Utrecht
University

Another critical look at inflationary cosmology

A.M. Fleuren
5928419

July 17, 2024

Supervisors:

Niels Martens
(History and Philosophy of Science)

Gil Cavalcanti
(Mathematical Sciences)

Antonio Ferreiro de Aguiar
(History and Philosophy of Science)

Abstract

I present an overview of the criticism of the theory of cosmological inflation of the past 25 years. I argue that the three main problems motivating the inflationary explanation are the flatness problem, the uniformity problem, and the structure formation problem. The flatness problem has been resolved. In analysing the other two problems, an alternative explanation emerges. This explanation relies on law-like assumptions about the initial conditions of the universe. I claim that this is a proper explanation. This places the inflationary explanation on par with an alternative explanation. I argue that different extra-empirical considerations lead cosmologists to prefer one above the other. The way forward lies in the middle: an amalgam of the two explanations promises to be the most fruitful.

Acknowledgements

A special thanks to:

- Antonio Ferreiro de Aguiar, for the many insightful conversations on which this thesis is built, for the patience to deal with my many questions, and for always having more papers ready for me to read;
- Niels Martens, for the warm supervision, the continuing support, and the many thought-provoking questions;
- Gil Cavalcanti, for helping me out with the peskiest problems, and for helping me appreciate the value of my background in mathematics;
- Philip Helbig, for being so kind to give feedback on this thesis;
- Maura Burke, for helpful conversations on the philosophy of explanation;
- the UPAC group, for a great many inspiring discussions.

Contents

Chapter 1

Introduction

About 25 years ago, philosophers John Earman and Jesús Mosterín published a critical evaluation of the theory of cosmological inflation (Earman and Mosterin 1999). Cosmological inflation proposes a period of exponential expansion of the early universe, driven by a scalar field. The central claim of Earman and Mosterín is that inflation developed into a dominant cosmological theory not because the standard model of cosmology is empirically inadequate, but because the explanations it provides are generally deemed unsatisfactory.[1] Originally, the explanations of standard cosmology were found wanting on account of three problems (Guth 1981). These were the horizon problem, the flatness problem, and the monopole problem. It is these problems that open the door for inflation.

To get a flavour of the criticism of Earman and Mosterín on inflation, consider their arguments on the monopole problem. Grand unified theories (GUTs) predict that at energy scales above 10^{14} GeV the electroweak force and the strong force are united. As the nascent universe cools to such temperatures at around 10^{-36} after the big bang, the two forces are separated. In this process, a theoretical GUT Higgs field, specifically introduced for this purpose, undergoes a spontaneous symmetry breaking. Such a process is not a smooth one. Imagine a pond of water freezing. Across many patches of water in the pond, freezing will start at the same time, forcing the ice crystals formed in the process to pick an orientation. Different patches will be oriented differently, and at their borders the orientations will be misaligned. Such misalignments are called topological defects. By analogy, as the symmetry of the GUT Higgs field spontaneously breaks, it forms point-like topological defects: magnetic monopoles. The absence of any sightings of such monopoles, in spite of their large mass, is taken to be problematic by the inflationary cosmologist. The solution, however, is easy. An accelerated expansion, taking place after the electroweak-strong force separates, would dilute the monopoles to such a low density that the probability of us ever finding one is effectively zero.

Of course, as is noted by Earman and Mosterín, the monopole problem is only a problem insofar as one subscribes to GUTs. Although GUTs bear formal similarities to the Glashow-Salam-Weinberg theory uniting the electromagnetic and weak interactions, only the latter enjoys empirical confirmation, such as the detection of the Higgs boson in 2012. Experimental verifi-

[1] I use the term 'standard cosmology' to refer to the standard hot–big-bang model of the universe, with its evolution governed by the Friedmann equations. There is no inflation in this model.

cation of GUTs, on the other hand, is nigh impossible, because its predictions concern energy scales many orders of magnitude above any experiments thus far realised on Earth. The non-observation of the decay of the proton even disfavours GUTs. So, the absence of monopoles is more properly construed as evidence against GUTs. Even if one is strongly committed to GUTs, belief in inflation is not necessary, since the monopole problem can be addressed within the framework of GUTs without invoking inflation (Earman and Mosterin 1999, p. 17). I agree with these remarks by Earman and Mosterín, so I will continue and dismiss the monopole problem.

I want to stress that this is a serious blow for inflation. Indeed, it would have been beautiful if the once-so-distant fields of particle physics and cosmology were working hand-in-hand in a theory of the early universe. Dismissing the monopole problem amounts to severing this connection, reducing the unificatory power of inflation. Despite efforts to restore this connection, most notably via Higgs inflation, it remains weak at best. Even the Higgs inflation model—one of the more promising inflationary models in which the false vacuum state of the Higgs boson provides the springboard for inflation—cannot survive without making a great deal of assumptions about the properties of the Higgs field, interpreted as a *classical* field, that have no experimental verification.

In this spirit, Earman and Mosterín evaluate the other problems, too. The point is that the extend to which these problems are problematic depends on one's attitude towards explanation. What kind of explanations are we looking for in cosmology? Do the flatness and horizon problems stand in need of more explaining? Can the inflationary explanation adequately solve these problems in the first place? In probing the standard inflationary story with such questions, Earman and Mosterín develop a rather negative assessment. It will be the goal of the first two chapters to evaluate these arguments in light of developments in the field over the last two decades.

I hold that the flatness problem has been resolved. A largely neglected body of literature advances arguments that are fatal to the problem. As for the horizon problem, I agree with Earman and Mosterín that this is not a problem *sensu stricto* in Friedmann–Lemaître–Robertson–Walker (FLRW) cosmology. There is a more general problem, however, called the uniformity problem. I find two different types of explanation able to resolve this problem. One makes suitable assumptions about the initial state of the universe (say, at the Planck time), the other introduces a new dynamical mechanism, namely inflation. The appearance of these seemingly opposing styles is a recurring theme of this thesis.

If this was the full story, then this thesis could have been quite short. But since the publication of the first paper on inflation by Guth, two new problems of standard cosmology have entered the scene. These are the entropy problem and the structure-formation problem, to be discussed in Chapter 4 and Chapter 5, respectively. Strikingly, the aforementioned rival styles of explanation appear again here. I carefully analyse the debate on these problems in the literature, from which a clearer picture of the two sides emerges. I refer to the advocates of either side as belonging to an explanatory camp: the initial-conditions camp and the dynamical camp.

The second objective of this thesis is to place these two explanations on equal footing. I will argue that different extra-empirical considerations may induce a preference for one explanation above the other. Both, however, are proper explanations, and allegiance to either camp is tenable. Since the two explanations are on equal footing, I suggest that the more conservative initial-conditions explanation, currently left out in the cold, deserves more attention in terms of research effort. Nevertheless, I believe that the most interesting explanations are hybrids, borrowing pieces from both styles of explanation. I present some examples that illustrate this.

I note that throughout this thesis I work in Planck units, where

$$1 = c = G = \hbar = k_{\mathrm{B}}. \tag{1.1}$$

Chapter 2

The flatness problem

The original motivation for the inflationary paradigm resides in its ability to resolve two classical problems of standard cosmology. Apart from the monopole problem and its connection with GUT phase transitions, Guth's original paper centers on straightening out the flatness and horizon problems (Guth 1981). I have yet to encounter a textbook introduction of inflationary cosmology that does not begin with a treatment of these problems. It is therefore only natural that we start with an analysis of these problems to evaluate the inflationary explanation, to begin with the flatness problem.

In the present chapter, I discuss the ideas of Marc Holman regarding the flatness problem (Holman 2018). Despite its presentation in the literature as a particularly thorny problem in non-inflationary cosmology, I argue that Holman's arguments takes the sting out of it. These arguments stand in a longer tradition of arguments against the flatness problem, for a historical overview see Helbig (2021). For the sake of clarity, I restrict my focus to Holman's article, which captures all the important points. Loosely speaking, the argument goes against the claim that flatness is somehow 'unnatural'. In FLRW spacetimes, flatness is seen to be a generic property. Note that the flatness problem is well defined only for FLRW spacetimes, because it is a statement about the flatness parameter k from the Friedmann equations. In the next chapter, I will also consider non-FLRW spacetimes.

I consider two different presentations of the flatness problem commonly found in the literature. Holman shows that each contains dubious arguments. Closer analysis of each reveals that there may not be much of a problem after all.

1. **The fine-tuning argument.** The fine-tuning argument is typically presented as follows. Start from the Friedmann equation in a singular FLRW spacetime, without cosmological constant for the sake of simplicity:

   $$H^2(t) = \big(\frac{\dot{a}(t)}{a(t)}\big)^2 = \frac{8\pi}{3}\rho(t) - \frac{k}{a^2(t)}. \tag{2.1}$$

   The singularity is characterised by the conditions that $a \to 0$, and $\rho \to \infty$ as $t \to 0$. We can write the Friedmann equation in terms of the matter density parameter.

   $$\Omega(t) - 1 = \frac{k}{\dot{a}^2(t)}, \qquad \Omega(t) := \frac{8\pi}{3H^2(t)}\rho(t) \tag{2.2}$$

If, as suggested by observations, we take the present matter density parameter $\Omega(t_0)$ to differ from unity by a small quantity $\epsilon > 0$, we have that $k = \epsilon \dot{a}^2(t_0)$. Substituting this back in Equation 2.2 for some time t' close to $t = 0$, we get

$$\Omega(t') - 1 = \epsilon \frac{\dot{a}^2(t_0)}{\dot{a}^2(t')}. \tag{2.3}$$

It follows from the second Friedmann equation,

$$\frac{\ddot{a}}{a} = -\frac{4\pi}{3}(\rho + 3P) \tag{2.4}$$

together with the strong energy condition

$$\rho + P \geq 0, \qquad \rho + 3P \geq 0, \tag{2.5}$$

that $\ddot{a} < 0$. If spacetime is taken to be singular, then $\dot{a}$ blows up as $t \to 0$. Therefore, by Equation 2.3, at some t' close to $t = 0$, the matter density parameter will be extremely close to 1. For example, if t' is taken to be a second, and we take $\epsilon = 1$, we have that $\Omega(t')$ is fine-tuned to 1 to an accuracy of $1 : 10^{16}$.

Holman makes the point that any singular FLRW universe will display flatness fine-tuning. This is seen immediately from Equation 2.2 in the limit as $t \to 0$. Clearly, any universe in which the Friedmann equation applies has the feature that $\Omega(t)$ becomes arbitrarily close to 1 near the big bang singularity.

The fine-tuning problem is: given that Ω is so close to 1 today, it is unlikely that Ω must have been so close to 1 to an incredible precision in the past. However, even if Ω was not close to 1 today, we would have that Ω must have been close to 1 to an incredible precision in the past. So, we cannot say that it is *unlikely* that Ω was so close to 1 in the past: in fact, any singular FLRW spacetime has this feature. Restricting ourselves to the FLRW context, as the coherency of the flatness problem requires us to, we find the problem of unnaturally fine-tuned flatness resolved.

2. **The dynamical-instability argument.** The dynamical-instability argument is typically presented as follows. Given that Ω was so close to 1 in the early universe, it is unlikely that it is still close to 1 a long time later (today). This can be seen from Equation 2.2 for non-zero k. Since $\ddot{a} < 0$ for all times, we have that $\dot{a}$ is always decreasing. This means that the difference of $\Omega(t)$ from 1 is always increasing. That is, Ω is always moving away from 1, so it is strange that after billions of years we observe it to be so close to 1.

   Holman refutes this argument as follows. Take as a starting point the empirical fact that Ω today is close to 1. So, by evolving the Friedmann dynamics in the past direction, we see that Ω must have been very close to 1 in the early universe. But, by the argument in the previous item, this is not unlikely in FLRW models. So, it is not justifiable to say that it is unlikely that after billions of years Ω is still close to 1; again, simply because the extreme proximity to 1 of Ω in the early universe is not unlikely.

   The slightly more technical version of the argument is as follows. Start again from the Friedmann equation, this time with a cosmological constant:

$$H^2 = \frac{8\pi}{3}\rho + \frac{\Lambda}{3} - \frac{k}{a^2}. \tag{2.6}$$

The equation of state and the fluid equation give

$$\rho = Ca^{-3(1+w)}, \tag{2.7}$$

where C is a constant. We can rewrite the Friedmann equation as

$$\underbrace{\frac{8\pi}{3H^2}Ca^{-3(1+w)}}_{:=\Omega_m} + \underbrace{\frac{\Lambda}{3H^2}}_{:=\Omega_\Lambda} = 1 + \underbrace{\frac{k}{\dot{a}^2}}_{:=\Omega_k}. \tag{2.8}$$

where we interpret the contributions of the cosmological constant and the curvature as energy densities. From this we see that

$$\Omega := \Omega_m + \Omega_\Lambda = \frac{8\pi C + \Lambda a^{3(1+w)}}{8\pi C + \Lambda a^{3(1+w)} - 3ka^{1+3w}}. \tag{2.9}$$

Observe that $\Omega = 1$ both at $a = 0$ and in the limit of $a \to \infty$. It is therefore unclear whether it is unlikely that the current value for Ω is close to 1.

To investigate this, Holman rewrites the Friedmann dynamics for a universe with only dust and cosmological constant ($w = 0$) in terms of the density parameters:

$$\Omega_m' = (\Omega_m - 2\Omega_\Lambda - 1)\Omega_m, \qquad \Omega_\Lambda' = (\Omega_m - 2\Omega_\Lambda + 2)\Omega_\Lambda, \tag{2.10}$$

where $'$ denotes differentiation with respect to the time parameter $\eta := \log\left(\frac{a}{a_0}\right)$. This system has been plotted in Figure 2.1. The black curves correspond to different universes: the red line is a flat universe with $\Omega_m + \Omega_\Lambda = 1$, to the left of it are universes with negative curvature, and to the right of it are universes with positive curvature. Each curve is characterised by a constant of motion α. It is defined as

$$\alpha := \pm\frac{27\Omega_m^2\Omega_\Lambda}{4\Omega_k^3}, \qquad \text{for } k = \mp 1. \tag{2.11}$$

It is constant along any trajectory that is a solution to the dynamical system described by Eq. 2.10. Curves with a higher value for α are closer to the red line, which corresponds to the limit as $\alpha \to \infty$. The black curve closest to the red line has $\alpha = 500$. Data from the CMB power spectrum suggest that the FLRW model (with dust and cosmological constant) that best describes our universe has $\alpha \gtrsim 3000000$.[1] Thus, it is not unlikely that after billions of years Ω is close to 1; in fact, this seems to be the case for the entire trajectory that we are on, because the high value for α implies that we are close to the red trajectory at all times. This dynamical-systems approach to the flatness problem comes from Lake (2005), who even shows that, for models with $k = 1$ and $\Lambda > 0$, it is a state of non-flatness that requires fine-tuning. That is, only if α is close to 1 will such a universe be one that has large values of Ω_m and Ω_Λ. This argument has been completed by Helbig (2012). Classifying all FLRW models by their value for k and the sign of their cosmological constant, he shows that in each class either flatness is generic or observations of Ω_m and Ω_Λ corresponding to non-flatness are anthropically disfavoured. Indeed, the

[1] Holman shows that this analysis still holds for a more realistic matter composition of the universe.

> remainder of the parameter space can be divided into models that collapse, and models that expand forever. If the universe undergoes a collapse, the values of Ω_m and Ω_Λ can become infinitely large. However, their sum remains significantly greater than 1 for only a brief and distinct period in the universe's history. In the scenario of eternal expansion, the weak anthropic principle provides an explanation. The value of Ω cannot be arbitrarily small here, because we do not expect to have observers in such a universe.

The arguments presented so far convincingly defuse the flatness problem. There is no fine-tuning problem because there is nothing unlikely about extreme flatness in the early universe. A dynamical-systems analysis even shows that flatness is a likely state for our universe to be in. Thus, I consider the flatness problem resolved.[2]

The arguments against the flatness problem can be found in an embryonic form in the article by Earman and Mosterín.

> Now it is not literally true that the standard big-bang model offers no explanation of why the value of Ω in the early universe is very nearly equal to 1. [For each value of k,] $\Omega \to 1$ as $t \to 0^+$; in this sense, the value of Ω must be nearly equal to 1 sufficiently near the big bang. The complaint the inflationary cosmologists mean to make is that if explanations of the present state of affairs are based on conditions at an early time, such as [the Planck time]—which is arguably the earliest time at which quantum gravitational effects will be quiescent enough to permit a classical spacetime description—then the explanation given by the standard big-bang model is found wanting because it must rely on special initial conditions. (Earman and Mosterin 1999, pp. 22–23)

Notice that they are aware of the fact that there is nothing strange about fine-tuned flatness in the early universe. They note, as Holman did, that any singular FLRW universe is very flat close to the big bang. Whether one thinks this is a problem depends on what they take as adequate styles of explanation. To Earman and Mosterín, an explanation relying on special initial conditions is perfectly fine. This important theme will resurface throughout this thesis.

[2] There is a another version of the flatness problem which is rarely mentioned. It goes: if the Universe were not very nearly flat, then soon after the big bang, it would have either collapsed or expanded so rapidly that no structure could have formed. This problem is addressed in Helbig (2020).

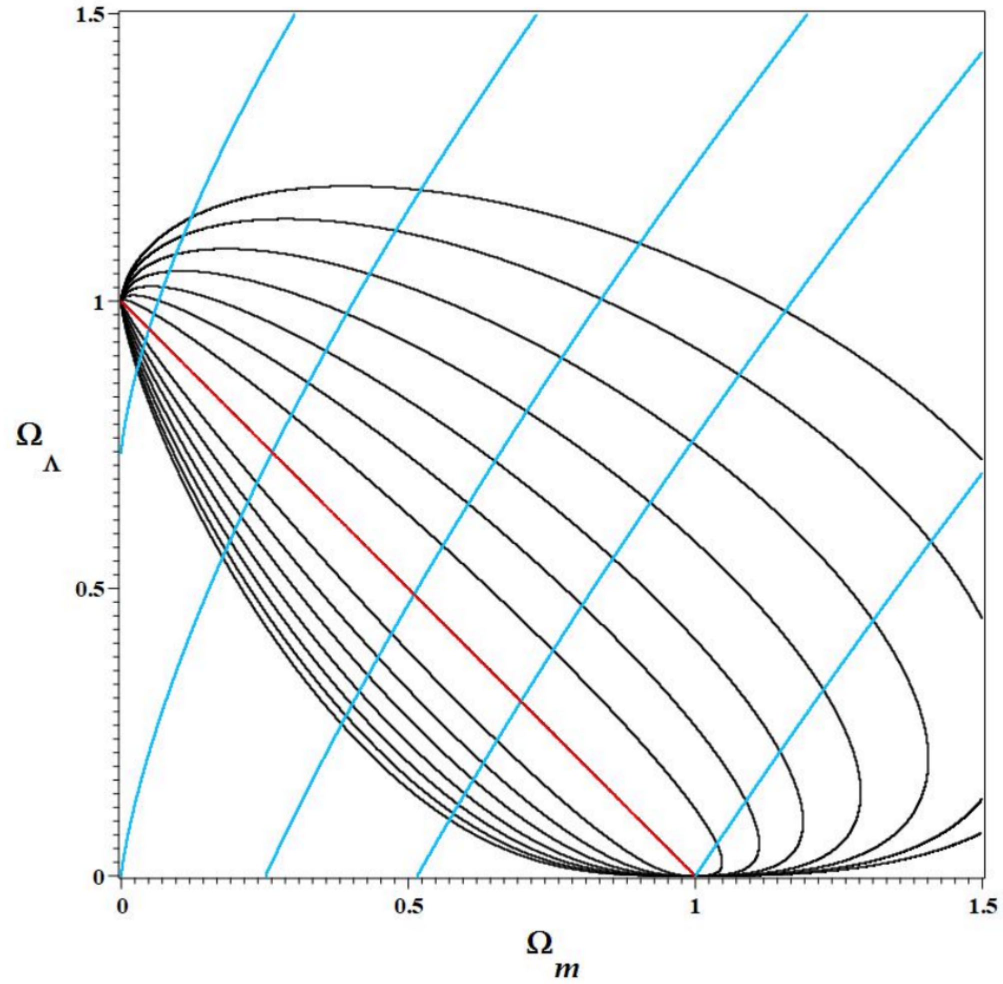


Figure 2.1: Phase portrait for density parameters in a universe with dust and cosmological constant. Note that trajectories start at $(\Omega_m, \Lambda) = (1, 0)$ and end at $(\Omega_m, \Lambda) = (0, 1)$. All the trajectories correspond to universes that expand forever. On each of the light-blue lines, the age of the universe is constant, corresponding, from right to left, to an age of $\frac{2}{3}$, $\frac{3}{4}$, 0.826, 1, and $\frac{3}{2}$ inverse Hubble parameters. The black trajectories correspond, from the outside in, to values of α of 6, 8, 12, 20, 40, 100, and 500. (Holman 2018, p. 11)

Chapter 3

Uniformity problem

The second classical problem of inflation is the horizon problem. After briefly giving its standard formulation, I show that it is not really a problem, much like the flatness problem. There is, however, a more general problem outside of the FLRW context. This is the uniformity problem. In the present chapter, I will evaluate the inflationary solution to this problem. I will give a rigorous proof of Wald's homogeneous no-hair theorem, and give arguments why this makes inflation an effective isotropising agent in spite of criticism by Earman and Mosterín. Nevertheless, an alternative explanation in terms of special initial conditions is also adequate. There is an ongoing debate on whether the onset of inflation in an inhomogeneous universe itself requires special initial conditions. I will analyse this debate, and show that it exposes underlying considerations of 'naturalness' on either side. The rivaling explanations of uniformity can be understood better in light of these considerations.

