] Following Rindler (1956), a horizon is the surface in a time slice t_0 separating particles moving along geodesics that could have been observed from a worldline γ by t_0 from those which could not. The distance to this surface, for signals emitted at a time t_e, is given by:

$$d = R(t_0) \int_{t_e}^{t_0} \frac{dt}{R(t)} \tag{1}$$

Different "horizons" correspond to different choices of limits of integration. The integral converges for $R(t) \propto t^n$ with $n < 1$, which holds for matter or radiation-dominated expansion. Thus the integral for the particle horizon ($\lim_{t \to 0}$) converges for the FLRW models. See Ellis and Rothman (1993) for a clear introduction to horizons.

[15] The Hubble radius is more aptly called the speed of light sphere, given that objects at that distance appear to move at speed c due to the expansion. See Ellis and Rothman (1993) for further discussion of the distinction between the Hubble radius and the particle horizon.

[16] Energy conditions are constraints on what is taken to be a reasonable source for the gravitational field equations. Roughly speaking, the strong energy condition requires that the stresses in matter will not be so large as to produce negative energy densities. Formally, $T_{ab}\xi^a\xi^b \geq \frac{1}{2}\mathrm{Tr}(T_{ab})$ for every unit timelike ξ^a; for a perfect fluid, this implies that $\rho + 3p \geq 0$, where ρ is the energy density and p is the pressure. As Bardeen notes, if the strong energy condition fails then there are solutions such that the integral in eqn. (1) diverges.

[17] For example, Blau and Guth (1987) compare the density contrast imposed at $t_i = 10^{-35}$ seconds to the fluctuations obtained by evolving backwards from the time of recombination implies $\Delta \approx 10^{-49}$ at t_i, nine orders of magnitude *smaller* than thermal fluctuations.

thermal fluctuations such that the amplitudes match observations. The Planck time is often singled out on dimensional grounds as the scale at which quantum gravity effects should become important. But in the absence of a successor theory, it is unclear how to delimit the boundary of applicability of classical GR and then choose a plausible "initial" perturbation spectrum.

By the late 70s and early 80s, several cosmologists had greater ambitions than merely giving a phenomenological account of structure formation. They sought to understand the origins of initial perturbations based on new physics applicable to the early universe. Those sharing this ambition could draw 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.[18] Sakharov (1966) was the first to propose a detailed quantum description of the initial perturbations — remarkably, before the discovery of the CMBR. 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 several theorists explored the implications of early universe phase transitions for structure formation, in particular the production of topological defects (discussed in more detail below). This work, along with studies of other possible impacts of phase transitions, illustrates that giving a physical account of the earliest stages of structure formation came to be regarded as a viable research topic. As of 1980 the field was wide open, with the potential to draw on ideas in general relativity and quantum gravity or the many novel ideas recently introduced in particle physics.

In addition to puzzles regarding the initial perturbations, both prevailing approaches to structure formation were threatened by tightening observational constraints based on the isotropy of the temperature of the CMBR. Partridge (1980) reached sensitivities of $\Delta T/T \approx 10^{-4}$ in isotropy measurements, and at this level he should have detected fluctuations according to either of the prevailing accounts of structure formation. This problem, along with other events such as experimental evidence in favor of a massive neutrino, led theorists to add hot and cold dark matter to their models of structure formation starting in the early 80s (see, for example, Peebles 1982 and Pagels 1984).[19] The early dark matter models established the compatibility between the observational upper limits on temperature anisotropies in the CMBR and the idea of structure formation via gravitational instability. Adding cold dark matter helps to reconcile the uniformity of the CMBR with later clumpiness of matter because, roughly speaking, the cold dark matter decouples from the baryonic matter and radiation early, leaving a minimal imprint on the CMBR, yet after recombination the cold dark matter perturbations regenerate perturbations in the baryonic matter sufficiently large to seed structure formation.

[18] 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.

[19] "Hot" vs. "cold" refers to the thermal velocities of relic particles for different types of dark matter. Hot dark matter decouples while still "relativistic," in the sense that the momentum is much greater than the rest mass, and relics at late times would still have large quasi-thermal velocities. Cold dark matter is "non-relativistic" when it decouples, meaning that the momentum is negligible compared to the rest mass, and relics have effectively zero thermal velocities.

3 Inflationary Cosmology

Many contemporary textbooks on structure formation use the puzzles regarding initial perturbations described above to set the stage for the entrance of inflationary cosmology. Rather than pulling the initial spectrum out of a hat, as one might suspect of the earlier proposals, the inflationary theorist can pull an HPZ spectrum with an appropriate amplitude out of the vacuum fluctuations of a quantum field. The performance is captivating because it displays the possibility of *calculating* the features of the initial spectrum from physical principles. This section will review the route by which the theorists discovered this feature of inflation, and assess the importance of this aspect of the idea by contrast with the other features of inflation emphasized by Guth (1981).

The essential idea of inflation is that the early universe went through a transient phase of de Sitter-like expansion.[20] During this phase the scale factor grows exponentionally with time, $R(t) \propto e^{\chi t}$, compared to the much more sedate radiation-dominated FLRW expansion with $R(t) \propto t^{1/2}$. The idea of modifying FLRW expansion in this way had been suggested several times prior to inflation (see Smeenk 2005), and the earlier proposals shared two common problems. First, what is the physical source of the accelerated expansion? I will refer to this as the source problem. The source could not be garden-variety matter or radiation, because to drive a stage of exponential expansion it would have to violate the strong energy condition typically assumed to hold for reasonable matter fields.[21] Second, how does the exponential expansion transition into the usual FLRW expansion with appropriate matter and energy densities? Solving this second problem, the transition problem, requires an explanation of how the physical source of the expansion ceased to be dynamically relevant and set the stage for the standard big bang model. Any matter or radiation present at the onset of exponential expansion is rapidly diluted away, leaving only the vacuum energy ρ_v, which remains constant throughout the expansion. One needs an account of how the universe is re-populated with normal matter and radiation after the stage of exponential expansion.

Guth (1981) launched a research program not by solving either of these problems but by making a compelling case in favor of inflation. Guth recognized that a stage of exponential expansion solves two fine-tuning problems of the standard model, the flatness and horizon problems.[22] On this basis he argued that the idea was worth pursuing despite his failure to give an account of the transition to the standard model. The source of exponential expansion in his original account was the vacuum energy of the Higgs field in a proposed Grand Unified Theory (GUT) trapped in a false minima during a first-order phase transition. Even though this solution of the source problem would not survive long, by contrast with earlier proposals Guth had shown how to link the idea of inflation with an active area of

[20] de Sitter spacetime is a solution to EFE with a stress energy tensor given by $T_{ab} = -\rho_v g_{ab}$, where ρ_v is the vacuum energy density. The scale factor then expands exponentially, with $\chi^2 = \frac{8\pi}{3}\rho_v$. During inflation the stress energy tensor has approximately this form. Given that the vacuum energy density remains constant during the expansion while "ordinary" matter and energy is rapidly diluted, the vacuum energy dominates the expansion and the solution, roughly speaking, approaches de Sitter spacetime.

[21] The stress-energy tensor stated in the previous footnote does not satisfy the strong energy condition formulated in footnote 16; the fact that the vacuum energy density does not dilute with expansion reflects this. A stress-energy tensor that violates this condition is a necessary condition for exponential expansion within classical GR.