For the sake of reference and completeness, I will now introduce the textbook presentation of the horizon problem, following the well-known GR textbook of Sean Carroll (Carroll 2019). For the sake of simplicity, assume a flat FLRW universe without cosmological constant which is matter dominated. Then a radially moving photon satisfies

$$0 = ds^2 = -dt^2 + a^2 dr^2. \tag{3.1}$$

Integrating gives the comoving distance that such a photon travels

$$\Delta r = \int_{t_1}^{t_2} \frac{1}{a} dt = \int_{t_1}^{t_2} \left(\frac{t}{t_0}\right)^{-2/3} dt = \left[\frac{t_0}{3}\left(\frac{t}{t_0}\right)^{1/3}\right]_{t_1}^{t_2} = \frac{t_0}{3}(\sqrt{a_2} - \sqrt{a_1}). \tag{3.2}$$

Taking $a_{\mathrm{cmb}} = 1/1200$, as Carroll does, we find that the comoving distance from a particle in the CMB to an observer on Earth is

$$\Delta r \approx 0.33 t_0, \tag{3.3}$$

whereas the comoving horizon size of such a particle is

$$r_{\mathrm{hor}} \approx 0.01 t_0. \tag{3.4}$$

So, two widely separated regions in the CMB are well outside each others horizon, which implies that they have never been in causal contact at the time of the CMB. Yet these regions have remarkably similar temperatures. This is the horizon problem.

For FLRW universes, the horizon problem suffers the same fate as the flatness problem. That is, the horizon problem is resolved for strict FLRW models of the universe, because these are isotropic and homogeneous by assumption. This uniformity is present throughout the entire evolution of the universe, so there is nothing conspiratorial about causally disconnected CMB patches displaying uniform properties.

Outside of the FLRW context, there could be a horizon problem. This would happen if the initial conditions of the universe were non-uniform, i.e. anisotropies or inhomogeneities exist, and if there was no causal mechanism that could have 'smoothed out' these wrinkles by the time of recombination. If this is the case, then there is need for an additional mechanism, such as inflation. I will refer to this problem as the uniformity problem.

The uniformity problem received a lot of attention well before the invention of inflation. Collins and Hawking (1973) already analysed the problem in a seminal paper. To elaborate on the uniformity problem, and as a nice historical introduction, I will sketch their ideas here. Let us start with the observation that the universe is isotropic and homogeneous to a very high degree, based on data from the CMB. How can we then explain the existence of local non-uniformities, such as galaxies? To answer this question, we could postulate an initial spectrum of perturbations to the FLRW metric. However, these perturbations grow too slowly to explain the formation of large-scale structure. Therefore, we have to postulate an initially non-uniform universe. To avoid an *ad hoc* solution to this problem, we have to assume a generic initial state, with anisotropies and inhomogeneities of all kinds.

Collins and Hawking show that there is no open set (with respect to a reasonable topology) in the space of all initial conditions, which contains at least one homogeneous initial configuration, and which produces a homogeneous and isotropic universe at late times. They claim the extra condition that the open set contains one homogeneous set of initial conditions is not a strong one, because they do not expect anisotropies or inhomogeneities to arise from homogeneity. The cosmological constant is assumed to be zero. Their strategy is to show that there exists a homogeneous but anisotropic perturbation mode ζ_0 of the metric that increases in amplitude with time. The universe's expansion rate compared to the escape velocity (the minimum expansion velocity required to avoid recollapse) now gives three cases to consider:

1. **The universe is expanding slower than the escape velocity.** In this case, the universe will exist for only a finite amount of time. Thus, there will not be enough time for the anisotropic perturbation mode ζ_0 to become arbitrarily small. This would not result in an anisotropic universe at late times if the initial amplitude of this mode is zero, but this specific situation has measure zero in the space of initial conditions.

2. **The universe is expanding with exactly the escape velocity.** This is a set of measure zero in the space of initial conditions. Any small deviation from its expansion rate will result in either recollapse or the growth of anisotropy.

3. **The universe is expanding faster than the escape velocity.** In this case, the amplitude of the anisotropic mode will grow because long-range gravitational effects will eventually dominate. Again, if its initial amplitude were zero, this would not result in anisotropy at late times, but this initial value has measure zero in the space of initial conditions.

Since our universe only has a finite age, it could be objected that the anisotropic mode has not had enough time to grow. Any tiny deviation from the escape velocity or from a zero initial amplitude of ζ_0 will, however, ultimately result in a very anisotropic universe. We would reach the conclusion that our universe has remained isotropic for all of its lifetime up until the present, but is ultimately destined to become anisotropic. This would be awkward, because it implies that the value of the amplitude of ζ_0 is special. The dilemma presented by Collins and Hawking is then the following. In order to explain the existence of large-scale structure in our universe, we have to postulate an initial spectrum of perturbations. These perturbations, however, have to be chosen in such a way that the amplitude of ζ_0 and the universe's escape velocity are zero. So, to explain both the formation of galaxy clusters and the isotropy of the CMB, we are forced to make the assumption of special initial conditions. In absence of an established dynamical mechanism to provide the desired explanation, Collins and Hawking resort to the anthropic principle to escape a commitment to such a fine-tuned initial state. At present, the prevailing solution to the dilemma described by Collins and Hawking is a dynamical one: inflation.

In the next section, I will first consider how non-uniformities can be smoothed out by introducing a positive cosmological constant, which is a surrogate for inflation, as we will see. I will thoroughly analyse Wald's no-hair theorem and evaluate its validity in the inflationary context. I postpone the question of structure formation from such a mechanism to Chapter 5. The more philosophically-inclined reader may wish to skip this next section.

3.1 The smoothing mechanism

The cosmic no-hair theorem is the statement that expanding universes with a positive cosmological constant—and in which some appropriate energy conditions are satisfied—asymptotically approach the de Sitter spacetime. In 1983, Wald proved this result for homogeneous spacetimes (Wald 1983). Loosely speaking, this proves that the cosmological constant functions as an effective way to isotropise a universe. More precisely, the statement being proved is the following. Let (M, g) be a manifold with a metric satisfying

$$G_{ab} = -\Lambda g_{ab} + 8\pi T_{ab}, \tag{3.5}$$

with $\Lambda > 0$. If:

- T satisfies the dominant energy condition (DEC) and the strong energy condition (SEC),
- $K > 0$ at $t = 0$ (to be defined),
- M has a foliation of space-like, homogeneous (to be defined) hypersurfaces,
- M is not of Bianchi type-IX,

then $K \to \sqrt{3\Lambda}$, $\sigma_{ab} \to 0$, and $T_{ab} \to 0$.

The energy conditions play a crucial role. I state them explicitly here. The DEC asserts that for any two co-oriented time-like vectors η^a and ξ^b, the stress-energy tensor satisfies

$$T_{ab}\eta^a\xi^b \geq 0. \tag{3.6}$$

It is motivated by the conviction that the flux of energy-momentum is causal. Note that the DEC implies the weak energy condition (WEC), which states that for any time-like vector η^a, the stress-energy tensor satisfies

$$T_{ab}\eta^a\eta^b \geq 0. \tag{3.7}$$

The SEC asserts that for a time-like vector η^a, the stress-energy tensor satisfies

$$(T_{ab} - \frac{1}{2}g_{ab}T)\eta^a\eta^b \geq 0. \tag{3.8}$$

It is motivated by the idea that gravity should be attractive for particles traversing null geodesics.

3.1.1 Preliminary I: extrinsic curvature

The following exposition is due to Wald (1984). Let (M, g) be a manifold with a metric satisfying Einstein's equation. Let $O \subseteq M$ be an open set. A congruence on O is a family of geodesics (with respect to g) such that through each point $p \in O$ there passes exactly one geodesic. Consider a smooth congruence of time-like geodesics, and denote its tangent vector field by ξ^a. Parameterising the geodesics in the congruence by their proper time gives $\xi_a\xi^a = -1$. Define the spatial metric as

$$h_{ab} := g_{ab} + \xi_a\xi_b. \tag{3.9}$$

Any vector V^a in the tangent space of the congruence will be projected onto the tangent subspace of vectors orthogonal to ξ^a. That is,

$$(h_{ab}V^a)\xi^b = V^a\xi_a + \xi_a\xi_b g^{ab}V^b\xi^b = 0. \tag{3.10}$$

In addition, h_{ab} is idempotent and acts like the metric for vectors that are already orthogonal to ξ^a. We can define the extrinsic curvature tensor as the change of the spatial metric as we move along an integral curve defined by ξ^a. Rewriting this gives:

$$\begin{aligned} K_{ab} :&= \frac{1}{2}\mathcal{L}_\xi h_{ab} \\ &= \frac{1}{2}\mathcal{L}_\xi g_{ab} \\ &= \frac{1}{2}(\xi^c\nabla_c g_{ab} + g_{cb}\nabla_a\xi^c + g_{ac}\nabla_b\xi^c) \\ &= \nabla_a\xi_b, \end{aligned} \tag{3.11}$$

where in the second line we used that the Lie derivative of ξ along itself vanishes, and in the fourth line we used that the connection ∇ is the Levi-Civita connection associated with g and the Leibniz rule for connections. The extrinsic curvature is clearly a spatial (orthogonal to ξ^a) and symmetric tensor.

How should we interpret K_{ab}? Let $\gamma_s(t)$ denote a smooth 1-parameter subfamily of the congruence. Define the vector field $X^a := (\frac{\partial}{\partial s})^a$, which represents an infinitesimal displacement to nearby geodesics. Exploiting the fact that an s-dependent affine reparameterisation of the geodesics changes X^a by adding a multiple of ξ^a, we can always reparametrise t so that X^a is orthogonal to ξ^b. Then,

$$(\mathcal{L}_\xi(X^a))\xi_a = \mathcal{L}_\xi(X^a\xi_a) - (\mathcal{L}_\xi(\xi_a))X^a = 0. \tag{3.12}$$

This implies that the Lie bracket $[\xi, X]^a = 0$, so we have

$$\xi^b\nabla_b X^a = X^b\nabla_b\xi^a = X^b K_b{}^a. \tag{3.13}$$

So, K_{ab} measures the failure of X^a to be parallelly transported along ξ. An observer on some geodesic γ_0 would see a linear deformation (stretching and rotating) of their surrounding geodesics, described by K_{ab}. We can decompose the extrinsic curvature as:

$$K_{ab} = \frac{1}{3}Kh_{ab} + (K_{(ab)} - \frac{1}{3}Kh_{ab}) + K_{[ab]} = \frac{1}{3}Kh_{ab} + \sigma_{ab}, \tag{3.14}$$

where $K_{[ab]}$ vanishes because the extrinsic curvature is symmetric, $K = K_{ab}h^{ab}$ is the trace of K_{ab}, and

$$\sigma_{ab} := K_{(ab)} - \frac{1}{3}Kh_{ab} \tag{3.15}$$

is the shear of the congruence. The trace K measures the expansion of the congruence around some γ_0, and σ_{ab}, being the traceless part of K_{ab}, measures its shear. Intuitively, the shear can we thought of as the deformation of a sphere in the tangent space of M into an ellipsoid as it is transported along the flow of ξ.

We derive the Raychaudhuri equation:

$$\begin{aligned}
\xi^c\nabla_c K_{ab} &= \xi^c\nabla_c(\nabla_a\xi_b) \\
&= \xi^c\nabla_a(\nabla_c\xi_b) + R_{cab}{}^d\xi^c\xi_d \\
&= \nabla_a(\xi^c\nabla_c\xi_b) - (\nabla_a\xi^c)(\nabla_c\xi_b) + R_{cab}{}^d\xi^c\xi_d \\
&= -K_a{}^cK_{cb} + R_{cab}{}^d\xi^c\xi_d.
\end{aligned}$$

We have used that the Riemann tensor measures the failure of commutativity of the second covariant derivative, the Leibniz rule, and the fact that $\xi^c\nabla_c\xi_b = 0$ because each vector ξ^a is tangent to a geodesic. Contracting this equation then gives

$$\dot{K} := \xi^c\nabla_c K = -\frac{1}{3}K^2 + \sigma_{ab}\sigma^{ab} + R_{cd}\xi^c\xi^d, \tag{3.16}$$

where we have used that $h^c{}_b h^b{}_c = 3$ and that σ_{ab} is symmetric. Rewriting using Einstein's equation gives

$$\dot{K} = \Lambda - \frac{1}{3}K^2 - \sigma_{ab}\sigma^{ab} - 8\pi(T_{ab} - \frac{1}{2}g_{ab}T)\xi^a\xi^b, \tag{3.17}$$

where we have used that the trace of Einstein's equation gives

$$R = 4\Lambda - 8\pi T. \tag{3.18}$$

Next, we derive the initial value constraint equation. Multiplying Einstein's equation with $\xi^a\xi^b$ gives us

$$0 = G_{ab}\xi^a\xi^b - \Lambda - 8\pi T_{ab}\xi^a\xi^b. \tag{3.19}$$

Rewriting G_{ab} in terms of K_{ab} and the Ricci scalar of the homogeneous hypersurface ${}^{(3)}R$ using Gauss's equation, we get the initial value constraint equation

$$K^2 = 3\Lambda + \frac{3}{2}\sigma_{ab}\sigma^{ab} - \frac{3}{2}{}^{(3)}R + 24\pi T_{ab}\xi^a\xi^b. \tag{3.20}$$

Gauss's equation is a relation between the intrinsic and extrinsic curvature of a hypersurface, see Carroll (2019, p. 451).

3.1.2 Preliminary II: curvature in Bianchi models

I follow Chapter 7 of Wald (1984). Let G be a Lie group. Let $\psi_h : G \to G$ denote left-translation by h. Vector fields satisfying

$$(\psi_h)_* v^a = v^a \tag{3.21}$$

for all $h \in G$ are called left-invariant. Here $(\psi_h)_*$ is the induced push-forward map, canonically defined as

$$(\psi_h)_*(v)(f) = v(f \circ \psi_h) \tag{3.22}$$

for scalar functions f on G. Since left-invariance is closed under summation and scalar multiplication, the left-invariant vector fields form a vector space. The commutator of two left-invariant vector fields is again left-invariant:

$$\psi_h^*[v, w] = [\psi_h^*(v), \psi_h^*(w)] = [v, w]. \tag{3.23}$$

In a coordinate basis, we have that the commutator depends linearly on v^a and w^a so there exists a left-invariant tensor field $c^a_{\ bc}$, called the structure-constant tensor, such that

$$[v, w]^a = c^a_{\ bc} v^b w^c. \tag{3.24}$$

Anti-commutativity and the Jacobi identity for the commutator bracket imply:

$$c^a_{\ bc} = -c^a_{\ cb}, \qquad c^e_{\ d[a} c^d_{\ bc]} = 0. \tag{3.25}$$

A finite-dimensional vector space together with a (1,2)-type tensor satisfying these identities is called a Lie algebra.

Next, let M be a spatially homogeneous spacetime with metric g_{ab}, and let G be the Lie group of isometries of M. In general relativity, homogeneity entails that there exists a foliation of hyperspaces Σ_t such that for every two points $p, q \in \Sigma_t$ there exists a $g \in G$ such that $q = g(p)$. We assume that this g is the unique isometry with this property.[1] Fixing a point $p_0 \in \Sigma_t$, this ensures that there is a bijection

$$G \to \Sigma_t, g \mapsto g(p_0). \tag{3.26}$$

We can define a global basis $(\sigma^1)_a, (\sigma^2)_a, (\sigma^3)_a$ of dual vector fields on a hypersurface Σ_0 as follows. Define an arbitrary dual basis at $p \in \Sigma_0$, and define a dual basis at an arbitrary point q using the push-forward isometry $g_* : (T_p\Sigma_0)^* \to (T_{g(p)}\Sigma_0)^*$. We can write the spatial metric h_{ab}, defined in the previous section as

$$h_{ab} = \sum_{i,j=1}^{3} h_{ij} (\sigma^i)_a (\sigma^j)_b. \tag{3.27}$$

Since h_{ab} is left-invariant by the definition of isometry and global basis is left-invariant by construction, we have that the components h_{ij} are constant across Σ_0.

[1] Restricting consideration to a simply transitive action involves almost no loss of generality. This is because the only exception where G does not act simply transitively or lacks a subgroup with a simply transitive action is the group $SO(3) \times \mathbb{R}$ acting on $S^2 \times \mathbb{R}$. Spatially homogeneous models with this isometry group can be addressed separately. See Wald (1984, p. 171) and the references stated there.