[22] Guth discovered inflation in connection with a third problem, the monopole problem. Grand unified theories from the late 70s predicted the existence of magnetic monopoles, and the relic abundance of the monopoles would be many orders of magnitude greater than the observed energy density of the universe. See Guth (1997a) for his account of how he discovered inflation.

research in particle physics. In effect, inflation exchanged various large-scale properties of the universe previously treated as initial conditions for features of the dynamical evolution of a scalar field in the early universe. This exchange was soon exploited in giving a solution to the transition problem and in giving an account of the origins of the seeds for structure formation. After reviewing Guth's case and critical responses to it, we will turn to the discovery of the inflationary account of structure formation at the Nuffield Workshop and briefly discuss the account itself in more detail.

3.1 Inflation and Initial Conditions

Guth (1981) identified two problems for the standard big bang model that inflation solved:

> The standard model of hot big-bang cosmology requires initial conditions which are problematic in two ways: (1) The early universe is assumed to be highly homogeneous, in spite of the fact that separated regions were causally disconnected (horizon problem) and (2) the initial value of the Hubble constant must be fine tuned to extraordinary accuracy ... (flatness problem). (p. 347)

A better label for the first problem is the "uniformity problem": there is an apparent conflict between the strikingly uniform temperature of the CMBR and the presence of particle horizons in the standard FLRW models. We have seen above that cosmologists working on structure formation had noted various puzzles raised by the presence of horizons, and Misner (1969) formulated the problem in terms similar to Guth's a decade earlier.[23] Suppose we take the Planck time as an appropriate boundary of GR's domain of applicability. Because of the huge difference between the particle horizon and the size of the observed universe, the universe at that time consists of 10^{83} causally disconnected regions. The fact that all of these regions have the same physical properties (such as the same temperature to one part in 10^5) cannot be explained in the standard model via causal interactions, but instead has to be treated as a posit. Including a stage of exponential expansion stretches the horizon length; for N "e-foldings" of expansion the horizon length d_h is multiplied by e^N. For $N > 65$ the horizon distance, while still finite, encompasses the observed universe; the observed universe could then have evolved, with an inflationary stage, from a single causal patch rather than 10^{83} patches with an astonishing degree of pre-established harmony.

What Guth called the "flatness problem" was not widely discussed in the literature prior to inflation.[24] The problem arises from the following feature of the FLRW dynamics. We can write the fundamental equation for the FLRW models (the Friedmann equation) in terms of the density parameter Ω, which represents the total energy density and is defined as the ratio $\Omega =: \frac{\rho}{\rho_c}$, as follows:[25]

$$\frac{|\Omega - 1|}{\Omega} \propto R^{3\gamma-2}(t), \tag{2}$$

[23] See Smeenk (2005) for a discussion of how these two features of the FLRW models were treated prior to Guth's identification of them as problems to be solved by inflation.

[24] Guth learned of the problem from lectures given by Robert Dicke (Guth 1997a). See Dicke (1969) and Dicke and Peebles (1979) for Dicke's formulation of the problem.

[25] The critical density is the value of ρ for the flat FLRW model, $\rho_c = \frac{3}{8\pi}\left(H^2 - \frac{\Lambda}{3}\right)$, where $H = \frac{\dot{R(t)}}{R(t)}$ is the Hubble "constant" and Λ is the cosmological constant. Note that the problem is sometimes reformulated, equivalently, as the question of why the age of the universe is so large compared to the time scales relevant to fundamental physics.

with $\gamma > 2/3$ for normal matter.[26] Provided that $\gamma > 2/3$, if the value of Ω differs from 1 it evolves rapidly away from 1; the value $\Omega = 1$ is an unstable fixed point under dynamical evolution. It is thus particularly surprising that observations indicate that Ω is still close to 1. If we imagine choosing an initial value $\Omega(t_i)$ at some early time, it must be *incredibly* close to 1.[27] (This is called the "flatness" problem because $\Omega = 1$ corresponds to the "flat" FLRW model, with zero spatial curvature.) By contrast, during inflation $\gamma = 0$ and the density parameter is driven *towards* one. An inflationary stage long enough to solve the horizon problem drives a large range of pre-inflationary values of $\Omega(t_i)$ close enough to 1 by the end of inflation to be compatible with observations that constrain the current value, Ω_0, to be close to 1.

Inflation appears to eliminate the need for special initial conditions in the standard big bang model. Before turning to the inflationary account of structure formation, we will briefly consider debates about whether this should be taken as evidence in favor of inflation. The standard model requires positing the same physical conditions in 10^{83} causally disconnected patches and delicately tuning the overall energy density very close to that of the "flat" FLRW model. By contrast, with inflation it appears that "anything goes," in the sense that a "generic" lumpy, non-uniform initial state produces a uniform, flat region large enough to encompass the observed universe. A word of caution is in order: it is not the case that inflation *eliminates* dependence on initial conditions entirely. 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 $\Omega_0 = 1$. Inflation *enlarges* the range of initial conditions compatible with observations.[28]

Guth's approach is an example of a general strategy: given a theory that apparently requires special initial conditions, augment the theory with new dynamics such that the dependence on special initial conditions is reduced. Introducing an inflationary stage eases an apparent conflict between a "natural" or "generic" initial state and the observed universe, in the following sense. Suppose that we imagine choosing a cosmological model at random from among the space of solutions of EFE. Even without a good understanding of this space of solutions or how one is chosen to be "actualized," it seems clear that one of the maximally symmetric FLRW models must be an incredibly "improbable" choice.[29] New dynamics in the form of inflation makes it possible for "generic" pre-inflationary initial conditions to evolve into the uniform, flat state required by the standard model. McMullin (1993) describes a preference for this approach as accepting an "indifference principle,"

[26]The equation of state of a perfect fluid is $p = (\gamma - 1)\rho$, where p is the pressure, ρ the density, and the index γ is used to classify different types of fluids. For radiation, $\gamma = 4/3$ and for "dust" $\gamma = 1$ (corresponding to zero pressure). "Normal" matter satisfies the energy conditions defined in footnote 16.

[27]Guth (1981) calculates the value for the Planck time, $t_p = 10^{-43}$ seconds: $|\Omega(t_p) - 1| < 10^{-59}$.

[28]This point was first made in response to Misner's "chaotic cosmology," which like inflation proposed new dynamics (in Misner's case, damping of anistropies due to neutrino viscosity) in order to insure that an isotropic universe emerges from a large range of anisotropic initial conditions. In response to Misner, Collins and Stewart (1971) showed that one can always pick an arbitrarily large anisotropy at a given time t_0 and find a solution of the relevant system of equations as long as there are no processes which could prevent arbitrarily large anisotropies at some $t_i < t_0$. A similar criticism applies to inflation, as Madsen and Ellis (1988) have emphasized. Guth (1997b) has acknowledged 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).

[29]For any reasonable choice of measure over the space of solutions, these models are presumably a measure-zero subset. There are some results supporting this claim. In particular, models lacking symmetry form a dense, open subset of the space of solutions to EFE; see Isenberg and Marsden 1982 .

which states that a theory that is indifferent to the initial state, i.e. robust under changes of it, is preferable to one that requires special initial conditions. Theorists who accept the indifference principle can identify fruitful problems by considering the contrast between "natural" initial states and the observed universe, given that any difference between the two should be reconciled via new dynamics.

An overwhelming majority of cosmologists found Guth's arguments persuasive. This line of reasoning is frequently endorsed as a motivation for inflation in the huge literature on the topic following Guth's paper. However, there are significant challenges to this case for inflation, initially raised in the early days of inflation and developed further by a number of critics. A first line of criticism focuses on the applicability of the dynamical approach to cosmology. One way of supporting the demand for new dynamics applicable in everyday situations does not carry over to cosmology. For a normal experimental system, it is possible to check, at least in principle, whether a large variety of initial states lead to the same final state; if so, there is evidence in favor of robust dynamics that reduces dependence on the initial state. Obviously such supporting evidence cannot be gathered in cosmology. The more fundamental question regards how to treat the initial state of the system, when the system is the universe. In the ordinary case this can be treated as fixed by the experimenter, with some variation in the initial state with each repetition of the experiment. But in cosmology it is not clear why we should suppose that the initial state of the universe is "chosen at random" from among the states compatible with the laws of physics, as the dynamical approach assumes.

There are two puzzling aspects of this picture. What laws should be considered in defining the space of possible initial states? One might extrapolate classical GR to define the space of possible states, but this requires extending past GR's expected domain of applicability. Problems defined based on this extrapolation are then vulnerable to being undermined. For example, if a full theory of quantum gravity incorporates a symmetry principle that limits allowed initial states to the FLRW models, there would be no uniformity problem left over for new dynamics such as inflation to solve. This feature of the problems targeted by the dynamical approach contrasts with the case of, for example, empirical anomalies as a target for theory development. Empirical anomalies may not wear their implications on their sleeves, but their existence is stable under the introduction of new theories. The fact that Newtonian theory failed to fully account for the anomalous advance of Mercury's perihelion did not change with the introduction of general relativity. By contrast, there is not such a clear conflict between contemporary observations and the standard model of cosmology; the conflict is instead between the standard model supplemented with a particular understanding of the initial state based on an educated guess regarding the full theory governing the initial state. Second, how should we make sense of the implicit probability judgments employed in these arguments? The assessment of an initial state as "generic," or, on the other hand, "special," is based on a choice of measure over the allowed initial states of the system. But on what grounds is one measure to be chosen over another? Furthermore, how does a chosen measure relate to the probability assigned to "actualization" of the initial state? It is clear that the usual way of rationalizing measures in statistical mechanics, such as appeals to ergodicity, do not apply in this case because the state of the universe does not "sample" the allowed phase space.

Critics of the dynamical approach typically advocate an alternative viewpoint, namely that the initial state of the universe should be taken as "special" rather than "generic" in any case. One motivation for this approach is the Boltzmannian explanation of time's arrow, which requires that the initial state of the universe was extremely special in the

sense of having low entropy. The dynamical approach seems to ignore Boltzmann's insight by supposing that the initial state is "generic." This approach shares the difficulty just mentioned of making sense of probability judgments regarding the initial state, or providing a definition of entropy applicable to the case at hand. But setting this problem aside, the methodology is diametrically opposed to the dynamical approach: rather than introducing a subsequent stage of dynamical evolution that erases the imprint of the initial state, on this alternative view the goal is to formulate a "theory of initial conditions" that accounts for its special features.

Setting aside this clash of methodologies, critics of inflation have further argued that inflation fails even by the standards of the dynamical approach. The question is whether inflation can deliver on the promise of producing a uniform, flat post-inflationary state from "generic" pre-inflationary conditions. A general line of argument originally due to Penrose (1986) suggests that the probability of inflation must itself be quite low.[30] Suppose that we are given a generic state in a universe that leads to a "big crunch" singularity in the future. It seems overwhelmingly unlikely that as the universe approaches the final singularity, it will "deflate" by converting all the gravitational energy of the collapsing matter into kinetic energy of a scalar field in just the right way to push it into a false vacuum state. But this is simply the time reverse of the account of inflation, and — on the assumption that the dynamics is time-reversal invariant — the argument concludes that our assessment that the probability of deflation is low should also apply to inflation itself. This raises the worry that inflation eliminates the dependence on special initial conditions in one sense, namely the degrees of freedom of the gravitational field, only by introducing special initial conditions for the field that drives the inflationary expansion. And there are indications that this is the case independent of Penrose's argument. Vachaspati and Trodden (1999) proved that the field driving inflation must be uniform over a region larger than the Hubble radius in order to trigger inflation.[31] In addition, a scalar field has to satisfy further constraints in order to drive exponential expansion during inflation.[32]

These critical responses to the dynamical approach Guth adopted seem to have had little impact on the development of the field. This is, I will argue, in part because of the answer that was shortly developed to a problem that Guth did see as a clear obstacle to the idea of inflation. Guth noted the advantages of inflation while at the same time admitting that his model failed to solve the transition problem (also called the graceful exit problem). Rather than smoothly joining onto the FLRW expansion, the phase transition Guth considered ended via bubble nucleation, leaving the early universe marred with non-uniformities. The model failed to achieve the delicate balance between overall uniformity and slight perturbations required for the account of structure formation via gravitational instability. At first blush, as Barrow and Turner (1981) noted, provided that bubble nucleation could be avoided, inflation may actually exacerbate the problem by too efficiently smoothing out the

[30] Penrose's original argument was quite brief, but he has discussed it further in Penrose (1989, 2004). This line of argument is also discussed and developed further in Unruh (1997); Earman and Mosterin (1999); Hollands and Wald (2002b,a).

[31] Their proof assumes the weak energy condition, trivial topology, and classical EFE.

[32] The stress energy tensor for a scalar field is given by

$$T_{ab} = \nabla_a \phi \nabla_b \phi - \frac{1}{2} g_{ab} \left(g^{cd} \nabla_c \nabla_d \phi - V(\phi) \right); \tag{3}$$

inflation requires that the field is "potential-dominated" in the sense that the field is sufficiently uniform that the derivative terms are negligible, $V(\phi) >> g^{cd}\nabla_c\nabla_d\phi$. If this condition holds, $T_{ab} \approx -V(\phi)g_{ab}$ as required to produce exponential expansion.

universe, leaving it without wrinkles to seed later structures. We will see shortly how this initial worry developed into an important success as theorists discovered a mechanism for generating perturbations during inflation.

3.2 New Inflation and the Nuffield Workshop

Guth's paper and talks based on it introduced many astrophysicists and particle physicists to the very idea of early universe cosmology. By admitting the flaws of his initial model, Guth also left his readers and audiences with a project: to find a working model of inflation. Paul Steinhardt, then a Junior Fellow in the Harvard Society of Fellows, exemplifies this reaction; he described 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 flatness, horizon, and monopole problems, Guth ended by explaining the fatal flaw of his initial model. Steinhardt recalls his reaction (Steinhardt 2002): "Here was this great idea and it just died right there on the table. So I couldn't let that happen."

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 led Andrei Linde in Moscow, a pioneer in the study of early universe phase transitions throughout the 70s, to reconsider the possibility of a first-order phase transition. Linde had considered the idea in collaboration with Chibisov, but had dismissed it as unworthy of publication — "there was no reason to publish such garbage" — due to the problem of inhomogeneities.[33] 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 than that used by Guth could solve the transition problem while providing sufficient inflation to solve the horizon and flatness problems (Albrecht and Steinhardt 1982; Linde 1982). Their proposal is usually called "new inflation."[34]

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. (The Lagrangian density for a classical scalar field is given by $\mathcal{L} = \frac{1}{2}\partial_\mu\phi\partial^\mu\phi - V(\phi)$, where $V(\phi)$ is the potential. The effective potential includes quantum corrections to the classical potential.)[35] This change leads to a dramatically different phase transition. Most importantly, inflation continues after the formation of an initial bubble: rather than tunnelling directly

[33]The collaborative work with Chibisov is mentioned in Linde (1979), pp. 433-34; the quotation is from a 1987 interview (Lightman and Brawer 1990, 485-86).