The global time coordinate can be defined as follows. Let t^a be a vector at p orthogonal to Σ_0. Let γ be the geodesic with tangent vector t^a at p. This geodesic is orthogonal to the spatial Killing vectors spanning the homogeneous hypersurface Σ_0. Once orthogonal to these Killing vectors, it will remain orthogonal to them for every hypersurface that γ intersects, and so the tangent to γ will be a normal of each hypersurface that it intersects. Our spacetime M will then be foliated by a family of hypersurfaces defined by the proper time t of γ. The vector field $t^a := -\nabla^a t$ defines a constant normal vector field for each hypersurface Σ_t, given by the tangent vector to γ at $p \in \Sigma_t$. It is everywhere orthogonal to Σ_t. Homogeneity of the hypersurface implies that the integral curves of this vector field are all geodesics. Constructing a global spatial basis $\sigma^i(0)$ for some hypersurface Σ_0 as before, we can define the global basis for subsequent hypersurfaces as

$$(\sigma^i)_a(t) := (\Phi_t)_*\big(\sigma^i(0)\big)_a, \tag{3.28}$$

where Φ_t is the flow of the vector field t^a. The upshot is that we can decompose our spacetime as $M = \mathbb{R} \times G$, and the metric takes the form

$$g_{ab} = -\nabla_a t \nabla_b t + \sum_{i,j=1}^{3} h_{ij}(t)(\sigma^i)_a(\sigma^j)_b. \tag{3.29}$$

Let α_a be any left-invariant dual vector field, and let v^a and w^a be left-invariant vector fields. Then, $\alpha_a v^a$ is a left-invariant scalar, so it is a constant. Then, using that $\nabla_b(\alpha_a v^a) = 0$, we get

$$\begin{aligned} 2v^a w^b \nabla_{[a}\alpha_{b]} &= (v^a w^b - v^b w^a)\nabla_a \alpha_b \\ &= -v^a \alpha_b \nabla_a w^b + \alpha_b w^a \nabla_a v^b \\ &= -\alpha_b [v, w]^b \\ &= -\alpha_b c^b{}_{cd} v^c w^d. \end{aligned} \tag{3.30}$$

So, we have that left-invariant dual vector fields satisfy

$$2\nabla_{[a}\alpha_{b]} = -\alpha_c c^c{}_{ab}. \tag{3.31}$$

Next, note that the spatial basis on a hypersurface Σ_t is invariant under spatial isometries

$$\begin{aligned} g_*(\sigma^i(t))_a &= g_*((\Phi_t)_*(\sigma^i(0))_a) \\ &= (\Phi_t)_* g_*(\sigma^i(0))_a \\ &= (\sigma^i(t))_a, \end{aligned} \tag{3.32}$$

using that the flow along t^a and spatial isometries commute by construction, and the fact that the spatial basis at Σ_0 is constructed to be left-invariant. Finally, note that $\nabla_{[a}(\sigma^i)_{b]}$ is spatial, in the sense that

$$\begin{aligned} 2t^a \nabla_{[a}(\sigma^i)_{b]} &= t^a \nabla_a (\sigma^i)_b - t^a \nabla_b (\sigma^i)_a \\ &= t^a \nabla_a (\sigma^i)_b + (\sigma^i)_a \nabla_b t^a \\ &= \mathcal{L}_t (\sigma^i)_b \\ &= \lim_{\epsilon\to 0}((\Phi_\epsilon)_*(\sigma^i)_b - (\sigma^i)_b \\ &= \lim_{\epsilon\to 0}((\sigma^i(t+\epsilon))_b - (\sigma^i(\epsilon))_b \\ &= 0, \end{aligned} \tag{3.33}$$

where we used the Leibniz rule in combination with the fact that $(\sigma^i)_a$ and t^b are orthogonal. Since $\nabla_{[a}(\sigma^i)_{b]}$ is spatial, and $(\sigma^i)_a$ is invariant under spatial isometries, we have that $(\sigma^i)_a$ is a dual vector field satisfying Eq. 3.31. Thus, we obtain an important relation between the spatial basis of the metric and the structure-constant tensor

$$2\nabla_{[a}(\sigma^i)_{b]} = -c^c{}_{ab}(\sigma^i)_c. \tag{3.34}$$

A spatially homogeneous cosmology is then fully characterised by its 3-dimensional Lie group G and the components of the spatial metric $h_{ij}(t)$. Classifying the 3-dimensional Lie groups has been achieved by Bianchi. By Lie's third theorem, Lie algebras stand in one-to-one correspondence with simply connected Lie groups. Thus, the goal is to find all 3-dimensional vector spaces and structure-constant tensors c^a_{bc} that are anti-symmetric in the lower indices and satisfy the Jacobi identity. First, define a dual vector and a symmetric tensor as follows.

$$A_a := c^b{}_{ba} \tag{3.35}$$

$$M^{ab} := \frac{1}{2}\epsilon^{acd}(c^b{}_{cd} - \delta^b{}_c A_d) \tag{3.36}$$

Here, ϵ is the Levi-Civita symbol. Contracting Equation 3.36 with ϵ_{aef} and using the anti-symmetric property of c^b_{cd} gives

$$c^b{}_{ef} = M^{ab}\epsilon_{aef} + \delta^b{}_{[e}A_{f]}. \tag{3.37}$$

Using the Jacobi identity for $c^b{}_{ef}$, we get that

$$M^{ab}A_b = 0. \tag{3.38}$$

The Lie algebras are classified by pairs (A_b, M^{ab}) satisfying this equation. Different types are characterised by the rank and signature (up to overall sign) of M^{ab}, and whether or not $A_b = 0$. We remark that the Bianchi IX type is characterised by $A_b = 0$ and M^{ab} having rank 3 with signature $(+,+,+)$ or $(-,-,-)$.

From Eq. 3.29 we can compute the corresponding Levi-Civita connection, and from there the Riemann and Einstein tensors. The Einstein equation, together with Gauss's equation, then gives an expression for the curvature scalar of a homogeneous hypersurface

$${}^{(3)}R = -c^a{}_{ab}c^{c\ b}_{\ c} + \frac{1}{2}c^a{}_{bc}c^{c\ b}_{\ a} - \frac{1}{4}c_{abc}c^{abc}, \tag{3.39}$$

where we recall that we have a relation between the metric and the structure-constant tensor, given by Eq. 3.34. Plugging in Equation 3.37 and simplifying using Equation 3.38 gives

$${}^{(3)}R = -\frac{3}{2}A_bA^b - h^{-1}(M_{ab}M^{ab} - \frac{1}{2}M^2) \tag{3.40}$$

If ${}^{(3)}R > 0$, then we must have that

$$M_{ab}M^{ab} < \frac{1}{2}M^2, \tag{3.41}$$

because A^b is a spatial vector. This inequality implies that M^{ab} must be a positive or negative definite matrix. To see why this is true, choose a basis in which M^{ab} is a diagonal matrix.

Assume for the sake of contradiction that there are two diagonal elements with opposing signs. Plugging the values of this matrix into Eq. 3.41 gives a contradiction. So, all eigenvalues of M^{ab} must have the same sign.[2] The signature of M^{ab} must be either $(+,+,+)$ or $(-,-,-)$, and Equation 3.38 then gives $A_b = 0$, because M^{ab} is definite. Hence, we have a type IX Bianchi universe. Contrapositively, if we have any Bianchi universe that is not type IX, then

$$^{(3)}R \leq 0. \tag{3.42}$$

3.1.3 Proving the no-hair theorem

The initial value constraint equation 3.20, together with Inequality 3.42, the dominant energy condition, and the fact that $\sigma_{ab}\sigma^{ab} \geq 0$ gives

$$K^2 \geq 3\Lambda. \tag{3.43}$$

Assuming that the universe is initially expanding, a positive Λ implies that K is always positive. This gives

$$K \geq \sqrt{3\Lambda}. \tag{3.44}$$

Next, the Raychaudhuri equation 3.17, together with the strong energy condition gives

$$\dot{K} \leq \Lambda - \frac{1}{3}K^2. \tag{3.45}$$

We can integrate this inequality to get[3]

$$K \leq \frac{\sqrt{3\Lambda}}{\tanh\left(t/\sqrt{\frac{3}{\Lambda}}\right)}. \tag{3.46}$$

Thus, K is squeezed between two limits, and from Eq. 3.46 we find that it approaches $\sqrt{3\Lambda}$ exponentially fast on a time scale $\sqrt{3/\Lambda}$. Next, the initial value constraint equation and Eq. 3.46 imply that

$$\sigma^{ab}\sigma_{ab} \leq \frac{2}{3}(K^2 - 3\Lambda) \leq \frac{2\Lambda}{\sinh^2(t/\sqrt{\frac{3}{\Lambda}})}. \tag{3.47}$$

Therefore, the shear quickly approaches zero. Finally, the initial value constraint equation and Eq. 3.46 also give

$$T_{ab}\xi^a\xi^b \leq \frac{\Lambda/8\pi}{\sinh^2(t/\sqrt{\frac{3}{\Lambda}})}. \tag{3.48}$$

By the dominant energy condition, we know that if we choose an orthonormal basis with ξ^a as the time-like vector, then $T_{ab}\xi^a\xi^b$ is an upper bound for the other components of T_{ab}. This follows from expressing the dominant energy condition (Eq. 3.6) in terms of this basis, and aptly choosing the components of η^a and ξ^b in this basis.[4]

[2] Thanks to Gil Cavalcanti for pointing this out to me.

[3] Although differentiating inequalities can be a difficult matter, in this case a simple separation of variables does the job. Thanks to Gil Cavalcanti for suggesting this to me.

[4] Thanks to Antonio Ferreiro for pointing this out.

The conclusion is that $K \to \sqrt{3\Lambda}$, $\sigma^{ab} \to 0$, and $T_{ab} \to 0$. In other words, thinking of K in terms of its effects on a congruence of geodesics, the decomposition of Eq. 3.14 gives, for late times:

$$\frac{\partial}{\partial t} h_{ab}(t) = \frac{1}{3}\sqrt{3\Lambda} h_{ab}(t). \tag{3.49}$$

Solving this gives

$$h_{ab}(t) = \mathrm{e}^{2(t-t_0)/\sqrt{3/\Lambda}}\, h_{ab}(t_0). \tag{3.50}$$

At late times, then, the constant value of K corresponds to exponential spatial expansion; the vanishing of σ_{ab} to an isotropic universe; and the vanishing of T_{ab} to an empty universe. Locally, then, the universe approaches the de Sitter spacetime. If Λ is sufficiently large, then this conclusion will also hold for Bianchi type IX universes (Wald 1983).

3.2 Criticism on the inflationary solution to the uniformity problem

Having seen the proof that a (temporary) cosmological constant acts as a smoothing mechanism, I now turn to the question: can it be applied to the inflationary context to solve the uniformity problem? This question is inspired by some critical remarks made by Earman and Mosterín on the uniformity problem. I will first scrutinise these remarks, and then argue that they are not very convincing.

3.2.1 Earman and Mosterín

I discern several arguments in the paper of Earman and Mosterín. I will treat each of them in turn. First, they claim that the no-hair theorem is not applicable to the inflationary context (Earman and Mosterin 1999, p. 30). In order to produce an effective cosmological constant, we must impose strong conditions on the properties of the inflaton field. To see this, compare the stress-energy tensors of a universe with cosmological constant $T^{\Lambda}_{\mu\nu}$ with that of a universe with an inflaton field $T^{\phi}_{\mu\nu}$:

$$T^{\Lambda}_{\mu\nu} = T^{\text{matter}}_{\mu\nu} + \Lambda g_{\mu\nu}, \qquad T^{\phi}_{\mu\nu} = T^{\text{matter}}_{\mu\nu} + \partial_\mu\phi\partial_\nu\phi - \frac{1}{2}g_{\mu\nu}\partial_\alpha\phi\partial^\alpha\phi - g_{\mu\nu}V(\phi). \tag{3.51}$$

(To get some insight into how to get these equations, see the final section of Chapter 5.) Earman and Mosterín argue that the inflaton field can produce an effective cosmological constant only when

$$\partial_\mu\phi\partial_\nu\phi = 0 = \frac{1}{2}g_{\mu\nu}\partial_\alpha\phi\partial^\alpha\phi, \tag{3.52}$$

resulting in a constant $V(\phi)$. Moreover, the proof of Wald requires that any matter in the universe should satisfy the SEC and DEC. Since the inflaton field—taken to be a form of matter—requires negative pressure to engender accelerated expansion, it does not comply with the hypotheses of Wald's theorem.

To analyse the severity of this criticism, let me first introduce the slow-roll conditions. From varying the action of the scalar field

$$S = \int dx^4 \sqrt{g}(-\frac{1}{2}\partial_\mu\phi\partial^\mu\phi - V(\phi)), \tag{3.53}$$

we get the Klein-Gordon equation of motion

$$\ddot{\phi} + 3H\dot{\phi} - \frac{1}{a^2}\nabla^2\phi + V'(\phi) = 0. \tag{3.54}$$

Assuming homogeneity, the gradient term—corresponding to spatial energy variations of the scalar field—vanishes. We can then compute the energy density and pressure of the homogeneous scalar field as:

$$\rho_\phi = T_{00} = \dot{\phi}^2 + \frac{1}{2}(-\dot{\phi}^2 + \frac{(\nabla\phi)^2}{a^2}) + V(\phi) = \frac{1}{2}\dot{\phi}^2 + V(\phi), \qquad P_\phi = T_{ii} = \frac{1}{2}\dot{\phi}^2 - V(\phi). \tag{3.55}$$

So, the generalised (non-isotropic) Friedmann equation is given by

$$H^2 = \frac{1}{3M_{\mathrm{Pl}}^2}\rho_\phi = \frac{1}{3M_{\mathrm{Pl}}^2}(\frac{1}{2}\dot{\phi}^2 + V(\phi)). \tag{3.56}$$

The slow-roll conditions are

$$\epsilon := -\frac{\dot{H}}{H^2} < 1, \qquad |\eta| := \frac{\dot{|\epsilon|}}{H\epsilon} < 1. \tag{3.57}$$

The smallness of ϵ is equivalent to accelerated expansion $\ddot{a} > 0$, and the smallness of $|\eta|$ guarantees the persistence of this state of affairs. Using the Klein-Gordon equation for the inflaton field and the Friedmann equation, we can rewrite

$$\epsilon = \frac{3\dot{\phi}}{\dot{\phi} + 2V(\phi)}. \tag{3.58}$$

Requiring $\epsilon < 1$ is equivalent to requiring that the inflaton field potential energy dominates its kinetic energy. So, to spark off inflation, we demand $V > \dot{\phi}$, and for its persistence we require that $\ddot{\phi}$ is small. This is what is needed for a (time-dependent) cosmological constant, and for inflation to solve the uniformity problem. One could still claim, as Earman and Mosterín do, that this demand is contrived. I agree that this imposes conditions on inflation that are in some sense 'engineered', which can be seen as an explanatory deficiency. Criticising the slow-roll inflation on these grounds, however, does not give sufficient reason to renounce its ability to smooth out anisotropies.

Next, let me address the criticism about the energy conditions more rigorously. The total stress-energy tensor in Wald's proof can be written as

$$T_{ab}^{\mathrm{tot}} = -g_{ab}\Lambda_0 + T_{ab}, \tag{3.59}$$

for some $\Lambda_0 > 0$, and where T_{ab} satisfies the SEC and DEC. There is some freedom in defining T_{ab} in this decomposition which can be exploited to set

$$\Lambda(t_i) = \frac{1}{3}K^2, \tag{3.60}$$

where t_i marks the onset of inflation. See Maleknejad and Sheikh-Jabbari (2012, pp. 17–18). Then, from the initial value constraint equation with ${}^{(3)}R \leq 0$, we find that

$$T_{ab}\xi^a\xi^b \leq 0, \tag{3.61}$$

at t_i. Since $\dot{K} \leq 0$ (expansion of a congruence decreases during inflation by the SEC, see Eq. 3.45), T_{ab} violates the DEC during the entire inflationary era. So, if we assume that T_{ab} satisfies the SEC, then it must violate the DEC, and the hypotheses of Wald's theorem are not met. Note, however, that T_{ab}^{tot} always satisfies the DEC.

In Maleknejad and Sheikh-Jabbari (2012) it is shown that the cosmic no-hair result still applies to general inflationary dynamics on a homogeneous background, in particular to slow-roll inflation. Any inflationary dynamics ($\dot{H} + H^2 \geq 0$, where H is the Hubble parameter of the non-isotropic Friedmann equations) satisfies the DEC but violates the SEC. Such dynamics can always be described by a stress-energy tensor T_{ab} that can be decomposed as

$$T_{ab} = -g_{ab}\Lambda(t) + T_{ab}^{+}, \tag{3.62}$$

where T_{ab}^{+} satisfies the SEC and WEC, and $\Lambda(t) \geq 0$ for all t, see Maleknejad and Sheikh-Jabbari (2012, pp. 16–17). Note that Λ is allowed to vary with time to justify this decomposition. This decomposition does not suffer from the problem described above, which exposed an incompatibility between the DEC and SEC. It is shown that inflation imposes an upper limit on the Hubble-normalized shear, which signifies anisotropy, given by

$$\frac{|\sigma^i{}_j|}{H(t)} \leq \sqrt{2\epsilon}, \tag{3.63}$$

where ϵ is the slow-roll parameter. Since ϵ can grow during inflation, in principle anisotropies can grow too, in contrast to the no-hair theorem. Nevertheless, the bound on their amplitudes implies that they will remain small. Imposing the slow-roll conditions enforces even stricter bounds. Thus, inflation generally evades the criticism of Earman and Mosteŕın concerning the energy conditions. Although they make a valid point, the arguments of Maleknejad and Sheikh-Jabbari (2012) show that the no-hair theorem can be generalised sufficiently.

I want to remark that, for this generalised no-hair theorem to hold, one still needs to assume the DEC. Although the stress-energy tensor corresponding to the inflaton field satisfies this condition, there could in principle be other matter violating this condition. Therefore, assuming the DEC, or at least initial domination of matter satisfying the DEC, is still vital to any inflationary explanation of the uniformity problem.

Second, Earman and Mosterin (1999, pp. 30–31) pose the question whether Wald's result can be generalised to non-homogeneous models . They criticise the generalisation of Wald's result by Jensen and Stein-Schabes (1987) on the basis that it assumes there to be no regions of positive curvature on any hypersurface in the spacetime foliation. Since regions of positive curvature could form, in spite of the imposed energy conditions, the conclusion of the generalised proof would not hold. However, in Kleban and Senatore (2016) it is proven that the no-hair theorem can be generalised significantly. Assuming a globally hyperbolic spacetime that is initially expanding, the WEC, the existence of a compact Cauchy surface (implying global hyperbolicity), and a 3-curvature scalar that cannot be everywhere positive on each hypersurface, it can be shown that there must be an expanding region on each surface, whose expansion rate is at least that of de Sitter spacetime in the flat foliation if $\Lambda > 0$. This suggests (but does not logically necessitate) that the universe will always have an increasing volume, from which it follows that vacuum energy will dominate. This implies that the potential energy of the inflaton field, acting as vacuum energy, will come to dominate its kinetic and gradient counterparts. So, inflation will be triggered somewhere no matter how inhomogeneous the early universe is. In other words, we have that the combination of the WEC, a flat or open spacetime, and initial expansion ensures that rapid expansion persists in some region, even in the presence of any inhomogeneities. Therefore, inflation is bound to spark off locally if $\Lambda > 0$. This result is corroborated by numerical simulations in Clough et al. (2017). This is a considerable generalisation of Wald's result. But fully dismissing the original criticism that the no-hair theorem cannot be generalised to the inhomogeneous case is a bridge too far. First, the result seems to be at odds with more sophisticated objections to the consistency of (the onset of) inflation in an initially inhomogeneous universe. I dive into these in the next section. Second, I want to remark that this result depends strongly on the presence of a positive cosmological constant which cannot be an effective cosmological constant coming from inflationary dynamics. This would beg the question: inflation would have somehow begun already, but this is precisely what we want to conclude. Third, despite every slice having an expanding neighbourhood, it could be the case that this neighbourhood itself is shrinking, possibly even faster than its expansion rate. Rather than dismissing this situation as pathological and unphysical, as is done in Kleban and Senatore (2016), I think we should be careful to make such claims in the face of substantial inhomogeneity. Fourth, perhaps the assumption that excludes the manifold from having ${}^{(3)}R$ positive everywhere on each hypersurface is not as innocent as it seems. In fact, it seems to do a lot of work in the theorem in finding the desired expanding region. Finally, note that result still depends on an energy condition.

The third criticism of Earman and Mosterín on the inflationary explanation of uniformity concerns the possible reappearance of non-uniformities, which poses a fine-tuning problem for inflation (Earman and Mosterin 1999, pp. 26–27). Consider two distantly spaced regions on the last scattering surface. According to the inflationary story, we observe similar conditions in these two regions because inflation has smoothed a portion of our universe that encompasses the last scattering surface. As times passes, however, the area of the last scattering surface increases as more light has had time to reach observers on Earth. So, the portion of the universe that inflation has made uniform will eventually be smaller than the last scattering surface. At this time, one could ask: what do we observe when comparing two regions outside of the inflated portion of the universe (which have non-overlapping past light cones)? If we observe uniform conditions, we are in the same position as before the appeal to inflation, except that now we cannot appeal to it again. If we observe non-uniform conditions, there must have been some special initial conditions bringing about these conditions. Or so the argument goes.

I find myself not particularly compelled by such an argument. It would be perfectly fine to find non-uniformities outside of the inflationary patch. As long as these non-uniformities are in some sense generic (i.e. featureless and uncorrelated), we cannot speak of *fine-tuned* conditions. In what sense would they be fine-tuned? In other words, we might at later times discover non-uniformities in the last scattering surface, but to say these are due to *special* initial conditions would be too strong. Rather, it would be the effect of initial conditions *tout court*.