[34]At roughly the same time, Stephen Hawking and Ian Moss proposed an alternative solution to the transition problem. Although Hawking and Moss (1982) is sometimes cited as a third independent discovery of new inflation, it differs substantially from the other proposals. 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 down the hill" (for an appropriate values of parameters in the effective potential), producing an inflationary stage long enough to match Guth's success.

[35]See, e.g., Coleman (1985), Chapter 5 for an introduction to the effective potential, and Kolb and Turner (1990) for a detailed discussion of the differences between old and new inflation.

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. 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 nonconserving 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:

$$\frac{d^2\phi}{dt^2} + 3H\frac{d\phi}{dt} + \Gamma_\phi \frac{d\phi}{dt} + \frac{dV(\phi)}{d\phi} = 0, \tag{4}$$

where t is the time coordinate in the FLRW model, and Γ_ϕ is the decay width of ϕ.[36] 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 that the Γ_ϕ term is negligible, then the evolution of ϕ can be approximately described by:

$$3H\dot{\phi} \approx -\frac{dV(\phi)}{d\phi}. \tag{5}$$

(The name is due to the similarity between the evolution of ϕ and that of a ball rolling down a hill, slowed by friction.) During slow roll the potential energy $V(\phi)$ dominates over the kinetic energy $\frac{\dot{\phi}}{2}$, 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 the process of 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. The reheating stage is necessary to "repopulate" the universe, given that any pre-existing matter or radiation is rapidly diluted during the inflationary expansion.

By the spring of 1982 several groups were at work fleshing out the details of the new inflationary scenario: a 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, 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

[36] 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.

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.

The first international conference focusing on "very early universe cosmology ($t < 1$ sec)" convened in Cambridge from June 21 - July 9, 1982.[37] 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 Δ, 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 $\Delta \approx 10^{-4}$, whereas Steinhardt and Turner initially estimated a magnitude of 10^{-16}. 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.

The calculations drew on an idea introduced prior to Guth's paper. Mukhanov and Chibisov (1981) had argued that a de Sitter phase could generate perturbations by "stretching" zero-point fluctuations of quantum fields to significant scales. This idea would become the basis for the generation of seed perturbations in inflationary cosmology. The details were worked out at the Nuffield Workshop, which seems to be a rare example of a scientific workshop that fulfilled the goal of bringing together the relevant research groups and successfully forging a consensus on an important problem.

Prior to the workshop, Hawking circulated a preprint which 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" of the scalar field. Hawking's preprint claimed that this results in a scale-invariant spectrum of adiabatic perturbations with $\Delta \approx 10^{-4}$, 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 per-

[37]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.

turbations with length scales greater than the Hubble radius.[38] 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 inflation since they scale as $R(t) \propto e^{Ht}$, but the Hubble radius remains fixed since H is approximately constant during the slow roll phase of inflation. 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.[39] 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 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 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.

Here I will not give a more detailed comparison of these three approaches. Despite the conflicting assumptions and other differences, the participants of the Nuffield workshop apparently lent greater credibility to their conclusions due to the rough agreement between the three different approaches. 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:

$$\Delta|_\lambda = AH\frac{\Delta\phi}{\dot{\phi}}, \tag{6}$$

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

[38] 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 $ds^2 = R^2(t)(dt^2 - h_{ij}dx^i dx^j)$, with $i, j = 1, 2, 3$. The coordinates break down if the geodesics of co-moving observers cross.

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

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 5) $\dot{\phi} \approx \frac{V'(\phi)}{3H}$, and V' depends on the coupling constants appearing in the effective potential. For a Coleman-Weinberg effective potential with "natural" coupling constants, $\dot{\phi} < H^2$; plugging this all back into the initial equation we have:

$$\Delta|_\lambda > A\frac{H^2}{2\pi H^2} \approx .1 - 1 \tag{7}$$

Inflation naturally leads to an *almost* 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 self-coupling for the Higgs field apparently needs to be on the order of 10^{-8}, although a "natural" value would be on the order of 1.[40]

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 an $SU(5)$ GUT. However, it was clear how to develop a "newer 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. Finding a particular particle physics candidate for the scalar field driving inflation would provide for an important independent line of evidence. The Nuffield workshop marked the start of a different approach, as the focus shifted to implementing inflation successfully rather than starting with a candidate for the field driving inflation derived from 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.[41] The inflaton may resemble the Higgs, but the rules of the game have changed: it is a new fundamental field distinct from any scalar field appearing in particle physics. The fact that inflation has not been closely tied to $SU(5)$ GUTs has been a boon to the field. 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). Following the demise of the minimal GUTs, there has been an ongoing effort to implement inflation within new models provided by particle physics.

Following the Nuffield workshop, inflation turned into a "paradigm without a theory," borrowing 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 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

[40] 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).

[41] 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.

> inflaton and nothing else can only be viewed as a toy, not a theory. (Olive 1990, p. 389)

In a similar vein, Dennis Sciama commented 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 nearly 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 as comic relief in conference talks.[42]

3.3 Horizon Crossing

The research throughout the 80s devoted to exploring different ways of implementing inflation led to a proliferation of models rather than consolidation around a single canonical model of inflation. Despite this lack of consensus regarding how inflation relates to particle physics, consensus was achieved regarding the consequences of inflation for structure formation.[43] Inflation provides one natural way to produce the appropriate seed perturbations to produce galaxies and other large-scale structures via hierarchical clustering.

The basic physical mechanism for producing density perturbations in inflation results from a feature called "horizon exit / re-entry." Start with a massless, minimally coupled scalar field ϕ evolving in a background FLRW model. Due to the symmetry of the FLRW models the Fourier modes ϕ_k of ϕ are uncoupled, and each mode evolves during slow-roll inflation according to the following equation:[44]

$$\frac{d^2\phi_k}{dt^2} + 3H\frac{d\phi_k}{dt} + \frac{k^2}{R^2}\phi_k = 0, \tag{8}$$

This equation is just that of a harmonic oscillator with a damping term. For modes such that $\frac{k}{R} << H$, the damping term is negligible, whereas those with $\frac{k}{R} >> H$ will evolve like an over-damped oscillator and "freeze in" with a fixed amplitude. The inflationary account runs very roughly as follows. One starts with a vacuum state: all the modes ϕ_k are assumed to be in their ground state prior to inflation. For $\frac{k}{R} << H$ the modes evolve adiabatically, remaining in their ground states, given that eqn. (8) is approximately the equation for a harmonic oscillator. This account is not sensitive to exactly when a given mode is assumed to be "born" in its ground state. During inflation the modes scale with the exponential expansion whereas H is approximately constant. Due to this scaling behavior, modes will reach the horizon scale $\frac{k}{R} \approx H$ — "horizon exit". The damping term in eqn. (8) is no longer negligible and the modes "freeze in" as they cross the horizon. Modes then "re-enter" the horizon later given that the Hubble radius grows more rapidly than the modes after the inflationary stage has ended. Finally, these modes are treated as "classical" density perturbations upon re-entering the horizon.[45] This evolution leads to a nearly

[42] Rocky Kolb used such a slide in a talk at the Pritzker Symposium (Chicago, 1998); for an example of such a list see **?**, figure 41.3.