To sum, the criticism by Earman and Mosterín against the applicability of the no-hair theorem to solve the uniformity problem is not very persuasive. I grant that imposing the slow-roll conditions makes inflation somewhat less generic. To claim that inflationary dynamics do not conform to the setup of the no-hair theorem, however, would be unfounded. As we have seen, very general inflationary dynamics can be shown to effectively isotropise a homogeneous universe. Importantly, this requires the assumption that matter in the early universe satisfies the DEC. Relaxing the assumption of homogeneity, the validity of the no-hair theorem is still pending. Although a local version of the theorem holds fairly generally, there are still a number of caveats. In particular, the connection between this result and the onset of inflation seems shaky. I will take a closer look at the inhomogeneous case in the upcoming section. In conclusion, inflation *cum* appropriate energy conditions is effective at isotropising homogeneous universes.

3.3 The initial-conditions problem

From the previous section it is clear that initial inhomogeneity might pose a problem for the efficacy of the inflationary solution to the uniformity problem. Despite serious efforts to generalise the no-hair theorem, there is still the worry that initial inhomogeneity may prevent the onset of inflation. Here, I will give a brief overview of these problems. The debate presented here goes to the heart of the tension between dynamical and initial-conditions explanations. For an analysis of this debate in connection to the scientificality of inflation, see Dawid and McCoy (2023).

First, I will briefly discuss chaotic inflation, first described by Linde in Linde (1983). Consider a closed universe with Planck length in a state with the Planck energy density. From the moment the universe has a lower energy density than the Planck energy density, it can be described classically. It is assumed that the sum of the kinetic, gradient, and potential-energy densities of the inflaton field is of the same order as the Planck density:

$$\frac{1}{2}\dot{\phi}^2 + \frac{1}{2}(\nabla\phi)^2 + V(\phi) \sim \rho_{\mathrm{Pl}}. \tag{3.64}$$

By appealing to these densities taking on 'natural' values, it is argued that

$$\frac{1}{2}\dot{\phi}^2 \sim \frac{1}{2}(\nabla\phi)^2 \sim V(\phi) \sim \rho_{\mathrm{Pl}}, \tag{3.65}$$

see Linde (1985). If, in a given domain, there occurs an energy fluctuation such that the potential energy dominates the sum of the kinetic and gradient energies, then inflation is triggered. Within a Planck time, the kinetic and gradient terms become much smaller than the potential term, so inflation will continue. Alternatively, if there occurs an energy fluctuation such that the sum of the kinetic and gradient energies dominates the potential energy, then the total lifetime of the domain will be of the order of the Planck time.

To the best of my knowledge, the earliest observation of an initial-conditions problem for inflation related to pre-inflationary inhomogeneity is found in Piran (1986). Assume that we have a minimally coupled scalar field in a flat FLRW universe without cosmological constant. To introduce linear perturbations, we decompose the scalar field into Fourier modes. Assume that the mode with wavenumber k dominates all others. Recall from the discussion around Eq. 3.55 that the energy density of the scalar field is given by

$$\rho_\phi = \frac{1}{2}\dot{\phi}^2 + \frac{(\nabla\phi)^2}{2a^2} + V(\phi). \tag{3.66}$$

Then, since $\phi \approx \phi_k\, \mathrm{e}^{i\mathbf{k}\cdot\mathbf{x}}$, and using the fact that ϕ is real, we have that

$$\rho = \frac{1}{2}\dot{\phi}_k^2 + \frac{k^2}{2a^2}\phi_k^2 + V(\phi_k\, \mathrm{e}^{i\mathbf{k}\cdot\mathbf{x}}). \tag{3.67}$$

The Friedmann equation is then

$$H^2 = \frac{8\pi}{3M_{\mathrm{Pl}}^2}\rho = \frac{8\pi}{3M_{\mathrm{Pl}}^2}(\frac{1}{2}\dot{\phi}_k^2 + \frac{k^2}{2a^2}\phi_k^2 + V(\phi_k\, \mathrm{e}^{i\mathbf{k}\cdot\mathbf{x}})). \tag{3.68}$$

The problem now is that when $\phi_k > M_{\rm Pl}$ (required for chaotic inflation), we get

$$H > \sqrt{\frac{4\pi}{3}}\frac{1}{M_{\rm Pl}}(\frac{k}{a}\phi_k) > 2\frac{k}{a}. \tag{3.69}$$

That is, the physical wavelength a/k corresponding to the dominant inhomogeneous perturbation is more than double the physical Hubble radius. This means that homogeneity is required at a very large scale to trigger inflation.

In the last decade, such arguments have been revisited and exacerbated. In Ijjas, Steinhardt, and Loeb (2013) the following version of the argument is put forward in light of the data from the 2013 Planck mission. Plateau-like models of inflation are favoured by the Planck data based on density perturbations in the CMB. The energy scale of the plateau is constrained to be many orders of magnitude below the Planck scale:

$$\frac{M_{\rm Pl}}{M_{\rm I}} \sim 10^3 \cdot (\frac{10^{16}\text{GeV}}{M_{\rm I}}), \tag{3.70}$$

where $M_{\rm I}$ is the energy scale associated with the plateau, and therefore also with the onset of inflation at $t_{\rm I}$. This means that the potential energy in some patch is sure to be dominated by the gradient and kinetic energy energies of the inflaton field. This will block the onset of inflation. More quantitatively, to start off inflation we need a homogeneous domain at the Planck time of radius r that grows into Hubble-sized volume by $t_{\rm I}$:

$$\Big(\frac{a(t_{\rm I})}{a(t_{\rm Pl})}\Big)r(t_{\rm Pl}) \geq \frac{1}{H(t_{\rm I})}. \tag{3.71}$$

This places a lower bound on the initial homogeneous volume of

$$r^3(t_{\rm Pl}) \geq \Big(\frac{a(t_{\rm Pl})}{a(t_{\rm I})H(t_{\rm I})}\frac{H(t_{\rm Pl})}{H(t_{\rm Pl})}\Big) \sim 10^9 \cdot \Big(\frac{10^{16}\text{GeV}}{M_{\rm I}}\Big)^3 H^{-3}(t_{\rm Pl}). \tag{3.72}$$

In other words, the onset of inflation requires at least a billion homogeneous Hubble volumes.

To make matters worse, Ijjas et al. argue that plateau-like inflation is unlikely by its own standards. To see this, consider the basic plateau-like potential

$$V(\phi) = \lambda(\phi^2 - \phi_0^2)^2. \tag{3.73}$$

In Figure 3.1, I have drawn such a potential. From the shape of the potential it is clear that inflation can occur in two ways. Either it slowly rolls down the plateau towards the local minimum, or it slowly rolls down the power-law side of the potential (imagine a slow-roll domain high up on the curve, outside the window of Figure 3.1). Plateau-like inflation will take place for values of ϕ less than the local minimum, which has energy of order $M_{\rm Pl}$. So

$$\Delta\phi(\text{plateau}) : \phi \lesssim M_{\rm Pl}. \tag{3.74}$$

Power-law inflation will take place for values of ϕ for which $V(\phi)$ is less than the Planck density $M_{\rm Pl}^4$, for which the gradient and kinetic terms of the inflaton field dominate. Since λ must be of order 10^{-15} to explain the large-scale density perturbations, this gives

$$\Delta\phi(\text{power-law}) : \phi \lesssim 10^3 M_{\rm Pl}. \tag{3.75}$$

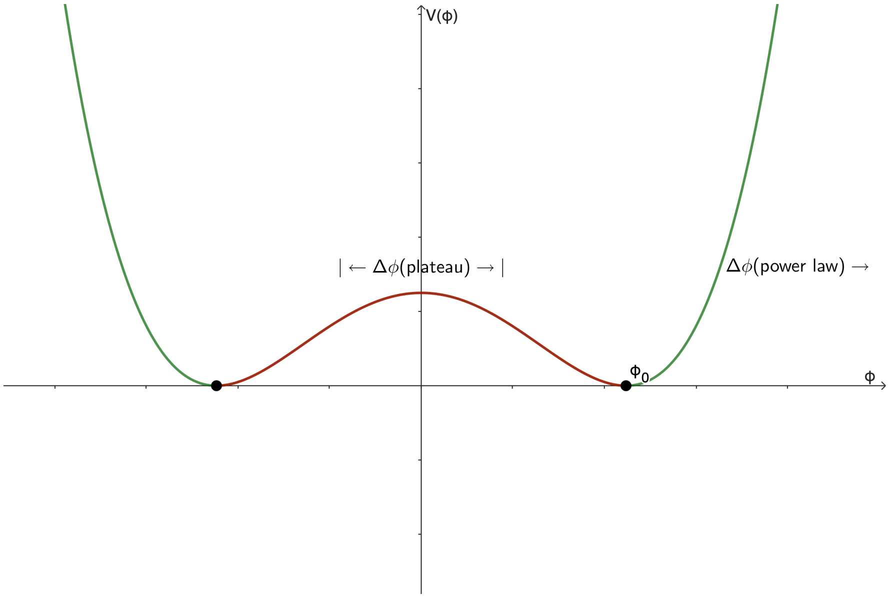


Figure 3.1: A basic plateau-like inflationary potential. The red part of the curve is the plateau-like part of the potential, the green part of the curve matches a power-law potential $V(\phi) = \lambda\phi^4$. The minima of the potential are $-\phi_0$ and ϕ_0.

The conclusion is that there is a much larger range of values for ϕ for which power-law inflation will take place (which is disfavoured by the Planck data) than for which plateau-like inflation will take place.

A vehement reply by Guth et al. follows (Guth, Kaiser, and Nomura 2014). Their response to the initial-conditions problem is twofold. First, they argue that Ijjas et al. make incorrect assumptions in their estimate. Guth et al. propose the following rectification. Assume a region of negative spatial curvature. This is plausible, because zero-curvature regions are special, and positive curvature regions will quickly recollapse unless the curvature is very close to 0. Since the curvature term in the Friedmann equation scales as a^{-2}, the gradient energy term scales as a^{-2}, and the kinetic energy term scales as a^{-6}, the a^{-2} terms will quickly dominate. Solving the Friedmann equation gives $a(t) \propto t$. Again, consider an initial region of homogeneity with radius r at the Planck time. In the worst case, inhomogeneities would travel inwards from the boundary of the homogeneous region with the speed of light. So, by the time of the onset of inflation t_{I}, we would have

$$\frac{r(t_{\mathrm{I}})}{a(t_{\mathrm{I}})} = \frac{r(t_{\mathrm{Pl}})}{a(t_{\mathrm{Pl}})} - \int_{t_{\mathrm{Pl}}}^{t_{\mathrm{I}}} \frac{1}{a(t)} dt. \tag{3.76}$$

Requiring the size of the homogeneous region to be one Hubble volume $r(t_{\mathrm{I}}) = H^{-1}(t_{\mathrm{I}})$, as Ijjas et al. do, and using that $a(t) \propto t$, so that $H(t) = a^{-1}(t)$, we can rewrite this equation as

$$r(t_{\mathrm{Pl}}) = \frac{1}{H(t_{\mathrm{Pl}})} \left(1 + \log\left(\frac{H(t_{\mathrm{Pl}})}{H(t_{\mathrm{I}})}\right)\right). \tag{3.77}$$

Using the relation

$$H^2(t) = \frac{8\pi}{3} M_{\mathrm{Pl}}^2, \tag{3.78}$$

together with the Planck 2013 constraint that

$$H(t_{\mathrm{I}}) < 3.7 \cdot 10^5 M_{\mathrm{Pl}}/\sqrt{8\pi}, \tag{3.79}$$

we get that

$$r^3(t_{\mathrm{Pl}}) > 2.6 \cdot 10^3 H^{-1}(t_{\mathrm{Pl}}), \tag{3.80}$$

which, even in this worst case scenario, is significantly less than what Ijjas et al. obtain.

Second, Guth et al. argue that the entire argument of Ijjas et al. is stated on false grounds. The problem is that Ijjas et al. make the crucial assumption that there is a single, uninterrupted inflationary phase stretching all the way from the Planck time until the last 60 e-folds—referred to by Guth et al. as "observable inflation". In other words, as the inflaton rolls down the potential, it encounters no local minima but the final one essential to reheating. Guth et al. claim that a potential with such complicated features is to be expected only at high energies, in light of developments in particle physics and string theory. In their view, the high-energy sector of inflation can have all sorts of properties that will not affect observable inflation in the sense that it changes the predictions for the values of the parameters of the density spectrum.

The unlikeliness problem is countered in a similar fashion. Guth et al. grant that the argument is convincing, given that we start out with a single scalar field and arbitrary initial conditions. If inflation is the result of a quantum-tunneling event, or if there are multiple fields at play, then the situation is different. So again, the gist of the counterargument is that inflation at non-observable high-energy scales could be more complex than is assumed by Ijjas et al.

The dust has not settled yet: the debate is alive and kicking, albeit in a more numerical direction (Aurrekoetxea et al. 2020) (Garfinkle, Ijjas, and Steinhardt 2023) (Elley et al. 2024). Although I am not in a position to judge who is right and who is wrong here, I do want to make an observation based on the above arguments: notice that the core theme of the debate is *naturalness*. What are natural assumptions to make about the state of the universe prior to inflation? Chaotic initial conditions are introduced for their genericity, but are shown to lead to the special (unnatural) requirement of a large homogeneous patch on which inflation can spark off. How large should this patch be? Again, the estimates made by Ijjas et al. make different assumptions of what is natural vis-à-vis those of Guth et al. First, Guth et al. assume that the initial region has negative curvature on the grounds that zero curvature would be *special*, i.e. unnatural. Second, Ijjas et al. assume the the inflationary potential is simple and featureless, which to them is natural. To Guth et al. a complicated potential, being the result of complex high-energy processes, is a natural one.

Is it natural that the kinetic, gradient, and potential energies of the inflaton field are of the same order prior to inflation? This, too, is taking place in an extremely high-energy background. Why should the distribution of energy types not be far more complicated than this, for the same reasons why the shape of the inflaton potential must be complicated? The arguments of Ijjas et al. show that the chaotic initial conditions and a simple inflaton potential together require special initial conditions for the onset of inflation. Rejecting simple potentials based on our ignorance of extremely high-energy physics, as Guth et al. do, but taking the simple chaotic initial conditions for granted, amounts to applying a double standard. Biting the bullet and rejecting the simple chaotic picture too, however, would spell trouble for the onset of inflation again. In fact, the chaotic initial conditions were imposed to address the concern that inflation may not start off anywhere. Without assuming such conditions, the problem of initial conditions for inflation returns.

Can we really say what conditions are natural at energy domains far beyond our experimental capacities? Why are complicated potentials, chaotic conditions, or non-flatness natural? To give any serious answers to these questions, one would need a measure. It is well known that constructing an appropriate measure is highly problematic (Earman and Mosterin 1999, pp. 31–34). In absence of a solution to the measure problem, it seems to me that there will be no unequivocal winner in the initial conditions debate. Each side makes strong assumptions on which states of affairs are natural (generic), and which are unnatural (special). As the debate shows, these assumptions lead to wildly different conclusions. Especially given the more numerical direction the debate has recently taken, I think it is vital that assumptions on naturalness are made explicit. The empirical inaccessibility of their domain of application demands caution in their use.

Despite any clear conclusions on the extent to which the initial-conditions problem impairs the inflationary explanation of uniformity, I can say the following. Consider again the two explanatory strategies for the uniformity problem. An advocate of the initial-conditions explanation wants to argue that inflation itself requires special initial conditions. If so, the inflationary explanation can be accused of hypocrisy: it rejects a special (uniform) initial state of the universe and proposes a dynamical mechanism to smooth out non-uniformities, but this very mechanism requires special initial conditions itself. Then, there is sufficient reason to jettison inflation and simply assume a special initial universe from the outset. This is what is at stake in the above debate. I have tried to show that each side of the debate is motivated by different ideas of what is natural. Without an objective standard of naturalness, one cannot make proper judgements about the claims made on either side. But, by extension, one then also cannot judge the special initial-conditions explanation to be special.[5]

3.4 Conclusions on the uniformity problem

Concisely stated, the uniformity problem refers to the incompatibility of a non-uniform initial universe and the absence of a causal smoothing mechanism before the time of recombination. The initial-conditions explanation of the uniformity problem rejects the non-uniformity of initial conditions, whereas the dynamical explanation postulates the existence of the required smoothing mechanism. The success of inflation as a candidate mechanism derives largely from Wald's homogeneous no-hair theorem, which I have extensively discussed. Although the theorem is originally proven not with reference to inflation but to a cosmological constant, the proof can be adapted to comply with general inflationary dynamics. An important ingredient of the proof is the dominant energy condition. In Chapter 6, I will return to the energy conditions.

Over the past decade, there has been considerable debate regarding the extension of the no-hair theorem to inhomogeneous initial conditions. The lesson of this debate is that chaotic initial conditions require that either the inflationary potential is featureless, or inflation itself requires a large homogeneous patch to start. To date, no consensus has been reached, as the disagreement on what background assumptions are natural or not persists. The conclusion that I

[5] The dynamical explanation can insist on robustness, however. I will discuss robustness and fine-tuning in due course.

draw from this is that either explanation is adequate, yet motivated by different intuitions about what it means for a state of the universe to be special. A different conception of naturalness may lead one to favour one explanation over the other.

Chapter 4 $\Longrightarrow$

The entropy problem

The topic of the present chapter is the entropy problem. In short, given that thermodynamical entropy is always future-increasing, supported by the non-observation of white holes, the initial state of the universe must have been one of extremely low entropy. This means that the universe started off in an incredibly special state. Note the similarities between this problem and the initial-conditions problem of the previous chapter. One may ask: why is the burden on inflation to make sense of this? The answer is that it is not, but since the initial-conditions explanation can easily make sense of the low-entropy initial universe, the inflationary explanation better not lag behind. Indeed, the dominant contribution to the total entropy is the *gravitational* entropy; a uniform state is one of extremely low gravitational entropy. The initial-conditions explanation is placed on firmer footing in this chapter. I will again analyse a debate, this time on the reversibility argument against inflation. Being forced to spell out their positions, the two rivaling explanations will admit to be more sharply defined here.

Earman and Mosterín discuss the entropy problem, albeit in a very rudimentary form:

> [U]nder the inflationary scenario a generic non-uniform state would be carried under the deterministic evolution of the coupled Einstein-scalar field equations into a currently uniform state. But whatever measure we use to gauge genericity of Cauchy data, it should have the property that measure is preserved under dynamical evolution. Thus, under the inflationary scenario, a generic non-uniform state is carried to a uniform state, which must be generic. But (the argument goes), a presently uniform state is highly non-generic. (Earman and Mosterin 1999, p. 28)

This captures the gist of the argument, but there is more to be said about this, as we will see in a moment. In the end, however, the inflationary solution to the entropy problem stands or falls by the conformity of inflation with measure preservation. This depends on the reversibility of inflationary dynamics and its compliance with Liouville's theorem—to be articulated and proved.

4.1 Penrose's challenge

The following argument is due to Penrose (Penrose 1989). Let's assume that the universe is closed, so that there will be a big crunch after a finite amount of time. This is not essential to the argument: similar consideration apply to a universe that is infinite and forever expanding. We estimate that the amount of baryons in the observable universe is at least 10^{80}. This is a lower bound; if there are more baryons, then this strengthens the argument. We can compute the entropy of the 2.7 K blackbody background radiation to be 10^8 per baryon. So, the total entropy of the CMB in the observable universe is 10^{88}.

Next, we consider the entropy of black holes. The Bekenstein-Hawking formula tells us how to compute the entropy of a black hole:

$$S_{\mathrm{BH}} = \frac{kc^3}{4Gh}A$$

where A is the surface area of the black hole. The surface area of a black hole is proportional the square of its mass m:

$$A = \frac{16\pi G^2}{c^4}m^2$$

So, we have that

$$S_{\mathrm{BH}} = \frac{4\pi kG}{hc}m^2$$

Therefore, the more massive a black hole, the larger its entropy. Now, since we assumed the universe to be closed, we know it will collapse. The final entropy of the universe can be estimated to be approximately the same as that of a black hole containing all the baryons in the universe, which is about 10^{123}.

The entropy of a system is proportional to the logarithm of the volume of phase space that consists of all microscopic configurations that are macroscopically indistinguishable. The total phase space, corresponding the the final entropy of the universe, is then $V = 10^{10^{123}}$. The region of phase space corresponding to the time of recombination (observed in the CMB) is $10^{10^{88}}$. Since entropy increases over time, we expect the entropy of the early universe to be even lower than this, corresponding to a smaller volume W. Then,

$$\frac{W}{V} \leq 10^{10^{88}}/10^{10^{123}} \approx 1/10^{10^{123}}$$

So, the early universe must have been fine-tuned to an accuracy of 1 part in $10^{10^{123}}$ in order for us to have a universe with at least 10^{80} baryons and with the second law of thermodynamics in place.

Importantly, entropy almost always increases (and for large systems this is overwhelmingly probable) due to the statistical nature of thermodynamic processes. This is a well-established principle of thermodynamics. In a closed system, there are exponentially more microstates corresponding to equilibrium macrostates than to non-equilibrium macrostates. When particles interact, they tend to move from less probable (low entropy) microstates to more probable (high entropy) ones, leading to an overall increase in entropy. Yet these statistical arguments are time-reversible. As a result, one would expect entropy to increase towards the past, too. If entropy were to increase towards the past, however, we would observe phenomena like white holes, the time-reversed counterparts of black holes, which expel matter and decrease entropy. However, such objects are not observed in our universe, suggesting a fundamental asymmetry in the direction of time. This asymmetry, the arrow of time, indicates that while individual microscopic processes are reversible, the overall macroscopic trend is an irreversible progression towards higher entropy.

4.1.1 Graviational entropy?

The use of Bekenstein-Hawking entropy in the argument above is somewhat puzzling. In ordinary thermal processes, we expect the increase of entropy to go hand in hand with the system approaching homogeneity. Imagine a gas being pumped into the corner of a box. The gas will not remain confined to a small corner, but spread out evenly throughout the box. In the case of gravitational entropy above it seems to be the other way around. Gravitational particles, initially spread out evenly, will eventually clump together under the influence of gravity. The particles will then confine themselves to a compact, probably spherical volume.

To make sense of this, Penrose invites us to think about the Sun. Low entropy energy coming from the Sun stands at the basis of many entropy-increasing processes on Earth from which we acquire energy. The Sun itself is a hot spot in a dark and cold surrounding. This thermodynamical unevenness in the sky is the result of the gravitational clumping of the matter making up the Sun. Thus, most thermodynamical processes on Earth are actually powered by the low entropy resulting from the gravitational field rather than some thermal phenomenon. It is in this sense that Penrose suggests that we should think about entropy primarily in terms of gravitational processes. The ultimate source of all entropy-increasing processes lies in the extremely low entropy of the initial state of the universe. This must have been an extremely uniform state so as to leave plenty of room for gravitational clumping to occur, i.e. entropy to increase.

4.1.2 The special initial state

Penrose asks in what sense the incredibly small part of phase space corresponding to a low-entropy initial state of the universe is special. First, one might ask whether 'special' could mean something different now than it meant in the nascent universe. To address this, we consider Liouville's theorem. For the sake of preciseness, I will present a proof which the more philosophically inclined reader may wish to skip.

Theorem 4.1.1 (Liouville's theorem)**.** *Let (M,ω) be a symplectic manifold with* $\dim(M) = 2n$*, where ω is the symplectic form. Let X_H be the Hamiltonian vector field associated with a Hamiltonian function $H : M \to \mathbb{R}$. Then, the volume form $\Omega = \frac{\omega^n}{n!}$ is preserved under the flow generated by X_H. In other words, the Lie derivative of the volume form Ω with respect to X_H is zero:*

$$\mathcal{L}_{X_H}\Omega = 0. \tag{4.1}$$

Proof. We start with the symplectic manifold (M,ω), where ω is a closed and non-degenerate 2-form. The volume form on this $2n$-dimensional manifold is given by

$$\Omega = \frac{\omega^n}{n!}. \tag{4.2}$$

Consider the Hamiltonian vector field X_H associated with a Hamiltonian function H. The Hamiltonian vector field is defined by the relation

$$\iota_{X_H}\omega = \omega(X_H, \cdot) = dH, \tag{4.3}$$

where ι denotes the interior derivative. The Lie derivative of the volume form Ω with respect to the vector field X_H can be computed using Cartan's magic formula:

$$\mathcal{L}_{X_H}\omega = d(\iota_{X_H}\omega) + \iota_{X_H}(d\omega) = 0, \tag{4.4}$$

using that $ddH = 0$ and that ω is closed. Since $\mathcal{L}_{X_H}$ is a derivation, the Leibniz rule gives

$$\mathcal{L}_{X_H}\Omega = n\omega^{n-1}(\mathcal{L}_{X_H}\omega) = 0. \tag{4.5}$$

This proves that the volume form Ω is invariant under the flow generated by the Hamiltonian vector field X_H, which implies that phase-space volumes are preserved during the dynamical evolution of the system. □

Since the Lie derivative of the volume form Ω with respect to the Hamiltonian vector field X_H is zero, the volume form does not change along the flow generated by this vector field. This implies that phase-space volume is preserved under the time evolution of Hamiltonian dynamics. Therefore, Liouville's theorem mathematically formalises the physical concept that phase-space volumes are invariant during the dynamical evolution of a Hamiltonian system. From this, Penrose concludes that the unlikeliness of the initial low entropy state is not some idiosyncrasy of the early universe. At any time, the state of our universe corresponds to a special volume in phase space.

Penrose proposes to explain the special initial state of the universe by means of the Weyl-curvature hypothesis (WCH). One may decompose the Riemann tensor as

$$R_{abcd} = C_{abcd} + Q_{abcd}, \tag{4.6}$$

where the tensor Q vanishes if the Ricci tensor vanishes, while the Weyl tensor C does not. An exactly uniform distribution of gravitational bodies has vanishing Weyl tensor, whereas the Weyl tensor blows up when a system gets increasingly clumped. The WCH says that the initial state of the universe has vanishing Weyl curvature. From this its extremely low (gravitational) entropy follows, and thereby also the second law of thermodynamics. It gives the universe its fundamental asymmetry, which explains the absence of white holes.

4.2 Inflation

A key goal of inflation is to explain the special initial conditions of standard cosmology. Similar to the uniformity problem, the proponent of inflation wishes to show that the extremely low entropy of the early universe is the result of whatever dynamics were playing out before that. It seems, however, that in this case this would lead to a bizarre description of the early universe, in the following sense. To avoid the pitfall of postulating new special initial conditions to explain the original ones, the inflationary account must assume a generic initial state, such as the one adopted in the chaotic inflation model. But such a generic initial state would be one of high entropy. By the above argument related to gravitational entropy, this corresponds to a situation of enormous gravitational clumping. It is the burden of inflation to furnish an entropy-lowering process, which is tantamount to an initial phase that is riddled with white holes and anti-thermodynamic behaviour. Penrose's point is that this is nonsensical. Rather, we ought to postulate law-like initial conditions, as the WCH does, to set up the very asymmetry needed for a universe obeying the second law of thermodynamics. I will make the concept of law-like initial conditions more precise in Chapter 6.

The question can be put in general terms as follows. How can something extraordinarily special follow from something especially ordinary? In order to create a special low-entropy universe, inflation itself must be a very special (read: unlikely) process. This argument can be made more precise by means of Liouville's theorem. I will introduce a similar argument by Hollands and Wald in the next section.

Albrecht tries to resolve these problems for inflation (Albrecht 2002). He argues that a generic primordial universe requires only a small patch with special initial conditions to trigger inflation. He draws the analogy with a block of ice containing an insulated box in which a gas is confined to a small corner. Because of the insulator, the gas will have time to reach an equilibrium in the box, after which it will slowly condensate on the boundary of the box, defining the initial conditions for whatever happens after this. Similarly, inflation starts off in a small, homogeneous patch of matter. The matter in this patch, much like the gas in the box, will then approach the de Sitter equilibrium state during the slow-roll phase of inflation. As the inflaton field rolls down the steep part of the potential, reheating sets in (analogous to the condensation of the gas), and the newly produced particles will be distributed uniformly in the inflated volume, thus giving rise to the desired low-entropy state. From there on, gravitational entropic processes (resembling the ice) drive the subsequent arrow of time. The crux of the argument is that inflation requires special conditions only in a small part of the universe. This is much better than Penrose's global initial low-entropy state, which concerns the entire universe. Albrecht's result has been been made more rigorous in Carroll and Tam (2010).

Proponents of initial-conditions explanations think differently. Wald (2006) is critical of the type of argument that Albrecht makes. He reiterates that the onset of inflation, even in a small patch, requires very special initial conditions. In an infinite universe, such conditions will nevertheless be realised somewhere with certainty. On this view, it is an undeniable possibility that our universe resulted from an inflating region. But, it is of course also a possibility that our universe resulted from a non-inflating region. The inflationary advocate is burdened with the task of showing that an observer in our universe is likely to see what we see, i.e. to turn possibility into prediction. Assuming an inflationary universe, what should a typical observer see? One can answer this question only relative to the fact that observers exist. In other words, if inflation is to predict what a typical observer in some region should see, then it needs to consider only regions that can have observers in the first place. Given that a region of the universe has certain features that make observers possible, inflation may then predict how likely it is that this region looks like the observable universe. Setting aside the issue that we have no idea what kind of features make observers possible, it is unclear what such a prediction can teach us. Wald concludes: "I fail to see in what sense [a calculation of the probability of having a region of the universe with observers be similar to our observable universe] would provide an "explanation" of why the observable universe is in the state we find it to be in." (Wald 2006, p. 5)

4.3 The reversibility debate

Having described the core tenets of each side of the low-entropy debate, I will now showcase one particular debate in the literature that juxtaposes the two positions. One the one side, we have Hollands and Wald (2002a) siding with an initial-conditions explanation, arguing against inflation. They put forward arguments against anthropic reasoning in inflation, much like the one by Wald that I just presented. Next to this, they unoriginally argue that inflation itself requires special conditions, and originally argue that inflation might not be necessary to explain structure formation, which I will elaborate on in the next chapter. The argument that goes to the heart of the debate is the reversibility argument. This alternative formulation of Penrose's challenge goes as follows.

Let's consider a universe that ultimately contracts into a big-crunch singularity. As we approach this big crunch, it appears highly unlikely that the matter within the universe would spontaneously transform into scalar field kinetic energy in a way that allows the scalar field to smoothly ascend a potential-energy hill and remain in near-perfect equilibrium at its peak during a long period of exponential contraction. Consequently, the subspace of the measure space corresponding to universes that did not deflate should be larger than the subspace corresponding to universes that did deflate. But the time reverse of a contracting universe that did not deflate is an expanding universe that did not inflate. This implies (by Liouville's Theorem) that the subspace of the measure space corresponding to non-inflating universes is larger than the subspace of the measure space corresponding to inflating universes. In other words, it is highly improbable that inflation will take place.

Kofman, Linde, and Mukhanov (2002) respond as follows. A key assumption in the argument of Hollands and Wald is that dynamical evolution is measure-preserving. This assumption typically holds in dynamical systems without particle production but fails in inflationary cosmology, which is an irreversible process. During inflation, the total energy and the entropy of the scalar field and the particles created by its decay are not conserved; initially, the chaotic universe contained negligible mass, but its total energy becomes exponentially large after an inflationary era. Preservation of the measure would imply that the number of particles is conserved, which is certainly not the case in inflation. The decay of the scalar field and subsequent particle production are irreversible, meaning that it is impossible to return to the initial conditions by reversing time. We would see more particles, and larger inhomogeneities, if we evolve backwards in time the state of the universe after reheating. Inflation could not be a measure-preserving process, for how could an initial Planck-size inflationary domain contain all the positions and momenta of particles of the universe it evolves into? All the particles and inhomogeneities produced by inflation are the result of quantum processes, which cannot be predicted from initial conditions.

Hollands and Wald reply to this as follows. They disagree with the claim that particle production is an irreversible process. In quantum field theory, particles correspond to modes of a field that are in excited states rather than their ground state. The irreversibility of particle creation is similar to the breaking of a coffee mug: significant, but not fundamentally altering the system's degrees of freedom. This type of irreversible dynamic is consistent with measure-preserving flows in classical statistical physics. Next, they agree with Kofman et al. that many irreversible processes have occurred throughout the universe's history. This is precisely the argument they want to make: the initial state of the universe seems to be uniform, while the final state is expected to be extremely messy, with countless black holes colliding with each other. The question is why the initial state of the universe is not messy as well. Why are there no white holes filling the universe? This must be explained by the the initial state of the universe being special.

What can we learn from the stalemate in this debate? Clearly, neither side is strictly wrong: both can provide empirically adequate accounts of the universe, and are represented by prominent physicists. The question is what fundamental assumptions drive the advocates of either side to their convictions. I distinguish two philosophical points of contention in this debate: one concerns the occurrence of non-unitary evolution, the other is the attitude towards what constitutes a good explanation.

Regarding the former, it is clear that the reversibility argument, or any version of Penrose's challenge invoking Liouville's theorem, depends on the assumption of unitarity. The question is whether inflation is a unitary process, and as such falls under the umbrella of Liouville's theorem. In the view of Hollands and Wald, the dynamical evolution of the universe under inflation is a unitary process. No new degrees of freedom are created, as is the case in any ordinary thermodynamical process. A coffee mug breaking on the floor is a dynamical process, subject to the laws of classical mechanics. Inflation is just like that. We cannot reverse it only insofar as we cannot reverse any entropy-increasing process. But it is still a unitary process. Inflation then merely passes the bucket of extremely low entropy to the pre-inflationary universe. What's more, it's unclear whether inflation can even spark off in such a state.

This picture is changed dramatically when inflation is viewed as a non-unitary process. It is the view of Kofman et al. that new degrees of freedom can be created in an expanding spacetime. They take the creation of particles by reheating to be an example of a non-unitary process. There is more quantum 'information' in the post-inflationary universe (the positions and momenta of innumerable particles) than in the pre-inflationary universe, which is just a Planck-sized patch (not able to hold the information required to specify the entire universe it will inflate into). The fact that new information arises during inflation implies the fundamental irreversibility of the process. The creation of new information comes from indeterministic quantum processes in an expanding spacetime. Particle creation is a non-unitary process, like wavefunction collapse, which is non-reversible.

To Hollands and Wald, inflation must occur in a closed system subject to dynamical laws, so that its degrees of freedom do not change, and the whole process is a unitary one. At the root of this view lies a notion of entropy that coincides with the standard one from statistical mechanics. The pre-inflationary patch and the universe that it becomes are just different configurations of the same system. On the other hand, Kofman et al. take a more quantum-mechanical perspective on entropy, akin to the Von Neumann interpretation of entropy. They aim to show that new quantum-mechanical information arises in inflation, which cannot occur in unitary processes. On the basis of this lies the idea that particle creation is a non-unitary quantum process. This, however, is contingent on one's interpretation of quantum mechanics. The Everettian, for example, holds that the wavefunction is objectively real, so that all quantum processes are unitary. The point I want to make is that the differences between the inflationary and non-inflationary sides in this debate partly stem from different attitudes toward unitarity. These attitudes are strongly derivative on conceptions of entropy and beliefs about the foundations of quantum mechanics. Finding appropriate definitions of entropy is still an active topic in research, and it is well known that one's preferred interpretation of quantum mechanics is largely a matter of explanatory taste. On these grounds, it is fair to say that neither side in the debate on the reversibility argument is wrong. Both positions are perfectly tenable in virtue of the plurality of definitions of entropy and attitudes towards fundamental quantum processes.

Apart from the question of unitarity, there seems to be another philosophical point of contention in the debate, having to do with views on explanation. In the account of Hollands and Wald, the explanandum is the second law of thermodynamics, and the explanans is a theory of initial conditions. Why do all thermal processes that we observe go from low to high entropy? The answer, extrapolating these observations, is the special low-entropy initial state of the universe. This explanation is criticised for not being an explanation at all. In the account of Kofman et al., the explanandum is the uniform low-entropy state of the universe given that its initial conditions must be generic rather than special, and the explanans is the second law of thermodynamics together with the additional dynamical mechanism of inflation. Why is our universe in a special state? The answer is that a small patch in the generic (high-entropy) pre-inflationary universe randomly fluctuated into a special (low-entropy) state, triggering inflation. The inflationary mechanism (dynamical and entropy-increasing) then gives the desired uniform special state of the universe in the way described by Albrecht. But, in this account, the second law of thermodynamics is left unexplained, because this is the result of the initial, very special fluctuation that causes inflation in the first place. There is thus a new challenge created, namely to explain the special initial conditions required for inflation. The discussion then proceeds to revolve around the question whether this new challenge is better than Penrose's original challenge or not. The discussion then reduces to that of the previous chapter.

So, on the one hand, we have an explanation that is allegedly not an explanation at all. It conjures up law-like initial conditions, which have no content over and above an extrapolation of the second law of thermodynamics back in time. On the other hand, we have an explanation that invokes a novel dynamical mechanism unlike anything we have ever seen. Although this explanation can make sense of the current special state of the universe, it cannot account for the second law of thermodynamics unless it answers the question why the conditions for inflation could ever occur in a messy high-entropy primordial universe.

4.4 Alternative fundamental law-like initial conditions

I have already mentioned one example of law-like initial conditions that tries to do better than just postulating an extremely low-entropy state of the early universe. The Weyl-curvature hypothesis makes the assumption that the Weyl tensor initially vanishes, from which the universe's low-entropy follows. There are many examples of cosmological models which satisfy the Weyl-curvature hypothesis, see Tod (2010). The question remains, of course, whether we are somehow better off if we impose the Weyl-curvature hypothesis than if we just postulate an extremely low-entropy initial state. It seems to me that the two are one and the same, in the sense that they imply each other. So, what does the initial-conditions explanation win except for a new name?

An alternative proposal is made in Lehners and Stelle (2019). This paper operates within the framework of quadratic gravity. While general relativity needs to be quantised but is non-renormalisable, gravity becomes renormalisable when quadratic curvature terms are included. Evidence suggests that quadratic gravity might be asymptotically safe, meaning it remains valid up to (and even beyond) Planck-scale energies. This makes it particularly relevant in studying the universe at its earliest stages. Whereas in general relativity solutions with inhomogeneous and anisotropic conditions at the singularity are allowed, in quadratic gravity such conditions cause the action to blow up. As a result, spacetimes with these irregularities at the initial singularity are ruled out, favouring a homogeneous and isotropic start. By imposing the fundamental law-like assumption that the action must be finite, homogeneous and isotropic spacetimes (with, accordingly, very low entropy) are selected as physical.

The exclusion of infinite-valued actions, however, also entails that the universe should undergo an accelerated expansion. This sets the stage for an inflationary phase in the early universe. This is an interesting case for this thesis: in trying to provide an initial-conditions explanation for the universe by excluding non-uniform spacetimes, the quadratic-gravity model also accommodates an inflationary era. Note also that modifying gravity amounts to a dynamical change of strategy. Perhaps prematurely, this suggests that any full explanation of the early universe must be a hybrid one, combining law-like initial conditions and dynamical mechanisms. I will say more on this in Chapter 6.

One last alternative explanation invoking law-like initial conditions I wish to mention here is the CPT-symmetric universe (Boyle, Finn, and Turok 2018). In such a universe, the big-bang singularity acts like a CPT mirror, in the sense that there was a CPT-reversed universe 'before' the big bang. The authors claim that this model can possibly explain structure formation, the low-entropy of the universe, and dark matter, see Boyle, Finn, and Turok (2022). More research needs to be done in this direction to make any definitive conclusions on these matters.