[43] The most comprehensive, and widely-cited review article, regarding structure formation and the implications of inflation is Mukhanov et al. (1992).

[44] This equation can be derived from the action for the scalar field minimally coupled to gravity. I am neglecting various details, including the need to include metric perturbations along with the perturbations of the scalar field.

[45] Although I do not have space to discuss the issue further here, this step involves a quantum to classical transition. The justification for treating quantum fluctuations as .

scale invariant spectrum. The spectrum is not *exactly* scale invariant because the Hubble radius is not truly constant throughout inflation. As described above, the amplitude of the perturbations that are frozen in at horizon exit depends upon the details of the particular inflationary model under consideration.

The basic features of this account were established in the early days of inflationary cosmology, as described in the previous section. But in the 90s there was a new impetus to understand the consequences of inflation in further detail, in terms of both generic consequences of inflation and specific signatures for particular models of inflation. The remarkable success of the COBE satellite measurements made it clear that precision observations of the microwave sky could provide an increasingly precise probe of the physics of structure formation. This raised the possibility of empirically deciding between inflation and other competing ideas regarding the origins of structure, such as topological defect theory, and, more ambitiously, of determining features of the field (or fields) driving inflation. Lyth and Riotto (1998), for example, comment that the COBE data spurred a "renaissance" of inflationary model building.

4 Topological Defects

Scientific theories rarely develop in isolation, and rival theories shape the assessment of important problems and help to set the research agenda. Throughout the 80s and 90s the most important alternative account of the origins of structure was based on topological defects. These ideas were first studied in the 70s prior to inflation. Unlike inflation, which is the consequence of a specific type of phase transition, topological defects are a nearly generic consequence of phase transitions. This brief discussion will focus on the contrasts between the two theories, with no attempt to give a a detailed account of the historical development of these ideas.[46]

The formation of topological defects is determined by properties of the vacuum manifold $\mathcal{M}$. The vacuum manifold consists of the degenerate vacuum states of the system after the phase transition. Suppose the theory initially has a symmetry group G that is then spontaneously broken to a subgroup H.[47] The symmetry is broken in the sense that the vacuum states of the theory are degenerate: although the vacuum state is not invariant under the action of some $g \in G$, these distinct vacuum states are degenerate in that the Hamiltonian has the same eigenvalue. The subgroup H consists of those elements of G under which the vacuum state remains invariant. The space of degenerate vacuum states is then in one-to-one correspondence with sets of elements of the form gH; in other words, the vacuum manifold $\mathcal{M}$ is topologically equivalent to the quotient space G/H. Topological features of the vacuum manifold then determine what kinds of topological defects may form

[46] I have left aside one important aspect of the comparison between inflation and topological defect theories, namely the role of different types of dark matter in each scenario. The mechanisms for structure formation are part of package deal, including assumptions about the overall matter budget and other factors more significant for later stages of structure formation.

[47] This means that, roughly speaking, for all $g \in G$ the Hamiltonian of the system is invariant under the action of g, but the vacuum or ground state of the system is not. (This is only a rough gloss; in quantum mechanics the action of a symmetry g is usually represented by a unitary operator on the Hilbert space, but in the case of broken symmetry there is not a well-defined operator mapping between degenerate vacua, as these each define different Hilbert spaces.) The degenerate vacuum states are labeled by different values of the "order parameter" of the transition. The order parameter is the thermodynamic quantity that changes discontinuously through the transition and characterizes different phases, corresponding to degenerate vacua in this case. In this case the order parameter is the vacuum expectation value of the relevant field(s).

in the course of the phase transition.[48]

Starting in the early 70s these ideas were applied to cosmology. Extrapolating the FLRW models, the early universe reaches arbitrarily high temperatures at early times. Kirzhnits (1972) suggested that symmetries in particle physics would be restored at sufficiently high temperatures, by analogy with symmetry restoration in condensed matter systems. Further calculations of symmetry restoration in the Standard Model of particle physics supported the idea that as the universe cooled it passed through a series of phase transitions that broke the symmetries between various interactions.[49] Many symmetry breaking phase transitions in condensed matter systems lead to the formation of topological defects, such as the formation of vortices in liquid helium, so it is natural to expect that defects may have formed in early universe phase transitions.

In a seminal paper, Kibble (1976) argued that topological defects would be produced due to the horizon structure of the early universe. (His account is sometimes referred to as the "Kibble mechanism.") Given that the correlation length of the order parameter is bounded by the horizon distance, the phase transition produces domains in which the order parameter takes on different values determined by random fluctuations, assuming that the dynamics is not completely adiabatic. Whether defects form depends on the topology of the vacuum manifold. For example, suppose that there is a curve through $\mathcal{M}$ that cannot be smoothly contracted to a point. Each point within the space $\mathcal{M}$ represents a different degenerate vacuum state, which is labeled by different values of the order parameter for the phase transition. Suppose that the values of the order parameter around a spatial loop take the same values given along the loop in $\mathcal{M}$. Given that the loop cannot be continuously contracted to a point within $\mathcal{M}$, it is also not possible to assign values of the order parameter continuously in the region bounded by the spatial loop while remaining in $\mathcal{M}$. This implies that there must be a "defect," namely a region of space in which the fields cannot reach the vacuum state and instead remain trapped in a state of higher energy. The nature of these regions of higher energy is fixed by the structure of $\mathcal{M}$. In the case at hand, with a non-simply connected vacuum manifold, the phase transition leads to two-dimensional defects called "cosmic strings." There are several other possibilities. A phase transition breaking a *discrete* symmetry leads to regions in which the order parameter takes on discrete values separated by domain walls, which are three-dimensional surfaces in spacetime. If the vacuum manifold has non-contractible two-spheres rather than circles, then the phase transition produces point-like defects (such as magnetic monopoles); for non-contractible three-spheres the corresponding zero-dimensional defects are called "textures," event-like defects that do not have a stable localized core.[50]

Early studies showed that domain walls and some types of monopoles had disastrous consequences, conflicting with observational constraints by several orders of magnitude. However, other types of defects — in particular, cosmic strings — were more plausible candidates for the seeds for structure formation. The defects are inherently stable regions of higher energy density, whose scale is set by the energy scale of the phase transition.

[48] The relevant structure is given by the homotopy groups of the space. For further discussion, see, e.g., Vilenkin and Shellard (2000).

[49] Perhaps brief comment re. worries related to phase transitions in the early universe...

[50] Additional types of defects arise due to the distinction between gauge and global symmetries and the possibility of "hybrid" defects. Defects formed in a transition breaking a global symmetry tend to have energy density distributed throughout a region, whereas those formed by gauge symmetry breaking are more localized. Hybrid defects are produced by a series of phase transitions, leaving an interacting network of defects of different kinds. See, e.g., Vilenkin and Shellard (2000), for further discussion.

The defects have an important impact on the dynamical evolution of the system following the phase transition, and in particular it is plausible that they will provide seeds that are subsequently enhanced via gravitational instability as described by linear perturbation theory. For GUT-scale phase transitions the energy density is the appropriate order of magnitude to seed large-scale structure. Some defect theories have "scaling solutions," in which the network of defects evolves such that there is no preferred length scale imprinted at a particular time. These theories then pass an important initial test, in that they lead to an approximately scale-invariant HPZ spectrum of perturbations.[51] They are thus compatible with the first generation of CMBR observations and the general picture of structure formation described above. However, there are important general differences between the inflationary account and that provided by topological defects that were clarified by a substantial research effort throughout the 80s and 90s.