4.5 Conclusions on entropy

The low entropy of the early universe is something that requires explanation, so as to save all thermodynamical phenomena. On the one hand, the initial-conditions explanation imposes this state as a law-like initial condition. Simply postulating low entropy, or equivalently, assuming the Weyl-curvature hypothesis to hold true, may seem like a non-explanation. I return to this issue in chapter 5. Note, however, that more sophisticated arguments are available to support such an explanation, like the requirement of finite action in quadratic gravity, or the demand that the universe is CPT-symmetric. On the other hand, inflation can deliver a low-entropy state by blowing up a small homogeneous patch. If so, the inflationary explanation once more needs to address the issue of special initial conditions raised in the previous chapter. Just how special these initial conditions should be is what is at stake in the reversibility debate. If inflation is an ordinary measure-preserving process, then, by Liouville's theorem, it cannot play a role in explaining the low entropy of the early universe. This is precisely the point of contention. I suggest that one's verdict on the question whether inflation is entropy-preserving depends on one's commitments with respect to interpretations of entropy and the foundations of quantum mechanics. I also claim that the sides of the reversibility debate are shaped by their stance on what constitutes a good explanation. Mainly, the initial-conditions explanation seems to amount to giving up, whereas the dynamical explanation must introduce new exotic physics, and cannot explain the second law of thermodynamics unless it can make sense of the initial-conditions problem. I recall from the previous chapter that the debate on special initial conditions for the onset of inflation is motivated by different attitudes towards what physical states and dynamics are natural. In Chapter 6, I will zoom in on these philosophical commitments that are gleaming in the background.

Chapter 5

The problem of structure formation

So far, I have mainly been concerned with the explanation of the uniformity of the CMB spectrum. Although this spectrum is indeed uniform to a very high degree, it still contains tiny variations. These provide loads of information about our universe: the fluctuations of the CMB effectively give us a map of a part of the universe. They show the distribution of energy density in the early universe, from which galaxy clusters have formed through gravitational clumping. The density perturbations of the CMB have certain statistical properties. In this chapter, I will focus on the (near) scale invariance of the spectrum of perturbations. The prediction of scale invariance is often proclaimed to be a major empirical success of the theory of inflation. I remark that are other relevant statistical properties, such as the phase coherence and Gaussianity of the perturbation modes, which I wish to examine in future research.

In absence of accurate data of the spectrum of density perturbations, Earman and Mosterín expressed themselves succinctly and conservatively: "[a] confirmation of the inflationary predictions together with a disconfirmation of alternative explanations of the density perturbations would constitute strong evidence that inflation did play an important role in the early universe [...]." (Earman and Mosterin 1999, p. 41) These inflationary predictions have indeed been confirmed, most recently by the results of the Planck mission. The disconfirmation of alternative explanations will feature centrally in the present chapter. I will first present a rigorous proof of the inflationary prediction of scale invariance. Second, I will analyse the assumptions of this proof, from which I will conclude that inflation indeed gives an attractive explanation for structure formation. Some suggestions have been made, however, that the adequacy of the inflationary explanation rests on specific initial conditions, and so the leitmotif of this thesis resurfaces. Finally, I examine some non-inflationary explanations of structure formation (i.e. scale invariance). I argue that the proposed non-inflationary models have promising features, but need to be developed further. Remarkably, the explanatory adequacy of one such model depends on the presence of an additional dynamical mechanism. This suggests again a blurring of the explanatory dichotomy introduced in this thesis.

5.1 The inflationary argument for structure formation

Intuitively, inflation produces the fluctuations in the CMB spectrum by 'blowing up' quantum fluctuations in the inflaton field. The early universe is dominated by the inflaton field ϕ. Locally, $\phi(t)$ specifies the evolution of inflationary expansion, described by the potential of the field. The uncertainty principle then tells us that the field has spatial fluctuations, as the inflaton is a quantum-mechanical object. Different regions of space will inflate by different amounts, resulting in differences in density after inflation. These density perturbations can be measured via the temperature spectrum of the CMB.

5.1.1 What is scale invariance?

Let me first try to make clear what scale invariance means, following McCabe (2018). Consider a Newtonian gravitational potential field Φ with small fluctuations. Its fluctuation field can be expressed as a Fourier transform

$$\delta_\Phi := \Phi - \langle\Phi\rangle = \frac{1}{(2\pi)^3}\int A(\mathbf{k})\, e^{i\,\mathbf{x}\cdot\mathbf{k}}\, d^3\,\mathbf{k}. \tag{5.1}$$

Here $A(\mathbf{k})$ is the amplitude of the mode with wavenumber $|\mathbf{k}|$. Since $\langle\delta_\Phi\rangle = 0$, we can express the variance as

$$\sigma_\Phi^2 = \langle\delta_\Phi^2\rangle = \frac{1}{(2\pi)^3}\int |A(\mathbf{k})|^2\, d^3\,\mathbf{k}. \tag{5.2}$$

Since we assume the background field to the perturbations to be homogeneous and isotropic, we have spherical symmetry about every point, so the expression of the variance simplifies to

$$\langle\delta_\Phi^2\rangle = \frac{1}{(2\pi)^3}\int_0^\infty |A(k)|^2 4\pi k^2\, d\,k. \tag{5.3}$$

The power spectrum is defined by

$$P_\Phi(k) := |A(k)|^2. \tag{5.4}$$

It describes the contribution of a mode $\mathbf{k}$ to the variance in the fluctuation spectrum of Φ. Inflation predicts a scale-invariant spectrum

$$P_\Phi(k) \propto k^{-3}. \tag{5.5}$$

Why such a power spectrum is called scale-invariant can be seen by rewriting Equation 5.3 as a logarithmic integral

$$\langle\delta_\Phi^2\rangle = \frac{1}{2\pi^2}\int_{-\infty}^\infty P_\Phi(k) k^3\, d\log k. \tag{5.6}$$

Now, define

$$\Delta(k) := \frac{1}{2\pi^2} P_\Phi(k) k^3, \tag{5.7}$$

which represents the contribution to the total variance per unit logarithmic interval. If Equation 5.5 holds, then $\Delta(k)$ is independent of k. This explains the term scale invariance.

Proving that the power spectrum P_ϕ of the inflaton field ϕ has the property that $P_\phi \propto k^{-3}$, implies that the power spectrum of the spatial curvature perturbations has the same form. This, in turn, is a substitute for the power spectrum of the perturbations of the metric, which is approximated by the Newtonian gravitational potential. So, showing that inflation predicts $P_\phi \propto k^{-3}$ is equivalent to showing that inflation predicts the scale-invariant power spectrum of the CMB.

5.1.2 The proof

This proof follows that of McCabe (2018) and Baumann (2012). I try to be as rigorous as possible to hedge my thesis against the accusation of not giving inflation a fair hearing, but I will omit the details of computations in linearised gravity for the sake of clarity. For details, which are tedious and hardly contain any interesting mathematics, I refer to Baumann (2012). We assume slow-roll inflation described by the following action

$$S = \int \mathrm{d}\,t\,\mathrm{d}\,\mathbf{x}^3 \sqrt{-g}\big(\frac{1}{2}R - \frac{1}{2}g^{\mu\nu}\partial_\mu\phi\partial_\nu\phi - V(\phi)\big). \tag{5.8}$$

The proof takes place in the framework of linearised gravity. We assume an FLRW background metric with small perturbations, both in the inflaton field and the metric

$$\phi(t,\mathbf{x}) = \phi(t) + \delta\phi(t,\mathbf{x}), \qquad g_{\mu\nu}(t,\mathbf{x}) = g_{\mu\nu}(t) + \delta g_{\mu\nu}(t,\mathbf{x}). \tag{5.9}$$

Importantly, this split into background and perturbations that live on top of it is not unique. There is no preferred choice of coordinates in an inhomogeneous universe. In making a choice in how to slice up spacetime into a foliation of hypersurfaces of constant time, one is also implicitly defining the perturbations. This is why we consider both perturbations in the metric and in the inflaton field, which are interchangeable via coordinate transformations. Specifically, we are interested in perturbations that are *physical*, meaning that they cannot be eliminated by means of a change of coordinates.

The most general form in which we can write the perturbed spacetime interval is

$$\mathrm{d}\,s^2 = -(1+2\Phi)\,\mathrm{d}\,t^2 + 2aB_i\,\mathrm{d}\,x^i\,\mathrm{d}\,t + a^2\Big((1-2\Psi)\delta_{ij} + E_{ij}\Big)\,\mathrm{d}\,x^i\,\mathrm{d}\,x^j, \tag{5.10}$$

where Ψ, Φ, B_i, and E_{ij} are functions of space and time. The spacetime interval is invariant under coordinate transformations. This gives a freedom of gauge, which we can exploit to simplify calculations. Our main interest is a gauge-invariant quantity called the comoving curvature perturbation

$$\zeta := \Psi + \frac{H}{\dot{\phi}}\delta\phi. \tag{5.11}$$

We work in the comoving gauge: a foliation of hypersurfaces in which the fluctuations of ϕ vanish ($\delta\phi = 0$). In this gauge, any perturbation is now described as a metric perturbation

$$\delta g_{ij} = a^2(1-2\zeta)\delta_{ij} + a^2h_{ij}. \tag{5.12}$$

So, non-zero perturbations in the 3-curvature of the hypersurfaces do not vanish in the comoving gauge. Since ζ is gauge invariant, and since there also exists a spatially flat gauge in which $\Psi = 0$, we have

$$\zeta = \frac{H}{\dot{\phi}}\delta\phi. \tag{5.13}$$

Next, we can solve the Einstein equations to obtain expressions for δg_{00} and δg_{0i} in terms of ζ. Substituting this into the action, expanding in terms of ζ, and dropping terms that are above linear order we obtain

$$S = \int \mathrm{d}\,t\,\mathrm{d}\,\mathbf{x}^3 a^3\frac{\dot{\phi}^2}{H^2}\big(\dot{\zeta}^2 - a^{-2}(\partial_i\zeta)^2\big). \tag{5.14}$$

Define the Mukhanov variable as

$$v := z\zeta, \qquad \text{with } z := \frac{a\dot{\phi}}{H}. \tag{5.15}$$

Conformal time is given by

$$\tau = -\int_t^\infty \frac{1}{a(t)} dt = -\int_t^\infty \mathrm{e}^{-Ht}\, dt = -\frac{1}{aH}, \tag{5.16}$$

where we take inflationary expansion to be that of de Sitter space: $a(t) = \mathrm{e}^{Ht}$. Rewriting the action in terms of the Mukhanov variable and conformal time gives

$$S = \frac{1}{2}\int \mathrm{d}\,\tau\,\mathrm{d}\,\mathbf{x}^3 \big((v')^2 - (\partial_i v)^2 + \frac{z''}{z}v^2\big), \tag{5.17}$$

where primes indicate differentiation with respect to conformal time. Varying the action gives the Mukhanov–Sasaki equation

$$v''_{\mathbf{k}} + \omega_k^2 v_{\mathbf{k}} = 0, \qquad \omega_k^2 := k^2 - \frac{z''}{z}, \tag{5.18}$$

where $v_{\mathbf{k}}$ represent the Fourier modes

$$v(\tau, \mathbf{x}) = \frac{1}{(2\pi)^{2/3}} \int v_{\mathbf{k}}(\tau)\, \mathrm{e}^{\mathrm{i}\,\mathbf{k}\cdot\mathbf{x}}\, \mathrm{d}^3\,\mathbf{k}. \tag{5.19}$$

In de Sitter space, the Mukhanov–Sasaki equation reduces to

$$v''_{\mathbf{k}} + \big(k^2 - \frac{2}{\tau^2}\big)v_{\mathbf{k}} = 0. \tag{5.20}$$

We note that for modes with sub-horizon wavelengths, $k^2 \gg |\frac{z''}{z}|$,[1] we get

$$v''_{\mathbf{k}} + k^2 v_{\mathbf{k}} = 0, \tag{5.21}$$

which has oscillating solutions $v_{\mathbf{k}} \propto \mathrm{e}^{\pm\,\mathrm{i}\,k\tau}$. For modes with super-horizon wavelengths, $k^2 \ll |\frac{z''}{z}|$, we have

$$\frac{v''_{\mathbf{k}}}{v_{\mathbf{k}}} = \frac{z''}{z}. \tag{5.22}$$

Solving this equation gives $v_{\mathbf{k}} \propto z$. This describes ‘frozen’ perturbations

$$\zeta_{\mathbf{k}} = \frac{1}{z} v_{\mathbf{k}} \propto 1. \tag{5.23}$$

[1] Indeed, in de Sitter space

$$k^2 \gg |\frac{z''}{z}| \iff k^2 \gg |\frac{2}{\tau^2}| \iff |k\tau| \gg 1 \iff |\frac{k}{aH}| \gg 1 \iff \frac{1}{|k|} \ll |\frac{1}{aH}|,$$

so the condition in the main text is equivalent to the condition that the wavelength is much smaller than the comoving Hubble horizon.

The general solution to the Mukhanov–Sasaki equation (Eq. 5.18) is given by

$$v_{\mathbf{k}}(\tau) = a^-_{\mathbf{k}} v_k(\tau) + a^+_{-\mathbf{k}} v^*_k(\tau), \tag{5.24}$$

because the frequency $\omega_k(\tau)$ depends only on k. The solutions are normalised by the condition that

$$\mathrm{Im}(v'_k v^*_k) = 1. \tag{5.25}$$

The time-independent integration constants can be written as

$$a^-_{\mathbf{k}} = \frac{v^{*\prime}_k v_{\mathbf{k}} - v^*_k v'_{\mathbf{k}}}{v^{*\prime}_k v_k - v^*_k v'_k}, \qquad a^+_{\mathbf{k}} = (a^-_{\mathbf{k}})^*. \tag{5.26}$$

Putting this into Eq. 5.19 gives

$$v(\tau, \mathbf{x}) = \frac{1}{(2\pi)^{2/3}} \int \left(a^-_{\mathbf{k}} v_k(\tau)\, \mathrm{e}^{\mathrm{i}\,\mathbf{k}\cdot\mathbf{x}} + a^+_{\mathbf{k}} v^*_k(\tau)\, \mathrm{e}^{-\,\mathrm{i}\,\mathbf{k}\cdot\mathbf{x}} \right) \mathrm{d}^3\,\mathbf{k}. \tag{5.27}$$

Next, we apply canonical quantization. This entails the promotion of v and its conjugate momentum $\pi = v'$ to quantum operators which satisfy the canonical commutation relations. The constants of integration $a^-_{\mathbf{k}}$ and $a^+_{\mathbf{k}}$ in Eq. 5.27 then also become operators. This gives

$$\hat{v}(\tau, \mathbf{x}) = \frac{1}{(2\pi)^3} \int \left(\hat{a}^-_{\mathbf{k}} v_k(\tau) + \hat{a}^+_{-\mathbf{k}} v^*_k(\tau) \right) \mathrm{e}^{\mathrm{i}\,\mathbf{k}\cdot\mathbf{x}}\, \mathrm{d}^3\,\mathbf{k}. \tag{5.28}$$

The commutation relations for v and π then imply that

$$[\hat{a}^-_{\mathbf{k}}, \hat{a}^+_{\mathbf{k}'}] = \delta(\mathbf{k} - \mathbf{k}'), \qquad [\hat{a}^-_{\mathbf{k}}, \hat{a}^-_{\mathbf{k}'}] = 0 = [\hat{a}^+_{\mathbf{k}}, \hat{a}^+_{\mathbf{k}'}]. \tag{5.29}$$

Then, $a^-_{\mathbf{k}}$ and $a^+_{\mathbf{k}}$ are interpreted as the annihilation and creation operators. The vacuum state is defined as the quantum state satisfying

$$\hat{a}^-_{\mathbf{k}} \left|0\right\rangle = \left|0\right\rangle. \tag{5.30}$$

Other states can be constructed through repeated application of the creation operator. The non-uniqueness of the vacuum state will be the topic of next two sections. For now, we continue the derivation and assume that the mode functions are those corresponding to the Minkowski mode functions in the asymptotic past:

$$\lim_{\tau \to -\infty} v_k(\tau) = \frac{1}{\sqrt{2k}}\, \mathrm{e}^{-\,\mathrm{i}\,k\tau}. \tag{5.31}$$

(This is the Bunch–Davies condition—more on this in the next section.) The general solution to the Mukhanov–Sasaki equation in de Sitter space (Eq. 5.20) is given by

$$v_k(\tau) = \alpha \frac{\mathrm{e}^{-\,\mathrm{i}\,k\tau}}{\sqrt{2k}} (1 - \frac{\mathrm{i}}{k\tau}) + \beta \frac{\mathrm{e}^{\mathrm{i}\,k\tau}}{\sqrt{2k}} (1 + \frac{\mathrm{i}}{k\tau}). \tag{5.32}$$

From Eq. 5.31 we get that $\alpha = 1$ and $\beta = 0$, so that there is a unique mode function given by

$$v_k(\tau) = \frac{\mathrm{e}^{-\,\mathrm{i}\,k\tau}}{\sqrt{2k}} (1 - \frac{\mathrm{i}}{k\tau}). \tag{5.33}$$

We consider the super-horizon limit, where $|k\tau| \ll 1$. Then, since τ takes on negative values only, we compute the limit

$$\lim_{|k\tau|\to 0} v_k(\tau) = \frac{1}{\mathrm{i}\sqrt{2}} \frac{1}{k^{3/2}\tau}. \tag{5.34}$$

From Eq. 5.28, the considerations from the previous section, and the assumption that the quantised field $\hat{v}$ was initially in its ground state, we can find the power spectrum of the field $\hat{v}$ by computing:

$$\begin{aligned}
\langle|\hat{v}|^2\rangle &= \langle 0|\, \hat{v}(\tau,\mathbf{0})\hat{v}(\tau,\mathbf{0})\, |0\rangle \\
&= \int \frac{\mathrm{d}^3\,\mathbf{k}}{(2\pi)^{3/2}} \int \frac{\mathrm{d}^3\,\mathbf{k}'}{(2\pi)^{3/2}} \langle 0|\, (\hat{a}^-_{\mathbf{k}} v_k + \hat{a}^+_{\mathbf{k}} v^*_k)(\hat{a}^-_{\mathbf{k}'} v_{k'} + \hat{a}^+_{\mathbf{k}'} v^*_{k'})\, |0\rangle \\
&= \int \frac{\mathrm{d}^3\,\mathbf{k}}{(2\pi)^{3/2}} \int \frac{\mathrm{d}^3\,\mathbf{k}'}{(2\pi)^{3/2}} v_k v^*_{k'}\, \langle 0|\, \hat{a}^-_{\mathbf{k}} \hat{a}^+_{\mathbf{k}'}\, |0\rangle \\
&= \int \frac{\mathrm{d}^3\,\mathbf{k}}{(2\pi)^{3/2}} \int \frac{\mathrm{d}^3\,\mathbf{k}'}{(2\pi)^{3/2}} v_k v^*_{k'}\, \langle 0|\, [\hat{a}^-_{\mathbf{k}} \hat{a}^+_{\mathbf{k}'}]\, |0\rangle \\
&= \int \frac{\mathrm{d}^3\,\mathbf{k}}{(2\pi)^{3/2}} \int \frac{\mathrm{d}^3\,\mathbf{k}'}{(2\pi)^{3/2}} v_k v^*_{k'} \delta(\mathbf{k}-\mathbf{k}') \\
&= \frac{1}{(2\pi)^{3/2}} \int |v_k|^2\, \mathrm{d}^3\,\mathbf{k}.
\end{aligned} \tag{5.35}$$

So, we have

$$P_v(k) = |v_k|^2. \tag{5.36}$$

By Equation 5.34, we have on super-horizon scales

$$P_v = |\frac{1}{\mathrm{i}\sqrt{2}} \frac{1}{k^{3/2}\tau}|^2 = \frac{1}{2k^3\tau^2} = (aH)^2 \frac{1}{2k^3}. \tag{5.37}$$

Since the Mukhanov variable is defined as $v := z\zeta$, we have

$$P_\zeta = \frac{1}{z^2} P_v = \frac{H^2}{a^2\dot{\phi}^2} \frac{a^2H^2}{2k^3} = \frac{H^4}{\dot{\phi}^2} \frac{1}{2k^3}. \tag{5.38}$$

Using Eq. 5.13, the power spectra of Eq. 5.38 and the scalar field are related by

$$P_\zeta = (\frac{H}{\dot{\phi}})^2 P_\phi. \tag{5.39}$$

It follows that

$$P_\phi = \frac{H^2}{2k^3}. \tag{5.40}$$

Recall that we are working in the super-horizon limit. Since ζ freezes at horizon crossing, the power spectrum for each k does so, too. For each mode, the value for H is then equal to the Hubble constant H_0 during inflation. We have proven, then, that $P_\phi \propto k^{-3}$, which is what we set out to show.

As a final remark, I wish to define the scalar spectral index. To this end, define the dimensionless power spectrum as

$$\Delta^2_s(k) := \frac{k^3}{2\pi^2} P_\zeta(k). \tag{5.41}$$

The scalar spectral index characterises the degree of scale invariance, where $n_s = 1$ corresponds to perfect scale invariance:

$$n_s := 1 + \frac{\mathrm{d}\log\Delta_s^2(k)}{\mathrm{d}\log k}. \tag{5.42}$$

We can evaluate the index in terms of the slow-roll parameters using Eq. 5.40 and Eq. 5.41 as

$$n_s = 1 - 2\epsilon - \eta, \tag{5.43}$$

which is a restatement of scale invariance.

5.1.3 The choice of the physical vacuum

I will now show that the vacuum state is not uniquely determined in an FLRW universe. This is a standard issue from quantum field theory in curved spacetime; I follow the exposition by Mukhanov and Winitzki (2007). Consider again the most general solution to the Mukhanov-Sasaki equation, given by Eq. 5.24. The mode functions $v_{\mathbf{k}}(\tau)$ fully specify the creation and annihilation operators via Eq. 5.26, which in turn form the basis of all quantum states. However, the mode functions are not uniquely determined by the Mukhanov-Sasaki equation. Explicitly, we can generate new mode functions

$$u_k(\tau) = \alpha_k v_k(\tau) + \beta_k v_k^*(\tau), \tag{5.44}$$

where α_k and β_k are complex constants. These new mode functions are also normalised solutions to the Mukhanov-Sasaki equation if $|\alpha_k|^2 - |\beta_k|^2 = 1$. The functions $u_k(\tau)$ specify a different set of creation and annihilation operators $\hat{b}_{\mathbf{k}}^-$ and $\hat{b}_{\mathbf{k}}^+$, which can also be used to build a basis for the Hilbert space. Each of the annihilation operators defines a different vacuum state:

$$\hat{a}_{\mathbf{k}}^- \left|{}_{(a)}0\right\rangle = 0, \qquad \hat{b}_{\mathbf{k}}^- \left|{}_{(b)}0\right\rangle = 0. \tag{5.45}$$

Since the basis constructed by the creation and annihilation operators is used to determine how many particles a system contains, we are left with the question which set of annihilation and creation operators corresponds to observable particles. We note that the vacuum state $\left|{}_{(a)}0\right\rangle$, containing no particles in the a-basis by definition, may still contain particles in the b-basis. Is there a preferred set of mode functions that describes the real world?

If the Hamiltonian of the system at hand depends explicitly on time, there can be no time-independent vacuum state. However, at each instant of time τ_0 there is a state $\left|{}_{\tau_0}0\right\rangle$ that minimises the expectation value of the Hamiltonian. For the action we have been considering, the mode functions that select this lowest-energy state are given by

$$v_k(\tau_0) = \frac{1}{\sqrt{\omega_k(\tau_0)}}\,\mathrm{e}^{\mathrm{i}\,\omega_k(\tau)\tau_0}, \qquad v_k'(\tau_0) = \mathrm{i}\,\sqrt{\omega_k(\tau_0)}\,\mathrm{e}^{\mathrm{i}\,\omega_k(\tau_0)\tau_0}, \tag{5.46}$$

if $\omega_k^2(\tau_0) > 0$. If $\omega_k^2(\tau_0) < 0$, then the instantaneous lowest-energy vacuum does not exist. In the context of Minkowski spacetime, we have $\omega_k = k$, so Eq. 5.46 gives a time-independent vacuum state

$$v_k(\eta) = \frac{1}{\sqrt{k}}\,\mathrm{e}^{\mathrm{i}\,k\eta}. \tag{5.47}$$

Despite the ambiguity in the definition of particles in the case when ω_k is time-dependent, there are no ambiguities in physical predictions. This is because there is no unique way to separate particles and vacuum polarisation contributions in the total energy-momentum tensor, which is the only relevant source of gravitational effects. So, given a specific quantum state of the scalar field, we are perfectly able to predict specific observables, even in curved spacetime.

5.1.4 The Bunch–Davies vacuum in the de Sitter spacetime

Our goal is now to find a vacuum state in de Sitter spacetime that is physical. Again following Mukhanov and Winitzki (2007), we assume a homogeneous and isotropic universe with a positive cosmological constant Λ. The effects of this cosmological constant correspond to a perfect fluid with equation of state

$$p_\Lambda = -\rho_\Lambda. \tag{5.48}$$

Assuming flatness, the Friedmann equation has solution

$$a(t) = H_\Lambda^{-1}\,\mathrm{e}^{H_\Lambda t}, \qquad H_\Lambda = \sqrt{\frac{8\pi G}{3}\rho_\Lambda}. \tag{5.49}$$

where t denotes physical time and H_Λ is the time-independent Hubble parameter. The metric of the static maximally symmetric de Sitter spacetime, using conformal time and spherical coordinates, is given by

$$\mathrm{d}\,s^2 = \frac{1}{H_\Lambda^2\tau^2}\Big(\mathrm{d}\,\tau^2 - \mathrm{d}\,r^2 - r^2(\mathrm{d}\,\theta^2 + \sin^2\theta\,\mathrm{d}\,\varphi^2)\Big), \tag{5.50}$$

where $-\infty < \tau < 0$ and $0 \leq r < \infty$. It can be shown that these coordinates cover only half of the de Sitter spacetime. We can make the following change of coordinates:

$$\tau = \frac{\sin\tilde{\tau}}{\cos\tilde{\tau} + \cos\tilde{r}}, \qquad r = \frac{\sin\tilde{r}}{\cos\tilde{\tau} + \cos\tilde{r}}. \tag{5.51}$$

In these new coordinates, the metric is given by

$$\mathrm{d}\,s^2 = \frac{1}{H_\Lambda^2\sin^2\tilde{\tau}}\Big(\mathrm{d}\,\tilde{\tau}^2 - \mathrm{d}\,\tilde{r}^2 - \sin^2\tilde{r}(\mathrm{d}\,\theta^2 - \sin^2\theta\,\mathrm{d}\,\varphi^2)\Big), \tag{5.52}$$

where $-\pi < \tilde{\tau} < 0$ and $0 \leq \tilde{r} \leq \pi$. This metric describes a closed de Sitter universe, and the new coordinates $\tilde{\tau}$ and $\tilde{r}$ cover the entire spacetime.

Continuing in the flat coordinates, we can quantize a massive scalar field $\phi(\mathbf{x}, \tau)$ with potential $V(\phi) = \frac{1}{2}m^2\phi^2$. Plugging in the scale factor $a(\tau) = -\frac{1}{H_\Lambda\tau}$ gives

$$\omega_k^2(\tau) = k^2 - \frac{m^2}{H_\Lambda^2\tau^2} - \frac{2/(H_\Lambda\tau^3)}{1/(H_\Lambda\tau)}, \tag{5.53}$$

so that the mode functions are subject to

$$v_k'' + \Big(k^2 - (2 - \frac{m^2}{H_\Lambda^2})\frac{1}{\tau^2}\Big)v_k = 0. \tag{5.54}$$

Take a wavenumber k, and consider the early time (large negative τ) condition $k|\tau| \gg 1$. Then the physical wavelength of this mode is

$$a(\tau)k^{-1} \approx \frac{H_\Lambda^{-1}}{k|\tau|}, \tag{5.55}$$

which is much smaller than the curvature scale H_λ^{-1}. It is not surprising that the mode is not affected by the curvature of spacetime, and that we get the Minkowski solution. Indeed, in the limit $k|\tau| \gg 1$, we have $\omega_k \approx k$, and the solution to Eq. 5.54 is (see Eq. 5.47)

$$v_k(\tau) \approx \frac{1}{\sqrt{k}}\,\mathrm{e}^{\mathrm{i}\,k\tau}. \tag{5.56}$$

When $k|\tau| \approx 1$, the physical length of the modes is comparable to the curvature scale. At this time, modes cross the event horizon: sub-horizon modes satisfy $k|\tau| \gg 1$ and super-horizon modes satisfy $k|\tau| \ll 1$. Super-horizon modes do not oscillate and therefore it is not possible to define particles for such modes. Also, if $m^2 < 2H_\Lambda^2$, there is no lowest-energy state because $\omega_k^2(\tau)$ is negative. Still, there exists a preferred vacuum state called the Bunch–Davies vacuum. Fix a wavenumber k and consider the corresponding mode. As we have seen, when $k|\tau| \gg 1$ this mode does not feel much curvature, so we can demand the initial condition that

$$\lim_{\tau \to -\infty} v_k(\tau) = \frac{1}{\sqrt{\omega_k}} \mathrm{e}^{\mathrm{i}\,\omega_k \tau} . \tag{5.57}$$

This means that we choose a minimal energy state in the asymptotic past. If $m < \frac{3}{2}H\Lambda$, then the mode functions with this asymptotic are given by

$$v_k(\tau) = \sqrt{\frac{\pi|\tau|}{2}}(J_n(k|\tau|) - \mathrm{i}\,Y_n(k|\tau|)), \qquad n := \sqrt{\frac{9}{4} - \frac{m^2}{H^2}}, \tag{5.58}$$

where J_n and Y_n are the Bessel functions. The annihilation operators determined by these functions define the Bunch–Davies vacuum $|0_{\mathrm{BD}}\rangle$.

5.1.5 Conclusion on scale invariance

Let me briefly recap what we have done so far. We considered an FLRW universe with a slow-roll scalar field. We introduced linear perturbations (inhomogeneities) in both the scalar field and the metric. Choosing an appropriate gauge, we were able to confine our analysis to a perturbation in the metric. This gave us equations of motion for each perturbation mode that do not depend on fluctuations of the scalar field. Assuming that during an inflationary era the universe is approximately de Sitter (i.e. homogeneous, isotropic, flat, and with constant exponential expansion), these equations simplified. We solved these equations for the perturbation modes, but noted that these solutions are not uniquely determined. Next, we quantised the field of perturbation modes, realising that the resulting annihilation and creation operators are not unique, given an ambiguity in the definition of a vacuum state. In de Sitter spacetime, we may asymptotically define the vacuum state: towards the past, each mode function approaches the (unique) Minkowski mode function. Equivalently, modes with a very large wavelength are not affected by the curvature. Having defined the vacuum as such, we got a unique solution to the equation of motion of perturbation modes. We could then compute the power spectrum as the variance of the (super-horizon) perturbations, using the assumption that the modes were initially in their ground state (this will become important in the final section of this chapter). This gave the power spectrum for the metric perturbations, from which we found the power spectrum of the inflaton field perturbations. We finally noted that the amplitude of a mode freezes when it crosses the horizon, and that the Hubble horizon is (approximately) constant during inflation, so that the power spectrum was seen to be scale-invariant.

Insofar as we take the theoretical framework of quantum field theory in curved spacetime for granted, the key assumptions made in this proof are the following: during the slow-roll inflationary era, the universe is approximately de Sitter and only contains small quantum inhomogeneities on a classical background; the vacuum state is given by the Bunch–Davies vacuum; perturbation modes are born in their ground state. The latter is not much of an assumption, as we will see later in this chapter. Modeling the inflationary era as a de Sitter universe is warranted in virtue of the no-hair theorem. This raises the stakes for the validity of this theorem, which I have extensively discussed in Chapter 3. The other two assumptions will be discussed in the next section.

5.2 Quantum problems faced by inflation

This section will be concerned with problems faced by inflation stemming from its quantum-mechanical nature, which is essential to the explanation of structure formation. One such issue is the following. The proposal of inflation theory is that quantum fluctuations act as the precursors to classical-field fluctuations. For this transformation to occur, a measurement-like interaction process must be at work throughout space, converting a field of quantum expectation values into a classical field. How does this quantum-to-classical transition occur? Do modes crossing the Hubble horizon collapse? These questions are strongly connected to foundational issues in the interpretation of quantum mechanics. I believe it is not the burden of the theory of inflation to provide solutions to these issues. Indeed, any theory dependent on quantum mechanics suffers from issues in the interpretation of its foundations. Whether there is something essentially quantum about cosmological density perturbations is a matter that I will take up in the next section.

Next, consider the Bunch–Davies vacuum assumption. As is pointed out in McCabe (2018), there is something strange about defining the Bunch–Davies vacuum via a condition in the remote past. As we have seen, the time coordinate τ only covers half of the de Sitter spacetime, and the boundary $\tau \to -\infty$ does not correspond to some physical boundary. Rather, this boundary is a coordinate artefact. Moreover, since the pre-inflationary universe was not Minkowksi, it is unclear why we can take the vacuum at $\tau \to \infty$ to resemble the Minkowski vacuum.

As it turns out, this problem can be readily resolved. Using the extended coordinates defined in Eq. 5.51, we can get rid of the unphysical coordinate boundary. In fact, the Bunch–Davies vacuum can be defined in the complete de Sitter spacetime as the unique state that satisfies the Hadamard condition and is invariant under the symmetries of the de Sitter spacetime (given by $O(4,1)$). The Hadamard condition is a mathematical requirement for quantum states which excludes ill-defined or unphysical states. Hence there is a physical motivation for the Bunch–Davies vacuum that does not appeal to conditions in a remote past defined on a coordinate boundary. In any case, the Bunch–Davies vacuum is the one that gives the empirically verified prediction of scale invariance, and it can be motivated as such.

Finally, consider the assumption that the seeds for structure formation are quantum fluctuations on a classical background. It is argued in Di Tucci et al. (2019) that the framework of quantum field theory in curved spacetime is not equipped to describe the (very) early universe. Any comoving volume that we are interested in becomes exponentially small towards the past. The spacetime background cannot be expected to behave classically anymore in this regime, and as such it is subject to quantum fluctuations. Taking a semi-classical path integral approach to gravity, in which the background spacetime is also quantised, Di Tucci et al. show that QFT in curved spacetime breaks down for small values of the initial scale factor. The only way to resolve this problem is to assume that the universe was initially sufficiently large (the initial patch that inflates should be larger than the physical Hubble radius). Even worse, the Bunch–Davies vacuum is not naturally selected during inflation anymore, but it should be imposed by fiat. The conclusion is that inflation cannot describe a universe of arbitrarily small initial size, and so inflation itself requires special initial conditions. These could be set up by a pre-inflationary phase, in which a sufficiently large homogeneous patch is created whose perturbations are prepared in the Bunch–Davies state.

Overall, I find that the argument for structure formation is rightfully claimed to be the cornerstone of the inflationary explanation. Despite the fact that some foundational problems of quantum mechanics carry over, it is remarkably economical in accounting for large-scale structure by taking its origin to lie in primordial quantum fluctuations. Nevertheless, I once again find indications that initial conditions cannot be neglected. Although the semi-classical approach to gravity using path integrals remains speculative, the idea that classical spacetime breaks down at very small scales is not. So, I take seriously the suggestion that inflation fails to give adequate descriptions at these scales and that specific initial conditions are required to avoid pathological perturbations plaguing the early universe. This substantiates the view that inflation cannot be an adequate explanation of the early universe (or structure formation in this case) without appealing to initial conditions. Whether initial-conditions explanations of structure formation are adequate without appeal to inflation is the subject of the next section.

5.3 Alternative explanations of structure formation

Is there something inherently quantum about the fluctuations in the CMB spectrum? If this were the case, this would be a strong point in favour of inflation, because classical mechanisms would be ruled out. To answer this question, Ashtekar, Corichi, and Kesavan (2020) analyse the notion of classical emergence in expanding universes. How do we get classical behaviour from a system that is essentially quantum? They identify three non-equivalent ways in which classicality can emerge without invoking any details of quantum measurement theory, such as decoherence or wavefunction collapse. These are: quantum non-commutativity becoming negligible; quantum squeezing; and the ability to approximate the quantum state by a distribution function on the classical phase space. The latter two notions of emergence of classicality are realised regardless of inflation. Hence, inflation does not play a crucial role in realising classical behaviour of matter fields by the time of recombination.

What about an initial-conditions explanation of structure formation? Simply evolve the matter density spectrum at the time of recombination back in time, and postulate these as the initial conditions of the universe. Then declare that this just happens to be the universe that we live in. Unfortunately, things are not this easy. The following two arguments are from Blau and Guth (1987). Assume a radiation-dominated universe. For the first argument, consider density perturbations similar to those in the proof of scale invariance above. The behaviour of such perturbation modes depends on the physical wavelength (associated to a comoving wavenumber k) and the physical Hubble radius $\frac{1}{H}$. The former grows with the scale factor as $\sqrt{t}$, whereas the latter grows as $2t$. Going back in time, then, the perturbations will eventually become bigger than the Hubble radius. This implies that, at early times, the size of the perturbations stretches beyond the Hubble radius, which in a radiation-dominated universe has the same order of magnitude as the horizon radius. The initial conditions of the universe must then contain perturbations that are somehow coherent beyond the limit of possible causal interaction.

Second, one can take an average galaxy today, and compute the size of its corresponding perturbation at a time as early as $t = 10^{-35}$ s, when the initial perturbation spectrum is supposedly imprinted. One then finds that

$$\frac{\delta\rho}{\rho} \sim 10^{-49}, \tag{5.59}$$

where $\delta\rho$ denotes a density inhomogeneity on top of a homogeneous background density ρ. Such perturbations are several orders of magnitude smaller than standard thermal fluctuations. Since we can reasonably expect such thermal fluctuations to be present in a radiation-dominated early universe, it is unclear why large-scale structure is not seeded by these fluctuations rather than by the much smaller density perturbations postulated as initial conditions. In sum, there are serious flaws in a naive initial-conditions explanation for structure formation. Can we do better?

5.3.1 Hollands and Wald

Hollands and Wald propose an alternative mechanism to account for structure formation without relying on inflation (Hollands and Wald 2002a). They describe a mechanism producing density perturbations emulating the proof above. Thus, Hollands and Wald claim that the structure formation of the universe can possibly be explained without inflation.

Consider again de Mukhanov–Sasaki Equation in de Sitter spacetime, and the limiting behaviour of its solutions (Equations 5.20–5.22). In inflationary models, the relevant modes have wavelengths smaller than the Hubble radius for early times. This means the modes evolve like ordinary harmonic oscillators. Their amplitudes are not frozen, so the time at which they came into existence does not affect the argument for scale invariance. Now, for non-inflationary models, the modes have wavelengths larger than the Hubble radius for early times. This means their amplitudes are frozen. So, the analysis of the spectrum of the modes depends sensitively on the initial conditions of the modes.