To determine whether topological defects suffice as the primary mechanism for producing seeds for structure formation, researchers had to tackle two challenging problems. The first is to describe the phase transition itself and determine the nature of the defects produced, with sufficient quantitative detail to determine the consequences for later stages of evolution. In principle these details should be calculable given a particular extension of the Standard Model of particle physics. But the shear complexity of the models, and the nature of the quantities needed to assess the implications for structure formation, have made it quite difficult in practice to carry out such calculations. Second, one has to describe the subsequent evolution of the network of defects left over following the phase transition over a wide range of dynamical scales. Solving this second problem requires determining the interactions among the defects and their gravitational effects. The problem is exceedingly difficult because the evolution of defects is non-linear, and researchers have relied primarily on numerical simulations. Physically plausible suggestions regarding evolution of defects have often been undercut by numerical work. Throughout the 80s, for example, the general picture of how strings seeded galaxy formation changed considerably in light of numerical simulations establishing details regarding the size of typical closed loops of strings and the behavior of open strings.[52] These two problems are exacerbated by uncertainty regarding the relevant fundamental physics. The details of the phase transitions depend on specific features of the physics — for example, the vacuum manifold is fixed by the full symmetry group G and its unbroken subgroup H, but these differ among proposed extensions of the Standard Model.

Despite these difficulties, by around 1997 there was a consensus regarding the generic consequences of structure formation via defects and the contrast with the consequences of inflation.[53] Structure formation via topological defects is "active" in the sense that the network of defects persists over time and continues to interact gravitationally with the other constituents. More precisely, in the evolution equation for perturbations of the cosmological model there is a source term, representing the stress-energy of the network of defects. Determining the evolution of the perturbations thus requires calculating the evolution of this source term, based on the non-linear dynamics of the network of defects. By way of

[51] However, the sense in which the two theories are scale-invariant is different; see, e.g., §5.1.1 of Martin and Brandenberger (2001). Many defect models are scale-invariant only over a limited dynamical range; for example, in models of defect formation via strings scale invariance is broken at the matter-radiation transition.

[52] See Vilenkin and Shellard (2000, Chapter 11) for an overview; the closing section (p. 342) emphasizes the changes in the account due to numerical simulations of the evolution of string networks.

[53] References...

contrast, inflation is a "passive" account of structure formation: there is no source term in the evolution equation, and in the linear regime the solution is fixed by the initial conditions. Roughly speaking, in inflation the perturbations evolve "on their own" after being imprinted at early times, whereas in the defect theories the network of defects persists and continues to seed structure formation. In addition, perturbations produced in defect theories "decohere" in the sense that fluctuations at all wave-numbers are not in phase. This is a consequence of the non-linear evolution of the source term, which leads to mixing of perturbations across different modes. The perturbations are also non-Gaussian due to the correlations that this mixing produces between perturbations. Finally, defects generate scalar, vector, and tensor perturbations of roughly equal magnitude, whereas inflation eliminates vector perturbations and produces tensor perturbations that are much smaller than the scalar perturbations.

These general features lead to a observational signature on the CMBR that differs quite strikingly from that produced by inflation. In particular, defect theories predict that there will not be strong secondary oscillations evident in CMBR observations. These features are "washed out" due to decoherence, whereas by contrast in inflationary accounts there are coherent standing wave oscillations in the baryon density that lead to strong secondary peaks. The position of the first peak also differs in inflation and topological defect models, with defect models generally predicting a primary peak at a larger multipole moment ($\ell \geq 300$) than inflation ($\ell \approx 200$).[54] Observational results starting in the late 90s and culminating in the WMAP results (3 year results published in 2003) provide decisive support for inflation with respect to both of these features.

In addition to the physical contrast between the mechanisms for structure formation, there are important methodological contrasts between the two approaches. First, despite uncertainty regarding the detailed physics of the phase transitions, the account of structure formation via defects is constrained enough by general theoretical principles to produce specific observational signatures. Physicists working on defect theory often highlighted this rigidity as a virtue of the theory, characterizing it as "falsifiable" in a vaguely Popperian sense. Second, topological defect theory did not address the problems related to initial conditions highlighted by Guth. In effect, the theory starts from the same initial conditions as the standard FLRW models, with the exception that the initial seed perturbations were produced dynamically rather than fixed by hand. One might expect criticisms of topological defect theory based on its failure to solve these problems. However, that expectation assumes that it is accurate to treat these accounts as incompatible rivals.

Topological defect theory and inflation are not straightforwardly rival theories. Both accounts are based on the idea of early universe phase transitions, and there are models incorporating an inflationary stage during one phase transition and defect formation in a later transition. Defect theorists often emphasized that inflation could still be invoked to solve the problems related to initial conditions (see, e.g. Vilenkin and Shellard 2000), as long as inflation set the stage for a phase transition at the appropriate energy scale. The two accounts are rivals if they are taken as the primary mechanism for the formation of structure. But there is no theoretical reason to insist on one mechanism to the exclusion of the other, and several models have been constructed which combine aspects of inflation and defects.

[54]The angular power spectrum characterizes the variations in temperature of the CMBR, i.e. the amount of temperature variation across different points of the sky versus the angular frequency ℓ. Small values of ℓ correspond to temperature variations with a large angular scale. See, e.g., Liddle and Lyth (2000), §5.2, for a definition and further discussion of the angular power spectrum.

5 Characterizing Empirical Success

The fate of topological defect theory illustrates the power of contemporary cosmological observations. Within the last 50 years cosmology has gone from being a field with only "2 1/2 facts"[55] to a field with data that is sufficiently rich to warrant conclusions regarding novel physics far beyond the reach of earthbound accelerators. By the turn of the millenium, topological defect theory was not only falsifiable but apparently falsified by cosmological observations.[56] Inflation did not share this fate; it is clearly compatible with the CMBR observations that ruled out defect theories. But in what sense does current observational data support inflationary theory? How should we characterize the empirical success or predictive power of the theory?

Debates in the physics literature have typically framed this question in terms of falsifiability. Has inflation avoided falsification just because it is unfalsifiable? Consider, for example, whether inflation could be falsified by finding that $\Omega_0 \not\approx 1$.[57] Flatness is often cited as an unambiguous, correct prediction of inflation. Guth (1997a), for example, emphasizes the extraordinary precision of the inflationary prediction – a correct value of Ω at the end of inflation to 15 significant figures! There are two reasons, however, to doubt that this is a clear prediction of inflation.[58] First, for *any* particular value of Ω_0 there is a corresponding "initial" value $\Omega(t_p)$, whether inflation occurred or not. Thus the prediction has to be regarded as a probabilistic claim: for "highly probable" or "reasonable" initial conditions inflation yields $\Omega_0 = 1$. But, as discussed in §3.1 above, it is not clear what to make of these probabilistic claims without a measure over the space of initial values of Ω. The second objection is that the inflationary paradigm is too flexible to yield falsifiable predictions. In the mid 90s theorists constructed "open models" of inflation that yield a lower value of Ω_0 (see, for example, Bucher et al. 1995). At most one might claim that a subset of inflationary models could be ruled out by finding $\Omega_0 \not\approx 1$, with further disagreement over whether this subset includes all of the "natural" or "reasonable" models of inflation. Rather than an unambiguous, falsifiable prediction, we are left with equivocal judgments regarding the probability assigned to initial conditions and the plausibility of different inflationary models.

Discussions of the falsifiability of inflation often draw Liddle and Lyth's distinction quoted in the introduction between inflation "as a theory of initial conditions" and inflation as a theory of structure formation.[59] The account of structure formation appears to have

[55] Peter Scheuer made this remark in the course of warning a student, Malcolm Longair, about the current status of cosmology in 1963; the list included (1) that the sky is dark at night, (2) that the galaxies recede, and (2 1/2) that the universe is evolving (qualified as a half fact due to its uncertainty).

[56] Observations seem to rule out topological defects as the primary mechanism for generating large-scale structure. However, defects might still play a role as part of the full account of the formation of structure or in other aspects of early universe cosmology, such as baryogenesis.

[57] The falsifiability of inflation, focusing in part on flatness, is addressed quite directly in a number of papers in Turok (1997), in particular the contributions by Linde, Steinhardt, Guth, and Albrecht. This has been a perennial subject of debate since the early days of inflation.

[58] The question was particularly pressing throughout the 90s, when the evidence seemed to favor open cosmological models with $\Omega_0 \approx 0.2 - 0.3$, although there was not a general consensus. See, e.g., Coles and Ellis (1997) for a detailed argument in favor of an open universe. However, the consensus had begun to shift in favor of a flat universe by 1998. Peebles and David Schramm were invited to convene a "great debate" on the issue in April of 1998. Due to Schramm's death the debate was rescheduled for October of 1998, with Michael Turner taking Schramm's place. But given that Peebles and Turner both agreed that the evidence decisively favored a flat universe, they changed the subject of the debate to "Is Cosmology Solved?" (Peebles 1999a; Turner 1999).

[59] The distinction is perhaps too quick, given that there are some predictions related to initial conditions.

definitive, falsifiable consequences. Several observational signatures — gaussianity, near scale invariance — follow directly from the description of the dynamical evolution of the modes of a quantum field through horizon-crossing. This dynamical mechanism for generating perturbations is a direct consequence of the defining feature shared by all inflationary models, given that it depends on the evolution of the Hubble constant during exponential expansion. Thus one might hope to avoid the above objections: the production of density perturbations is independent of assumptions regarding initial conditions, and the account is generic in the sense of being common to all models of inflation. But does the success of inflation simply exploit the malleability of the "inflaton" field and its potential? Note, for example, that the amplitude of the density perturbations needed for accounts of structure formation is used to constrain the parameters of the inflaton field. Peebles (1999b) classifies the amplitude of the density perturbations as a "diagnostic" rather than a successful prediction for this reason.

Hollands and Wald (2002b) have recently illustrated that there is not such a clear contrast in terms of initial conditions. In particular, the inflationary account of the dynamical evolution of the modes of a quantum field through horizon crossing assumes that the modes are initially in their ground state. This is a plausible assumption given that the modes with cosmologically significant length scales will be well inside the Hubble radius prior to the inflationary phase. Since the modes evolve adiabatically before horizon crossing the exact time at which they are taken to be "born" in their ground state is unimportant. Hollands and Wald (2002b) construct a simple model that produces a similar spectrum of density perturbations *without an inflationary phase* based on a different *Ansatz* for the initial conditions for these modes. Their model describes quantized sound waves in a perfect fluid, with the same "overdamping" of modes with $\lambda >> H^{-1}$ as in inflation. By contrast with inflation, there is no horizon crossing, so it is significant precisely when the modes are taken to be in a vacuum state. Hollands and Wald (2002b) propose to take the modes to be "born" in a ground state when their proper wavelength is equal to the Planck scale, motivated by considerations of the domain of applicability of semi-classical quantum gravity.[60] This hypothesis combined with the dynamics governing the evolution of the modes leads to a scale-invariant perturbation spectrum. The significance of this result for present purposes is that it undermines claims that the theory of structure formation does not depend on arguments regarding plausible initial conditions.

Stepping back from the details of inflation for a moment, it should be clear that there are important questions regarding both how to characterize a theory's empirical success and what a given degree of success establishes. It is unfortunate that these questions are still treated in the physics literature in terms of "falsifiability," and I will briefly sketch an alternative drawing on recent studies of Newton's methodology (Harper 2002; Smith 2002). On this approach, empirical success is defined in terms of the ability to determine consistent values of theoretical parameters from multiple, independent bodies of data. Consider, for example, Newton's argument in favor of a universal force of gravity in the *Principia*. Newton

For example, inflation predicts that the observed universe is topologically simply-connected; inflation is incompatible with compact topology at sub-horizon scales. Evidence that the universe is multiply connected would rule out inflation.

[60]The modes will be "born" at different times, continually "emerging out of the spacetime foam" (or whatever description the full theory of quantum gravity provides), with the modes relevant to large-scale structure born at times much earlier than the Planck time. By way of contrast, in the usual approach the modes at all length scales are specified to be in a ground state at a particular time, such as the Planck time. But the precise time at which one stipulates the field modes to be in a vacuum state does not matter given that the sub-horizon modes evolve adiabatically.

takes the theoretical framework provided by the laws of motion to be exact, and the array of mathematical results applying to forces in general then allows him to infer properties of the gravitational force from the observed motions of the planets, their satellites, and various other bodies (such as pendulums). The famous precession theorem is a particularly beautiful example: Newton shows that for approximately circular orbits, the motion of the apsides measures the exponent of the power law.[61] Taking the exponent in the power law for gravity as our example of a theoretical parameter, there are several lines of argument from diverse, independent bodies of data that fix the value as very close to -2. This account acknowledges that the theory requires some data as "input" to enable further predictions. Other bodies of data that can be used to constrain the same parameter value then provide independent checks. Harper (1990, 2007) argues that Newtonian characterization of empirical success is much more demanding than mere predictive accuracy. A theory that achieves predictive accuracy by "curve-fitting" (exploiting theoretical flexibility) will suffer by comparison with a more rigid theory on the Newtonian account.

The strength of Newton's empirical argument for universal gravitation is important in answering two objections. First, why should one accept gravity as a "real force" given that it apparently involved action-at-a-distance? Although the issue is complicated, Newton clearly held that the empirical case was sufficient to establish the reality of gravitational force despite uncertainty regarding its underlying cause and certainty that it is not a "mechanical" cause (i.e., due to contact action). Second, from a modern perspective, why should Newton's theory be preserved as a limiting case of general relativity? If we regarded the theory merely as a predictively accurate curve-fit, rather than an accurate systematic treatment of physical relationships within a limited domain, there would be no reason to expect general relativity to recover anything more than the predictions themselves. Speaking more generally, the first kind of objection relates to unresolved problems. In some cases the empirical success of a theory is sufficient to warrant acceptance even in light of open physical questions. The second challenge regards the use of a theory as a step towards further theories. Sufficient empirical success warrants preserving not just the predictions of the theory but the physical relationships it ascribes to systems within its domain.

Returning to the case of inflation, there are two similar challenges faced by inflation. First, there are various open problems regarding the place of an "inflaton" field within particle physics at the appropriate energy scales and the coupling of a scalar field to gravity. The cosmological constant problem is sometimes characterized as the Achilles heel of inflation. Inflation is built on the assumption that the false vacuum energy of the inflaton field couples to gravitation. But if this is so, the vacuum energy density of other quantum fields should contribute to gravity as an effective cosmological constant. A comparison between the vacuum energy density calculated in QFT and observational limits on the cosmological constant in GR reveals an incredible discrepancy of some 120 orders of magnitude! As Frank Wilczek commented in a review of the Nuffield workshop:

> It is surely an act of cosmic *chutzpah* to use this dismal theoretical failure [in understanding the cosmological constant] as a base for erecting theoretical su-

[61] The apsidal angle θ is the angle through which the radius vector rotates between two consecutive apsides, which are points on the orbit of maximum (aphelion) or minimum (perihelion) distance from the force center. Newton establishes (Book I, Proposition 45) that for approximately circular orbits under a centripetal force varying as $f \propto r^{n-3}$, the apsidal angle is given by $n = \left(\frac{\theta}{\pi}\right)^2$. For stable orbits, the radius vector rotates through π between the aphelion and perihelion, such that $n = 1$ and $f \propto r^{-2}$; and for nearly stable orbits, the force is approximately $f \propto r^{-2}$.

> perstructures, but of course this is exactly what is done in current inflationary models (Hawking et al. 1983, p. 476, original emphasis).

Second, cosmologists have often suggested that the requirement to find an inflaton field should serve as a constraint on particle physics. This is certainly appealing, as a successful case for inflation would provide a strong constraint at energy scales with few observational constraints from earthbound accelerators.

On this approach, the question to ask regarding inflation is not whether it makes various "falsifiable" predictions, but to what extent do the observational data allow us to infer the details of inflation? On the assumption that inflation is correct, what do the data allow us to infer about the inflaton field, and its effective potential $V(\phi)$? In these terms the account of inflation as a theory of structure formation provides a richer set of constraints on the theory. The solution of the horizon and flatness problems constrains the duration of the inflationary phase: the pre-inflationary patch has to grow larger than the observed universe, at a minimum. The inflationary stage will last sufficiently long if the potential $V(\phi)$ is suitably flat, and satisfies the "slow-roll" conditions described in §3.2 above. The account of structure formation, by contrast, provides more detailed constraints. The fluctuation modes that seed the formation of structure depend on the properties of the effective potential $V(\phi)$ at the time when they cross the horizon. (There is a limit on the part of the potential that can be constrained in this way, given that only some of the modes will have re-entered the horizon as observable density perturbations.) This opens up the prospect of reconstructing the inflaton potential based on observations of the CMBR. Whether the reconstruction provides sufficient empirical warrant to answer the challenges above is another question.

References

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

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. D. (1980). Galaxy formation - The first million years. *Royal Society of London Philosophical Transactions Series A*, 296:273–288.

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.

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.

Bucher, M., Goldhaber, A. S., and Turok, N. (1995). An open universe from inflation. *Phys. Rev.*, D52:3314–3337.

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

Coles, P. and Ellis, G. F. R. (1997). *Is the universe open or closed?* Cambridge University Press, Cambridge.

Collins, C. B. and Stewart, J. M. (1971). Qualitative cosmology. *Monthly Notices of the Royal Astronomical Society*, 153:419–434.

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.

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

Ellis, G. F. R. and Rothman (1993). Lost horizons. *American Journal of Physics*, 61(10):883–893.

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.

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. H. and Pi, S. Y. (1982). Fluctuations in the new inflationary universe. *Physical Review Letters*, 49:1110–1113.

Harper, W. (1990). Newton's classic deductions from phenomena. *Proceedings of the 1990 Biennial Meeting of the Philosophy of Science Assocation*, 2:183–196.

Harper, W. (2002). Newton's argument for universal gravitation. In Cohen, I. B. and Smith, G. E., editors, *Cambridge Companion to Newton*, pages 174–201. Cambridge University Press, Cambridge.

Harper, W. (2007). Newtons Methodology and Mercurys Perihelion Before and After Einstein. *Philosophy of Science*, 74:932–942.

Harrison, E. R. (1970). Fluctuations at the threshold of classical cosmology. *Phys. Rev.*, D1:2726–2730.

Hawking, S. W. (1982). The development of irregularities in a single bubble inflationary universe. *Physics Letters B*, 115:295–297.

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.

Hollands, S. and Wald, R. (2002a). Comment on inflation and alternative cosmology. hep-th/0210001.

Hollands, S. and Wald, R. (2002b). Essay: An Alternative to Inflation. *General Relativity and Gravitation*, 34:2043–2055.

Isenberg, J. and Marsden, J. (1982). A slice theorem for the space of solutions of Einstein equations. *Physics Reports*, 89:180–222.

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

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

Kragh, H. (1996). *Cosmology and Controversy*. Princeton University Press, Princeton.

Liddle, A. and Lyth, D. (2000). *Cosmological Inflation and Large-Scale Structure*. Cambridge University Press, Cambridge.

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

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.

Longair, M. (2006). *The cosmic century: a history of astrophysics and cosmology*. Cambridge University Press.

Madsen, M. S. and Ellis, G. F. R. (1988). The evolution of ω in inflationary universes. *Monthly Notices of the Royal Astronomical Society*, 234:67–77.

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, J. and Brandenberger, R. H. (2001). The trans-Planckian problem of inflationary cosmology. *Physical Review D*, 63:123501.

McMullin, E. (1993). Indifference principle and anthropic principle in cosmology. *Studies in the History and Philosophy of Science*, 24(3):359–389.

Misner, C. W. (1969). Mixmaster universe. *Physical Review Letters*, 22:1071–1074.

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.

Nanopoulos, D. V., Olive, K. A., and Srednicki, M. (1983). After primordial inflation. *Physics Letters B*, 127:30–34.

Olive, K. A. (1990). Inflation. *Physics Reports*, 190:307–403.

Peacock, J. R. (1999). *Cosmological Physics*. Cambridge University Press, Cambridge.

Peebles, P. (1999a). Is Cosmology Solved? An Astrophysical Cosmologist's Viewpoint. *Publications of the Astronomical Society of the Pacific*, 111:274–284.

Peebles, P. (1999b). Summary; Inflation and Traditions of Research. *Arxiv preprint astro-ph/9905390*.

Peebles, P. J. E. (1980). *Large-scale Structure of the Universe*. Princeton University Press, Princeton.

Peebles, P. J. E. and Yu, J. T. (1970). Primeval adiabatic perturbation in an expanding universe. *Astrophys. J.*, 162:815–836.

Penrose, R. (1989). Difficulties with inflationary cosmology. *Annals of the New York Academy of Sciences*, 271:249–264.

Penrose, R. (2004). *The road to reality*. Jonathan Cape.

Press, W. H. and Vishniac, E. T. (1980). Tenacious myths about cosmological perturbations larger than the horizon size. *Astrophysical Journal*, 239:1–11.

Rindler, W. (1956). Visual horizons in world models. *Monthly Notices of the Royal Astronomical Society*, 116:662–677.

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*.

Shafi, Q. and Vilenkin, A. (1984). Inflation with SU(5). *Physical Review Letters*, 52:691–694.

Smith, G. E. (2002). 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.

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.

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.

Turner, M. (1999). Cosmology solved? Quite possibly! *Publications of the Astronomical Society of the Pacific*, 111:264–273.

Unruh, W. G. (1997). Is inflation the answer? In Turok, N., editor, *Critical Dialogues in Cosmology*, pages 249–264, Singapore. World Scientific.

Vachaspati, T. and Trodden, M. (1999). Causality and cosmic inflation. *Physical Review D*, 61(2):23502.

Vilenkin, A. and Shellard, E. (2000). *Cosmic strings and other topological defects*. Cambridge University Press.

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. 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.