To get an intuition for why this is true, consider Figure 5.1. The comoving Hubble radius is decreasing during inflation, and increasing during reheating, when spacetime is FLRW. The modes are frozen when their corresponding comoving wavelengths are larger than the Hubble radius. The relevant temperature fluctuations measured in the CMB obviously correspond to sub-horizon modes. Going back in time, these modes are seen to come from the super-horizon

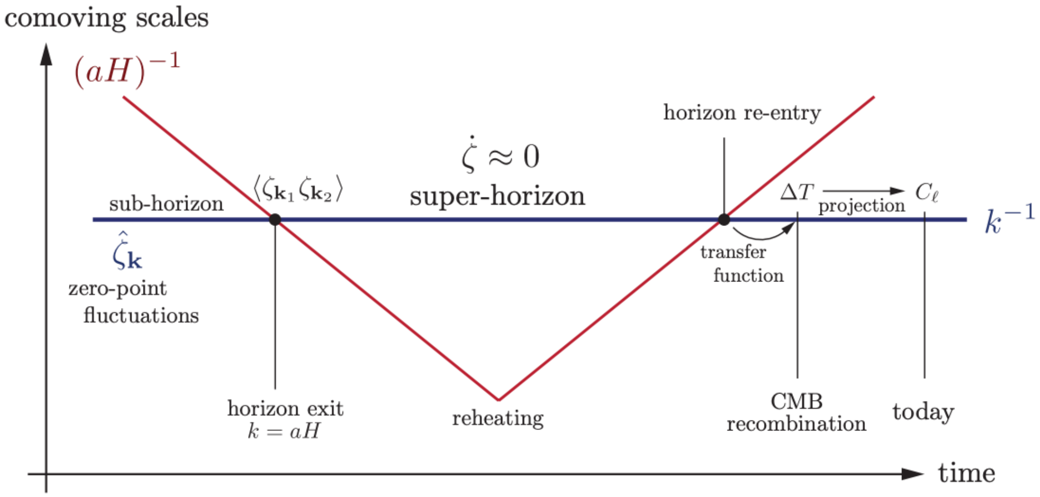


Figure 5.1: Evolution of perturbation modes. The red line shows the size of the comoving horizon, the blue line shows the size of the comoving wavelength. (Baumann 2012)

regime. Going back in time even further, and given that inflation lasted sufficiently long, the Hubble radius is seen to become large enough for the modes to be in the sub-horizon regime. Without inflation, the relevant modes would be in the super-horizon regime for most of their history. We therefore need to address the question in what state they were frozen when they came into existence.

Hollands and Wald make the assumption that the modes come into existence in their ground state when their proper wavelength equals some fundamental scale l_0. So, modes would be continuously created over all time, even at times earlier than the Planck time. When their wavelength is equal to l_0, they pop into existence, and since l_0 is larger than the Hubble radius, the amplitudes of the modes are frozen immediately.

Let me try to formalise this argument. Hollands and Wald consider a massless, minimally coupled scalar field in a spatially flat FLRW universe. The action is

$$S = \frac{1}{2}\int \mathrm{d}\, x^4 \sqrt{-g} g^{\mu\nu}\partial_\mu\phi\partial_\nu\phi, \tag{5.60}$$

with metric

$$\mathrm{d}\, s^2 = -\,\mathrm{d}\, t^2 + a^2(t)(\mathrm{d}\, x^2 + \mathrm{d}\, y^2 + \mathrm{d}\, z^2). \tag{5.61}$$

The Fourier decomposition of the scalar field is

$$\phi(t,\mathbf{x}) = \int \frac{\mathrm{d}\,\mathbf{k}^3}{(2\pi)^{3/2}}\phi_{\mathbf{k}}(t)\,\mathrm{e}^{\mathrm{i}\,\mathbf{k}\cdot\mathbf{x}}\,. \tag{5.62}$$

Plugging this into the action, and using that $\sqrt{-g} = a^3$ gives

$$S = \int \mathrm{d}\, t \frac{a^3}{2}(\dot{\phi}_k^2 - \frac{k^2}{a^2}\phi_k^2). \tag{5.63}$$

Varying with respect to ϕ_k gives

$$\ddot{\phi}_k + 3H\dot{\phi}_k + \frac{k^2}{a^2}\phi_k = 0. \tag{5.64}$$

We define $\omega_k := \frac{k}{a}$. Since the conjugate momentum is given by $\pi_k = a^3\dot{\phi}_k$, the Hamiltonian is

$$\begin{aligned}\mathcal{H} &= \pi_k\dot{\phi}_k - L \\ &= \frac{\pi_k^2}{2a^3} - \frac{a^3}{2}\omega_k^2\phi_k^2.\end{aligned} \tag{5.65}$$

We can define the annihilation and creation operators

$$a_k^- := \sqrt{\frac{a^3\omega_k}{2}}(\phi_k + \frac{\mathrm{i}\,\pi_k}{a^3\omega_k}), \qquad a_k^+ := (a_k^-)^*. \tag{5.66}$$

It is clear that these operators satisfy the canonical commutation relation and that

$$\mathcal{H} = \omega_k(a_k^+ a_k^- + \frac{1}{2}). \tag{5.67}$$

Rewriting gives

$$\phi_k = \sqrt{\frac{1}{2a^3\omega_k}}(a_k^- + a_k^+). \tag{5.68}$$

Then, we may compute the expectation value of $(\phi_k)^2$ as

$$\begin{aligned}\langle\phi_k^2\rangle &= \langle 0|\,\phi_k^2\,|0\rangle \\ &= \frac{1}{2a^3\omega_k}\,\langle 0|\,([a_k^-, a_k^+] + a_k^+ a_k^-)\,|0\rangle \\ &= \frac{1}{2a^3\omega_k}.\end{aligned} \tag{5.69}$$

In inflationary models, the modes are assumed to be born in their ground state, at a time when their wavelength is smaller than the Hubble radius. This means that the modes satisfy Equation 5.69. The proper wavelengths of the modes grow with a, while the Hubble radius is constant H_0 during inflation. When $a = a_0$, the wavelengths of the modes cross the Hubble radius:

$$\frac{a_0}{k} = \frac{1}{H_0}. \tag{5.70}$$

The fluctuation amplitude of the modes then freezes at

$$\langle\phi_k^2\rangle = \frac{1}{2a_0^3(k/a_0)} = \frac{H_0^2}{2k^3}, \tag{5.71}$$

which gives the desired scale-free spectrum.

Inflation is generally taken to be an indispensable part of this mechanism. Without inflation, the relevant modes have wavelength larger than the Hubble radius for a significant part of their life in the early universe. This means that their fluctuation amplitudes are frozen. Assuming all the modes were born at the Planck time in their ground state, their fluctuation amplitudes would go like $1/k$, by Eq. 5.69. This would contradict the observation of the scale-free power spectrum of the CMB.

Instead of this, Hollands and Wald propose that the modes appear (continuously) whenever their length reaches some fundamental length l_0. So, a mode ϕ_k pops into existence in the ground state when $a/k = l_0$. The fluctuation amplitude is then frozen at

$$\langle\phi_k^2\rangle = \frac{1}{2(kl_0)^2 k} = \frac{1}{k^3 l_0^2}, \tag{5.72}$$

as desired.

Hollands and Wald then produce a concrete fluid model of the universe that illustrates their theory, which suggests the value of l_0 to be of order $10^{-5} l_{\mathrm{Pl}}$, where l_{Pl} is the Planck length. Kofman et al. criticize the specific model proposed by Hollands and Wald for being too speculative (Kofman, Linde, and Mukhanov 2002). They consider a radiation-dominated universe, and compute the energy density at the time t^* when fluctuations corresponding to the present cosmological horizon froze. They find that it exceeds the Planck density by many orders of magnitude. This would render any (semi-classical) description of spacetime invalid. Additionally, Kofman et al. criticize the arbitrariness of the length scale l_0. Since the Hubble radius is much smaller than l_0 (and in radiation-dominated models also the wavelength of the photons) at t^*, there appears to be no reason why l_0 is fundamental. The amplitude of the fluctuations of the modes should then be determined by $H(t^*)^{-1}$, and not by l_0.

In response, Hollands and Wald admit that their model has unrealistic properties (Hollands and Wald 2002b). But, they claim that it was never intended as a realistic model, but merely as a demonstration of the existence of a non-inflationary model of the universe explaining the scale-free spectrum of the CMB. Next, they grant that there could be scales other than l_0 that could be of more dynamical importance in determining the early evolution of the modes. But, since the only scale that enters the Lagrangian of these modes is their proper wavelength, they think that their assumption that the modes come into existence when their wavelength equals l_0 is not an unnatural one.

Most significantly, Hollands and Wald agree that their theory is speculative in the sense that it makes assumptions about pre-Planckian times. Since there is no quantum gravity theory available yet, they have to make assumptions about the initial states of the modes. Hollands and Wald, however, claim that inflation suffers from similar speculation. Indeed, the modes responsible for the fluctuations in the CMB had wavelengths smaller than the Planck length at the start of inflation. So, if physics would break down at the sub-Planck scale, then inflationary models would be in the same amount of trouble as non-inflationary models.

Indeed, the trans-Planckian problem for inflationary physics has been noted already in Brandenberger and Martin (2013). To explain structure formation, the large-scale density perturbations observed in our universe are required to have wavelengths smaller than the Hubble radius at the start of inflation. Together with the requirement that inflation solves the uniformity problem, this sets a lower bound on the duration of inflation. If inflation lasts marginally longer than this, then a typical perturbation in our present Hubble horizon (corresponding to large-scale structures like galaxy clusters) must have been smaller than the Planck length at the start of inflation. As Kofman et al. rightly point out, we lack a thorough understanding of physics at these trans-Planckian scales. Fluctuation modes emerge from this poorly understood domain, necessitating new physics to fully comprehend the origin and early development of these fluctuations. What's more, it is plausible that Planck-scale physics could alter the spectrum of cosmological perturbations. Imagine that all perturbation modes start on a fixed initial surface at the beginning of inflation. Short-wavelength modes will be affected by the new physics differently from long-wavelength modes. Their spectrum would be boosted compared to that of long-wavelength modes, resulting in a tilt in the fluctuation spectrum, and therefore a deviation from scale invariance. Although solutions to the trans-Planckian problem have been proposed, these remain controversial, or they introduce more fine-tuning to the theory of inflation. A brief overview can be found in Wolf and Thébault (2023).

In sum, I think the drawbacks of the model pointed out by Kofman et al. are serious. Although the first problem described at the beginning of this section (the coherency of perturbation modes across causal horizons) is adequately addressed, the main problem is essentially an exacerbation of the second problem (scale comparison of perturbations with thermal fluctuations) described there. For the model of Hollands and Wald the situation is aggravated, because it needs to make assumptions about processes occurring in the trans-Planckian domain, in which spacetime itself is expected to break down. Nevertheless, the inflationary explanation of structure formation suffers the same deficiency. I am not in a position to judge whether the inflationary trans-Planckian problem is somehow worse or better than that of the non-inflationary one. In any case, since we have no experimental access to these energy scales, I think it is safe not to ascribe too much weight to trans-Planckian problems in either explanation.

I think Hollands and Wald rightly claim that their model cannot be expected to be realistic in every respect. Inflation has been extensively studied since the 1980s, and therefore much more research effort has been spent on tying up its loose ends. In fact, the first models of inflation suffered from a 'graceful exit' problem, referring to the fatally wrong prediction that accelerated expansion will continue forever. The model of Hollands and Wald should accordingly be treated as an immature idea in need of development, also in light of its restricted application to the case of massless scalar fields. Some steps in this direction have already been taken, as I will show in the next section.

5.3.2 Zero–active-mass fluid

In the present section I wish to discuss a non-inflationary explanation of structure formation by Melia (2021), which expands on the ideas of Hollands and Wald. Crucial to Melia's model of the early universe is a 'numen field', which is cosmic fluid with equation of state

$$\frac{p}{\rho} = -\frac{1}{3}. \tag{5.73}$$

Melia references the rich experimental history of analysing the CMB to argue for a cut-off value for k in the power spectrum; that is, there exists a smallest wavenumber k at which perturbations to the scalar field occur. His argument goes as follows. Three independent satellite missions have confirmed a lack of large-angle correlations in the CMB anisotropies. There is compelling observational evidence that this feature of the CMB is real, in spite of the common claim that it is the result of a systematic measurement error. Indeed, a study of the Planck data shows that the most likely explanation of this lack of correlation is a minimum value for k, predicted to have a value of

$$k_{\text{min}} = \frac{4.34 \pm 0.5}{r_{\text{cmb}}}, \tag{5.74}$$

where r_{cmb} is the comoving distance to the surface of last scattering. A value of $k_{\text{min}} = 0$ is ruled out at a confidence level of more than eight standard deviations.

Melia contends that this means trouble for inflation. A non-zero minimum value of k implies the existence of a definite starting moment for inflation. In a simplified model, the mode corresponding to k_{min} satisfies an equation like Eq. 5.70 at the start of inflation. This gives inflation an initial condition in terms of a definite value for $\dot{a}$. Melia shows that various slow-roll models for inflation are disfavoured by this condition. To be clear, this does not eliminate inflation altogether, but questions the 'ordinary' slow-roll shape of the inflaton field's potential, much like we have seen in Chapter 3. Melia proposes an alternative theory of the early universe. He contends that there is serious observational evidence for a cosmic fluid with equation of state given by Eq. 5.73. Moreover, he claims this is the only expanding FLRW model which is consistent with the strong equivalence principle, stating that at every point in spacetime, there is a local, inertial frame that allows one to measure the spacetime curvature from the observer's perspective. Although I am somewhat sceptical of these claims, it is not my business to evaluate them. Rather, I want to point out that the ideas presented here *could* be the starting point of a new approach to the structure-formation problem, and as such deserve scrutiny.

Let ϕ be a scalar 'numen' field with an equation of state as in Eq. 5.73. In what follows, I will go through Melia's derivation of the scale-free spectrum it produces. The early universe is modeled by the action

$$S = \int \mathrm{d}^4\, x\sqrt{-g}(\frac{1}{2}R - \frac{1}{2}g^{ab}\nabla_a\phi\nabla_b\phi - V(\phi)). \tag{5.75}$$

Then,

$$\begin{aligned}
T_{ab} &= -\frac{2}{\sqrt{g}}\frac{\delta(\int \mathrm{d}^4 x\sqrt{-g}(-\frac{1}{2}g^{ab}\nabla_a\phi\nabla_b\phi - V(\phi)))}{\delta g^{ab}} \\
&= -\frac{2}{\sqrt{-g}}(\frac{1}{4}\sqrt{-g}g_{ab}\nabla^c\phi\nabla_c\phi - \frac{1}{2}\sqrt{-g}\nabla_a\phi\nabla_b\phi + \frac{1}{2}\sqrt{-g}g_{ab}V(\phi)) \\
&= -\frac{1}{2}g_{ab}\nabla^c\phi\nabla_c\phi + \nabla_a\phi\nabla_b\phi - g_{ab}V(\phi).
\end{aligned}$$

So,

$$\rho = T_{00} = \frac{1}{2}(-(\partial_t\phi)^2 + \frac{1}{a^2}(\nabla\phi)^2) + (\partial_t\phi)^2 + V(\phi) = \frac{1}{2}\dot{\phi}^2 + V(\phi) + \frac{(\nabla\phi)^2}{2a^2}, \tag{5.76}$$

and, similarly,

$$p = T_{ii} = \frac{1}{2}\dot{\phi}^2 - V(\phi) - \frac{(\nabla\phi)^2}{6a^2}. \tag{5.77}$$

The last terms, containing $\nabla\phi$, vanish, because the numen field is homogeneous. The equation of state implies that

$$V(\phi) = \dot{\phi}^2. \tag{5.78}$$

From the Friedmann equations, we have that

$$a \propto t^{2/(3+3w)}. \tag{5.79}$$

Using $w = -1/3$ and $a(t_0) = 1$, we have that

$$a(t) = \frac{t}{t_0} = \mathrm{e}^{\tau/t_0}, \tag{5.80}$$

where τ denotes conformal time. Recall the Mukhanov-Sasaki equation (Eq. 5.18). We compute, using the Friedmann equation:

$$z := \frac{a\dot{\phi}}{H} = \frac{1}{\sqrt{4\pi G}}\frac{\dot{a}}{a}\frac{a^2}{\dot{a}} = \frac{1}{\sqrt{4\pi G}}\,\mathrm{e}^{\tau/t_0}\,. \tag{5.81}$$

Then,

$$\omega_k^2 = k^2 - \frac{z''}{z} = k^2 - \frac{1}{t_0^2} = \frac{1}{t_0}\sqrt{\left(\frac{1/H}{a/k}\right)^2 - 1}, \tag{5.82}$$

using that the Hubble radius $1/H$ equals t. Note that both the Hubble radius and the proper wavelength depend linearly on t. Therefore, the fluctuations are time-independent. Since the ratio of the wavelength to the Hubble radius is fixed, the modes do not cross the horizon. The solution to the Mukhanov-Sasaski equation is

$$v_k(\tau) = \begin{cases} A(k)\,\mathrm{e}^{\pm\,\mathrm{i}\,\omega_k\tau} & 1/H > a/k \\ A(k)\,\mathrm{e}^{\pm|\omega_k|\tau} & 1/H < a/k. \end{cases} \tag{5.83}$$

This shows that the sub-horizon modes oscillate, whereas the super-horizon modes do not.

Consider the mode that satisfies $a/k = 1/H$. This is the mode with the longest wavelength relevant for structure formation. It has wavenumber $k_{\text{min}} = 1/t_0$. Why should this mode correspond to the cut-off value measured in the CMB? In the spirit of Hollands and Wald, Melia makes the assumption that modes are born at a specific length scale, which in his case is the Planck scale. Thus, the modes emerge from the Planck regime into the semi-classical universe when $a/k = l_{\text{Pl}}$. In units where $c = 1$, the Planck length equals the Planck time, so we have

$$t_{\text{Pl}} = \frac{t_k}{t_0}\frac{1}{k}, \tag{5.84}$$

where t_k refers to the time of birth of the mode with wavenumber k. Define t_{min} to be the time of birth of mode k_{min}. Suppose that the cut-off value of k consistent with observation from Eq. 5.74 corresponds to k_{min}. Plugging in the value for r_{cmb} for the universe under consideration, we find that

$$t_{\text{min}} = t_0 t_{\text{Pl}} k_{\text{min}} = \frac{4.34 t_{\text{Pl}}}{\log(1 + z_{\text{cmb}})}, \tag{5.85}$$

from which we conclude that $t_{\text{min}} \sim t_{\text{Pl}}$. In a nutshell, if we assume that modes are born when their proper wavelength is at the Planck scale, then the value of k_{min} (determined by observation) corresponds to that of a mode born at the Planck time.

Since the frequency ω_k is time-independent, the Mukhanov-Sasaki equation reduces to that of Minkowski space. Here, the instantaneous vacuum of Eq. 5.46 unambiguously determines the vacuum for all times. Comparing to Eq. 5.83, we find that

$$A(k) = \frac{1}{\sqrt{\omega_k}}. \tag{5.86}$$

The metric perturbations for sub-horizon modes are given by

$$|\zeta_k|^2 = \frac{|v_k|^2}{|z|^2} = \frac{4\pi G}{\omega_k a(t)^2}, \tag{5.87}$$

using the fact that $\omega_k > 0$ for sub-horizon modes. The dimensionless power spectrum (see Eq. 5.41) is then

$$\begin{aligned} P_\zeta(k) &= \frac{k^3}{2\pi^2}|\zeta_k|^2 \\ &= \frac{G}{2\pi}\frac{1}{a(t)^2}k^2\frac{1}{\sqrt{1 - \frac{1}{k^2 t_0^2}}} \\ &= \frac{1}{8\pi^2}\Big(\frac{a(t_k)}{a(t)}\Big)^2\frac{1}{\sqrt{1 - \frac{k_{\text{min}}^2}{k^2}}}, \end{aligned} \tag{5.88}$$

using that the Planck time is $t_{\text{Pl}} = \sqrt{4\pi G}$. Melia proposes that at some length scale L_0, the fluctuations transform into the particles of ordinary matter. The fluctuations would then freeze when their proper wavelengths equal this length scale, i.e. when $a(t)/k = L_0$. The power spectrum is then

$$P_\zeta(k) = \frac{1}{8\pi^2}\Big(\frac{\lambda_{\text{pl}}}{L_0}\Big)\frac{1}{\sqrt{1 - (\frac{k_{\text{min}}}{k})^2}}. \tag{5.89}$$

The scalar spectral index is

$$n_s = 1 + k\frac{d\log P_\zeta(k)}{dk} = 1 - \frac{1}{(\frac{k}{k_{\min}})^2 - 1}. \tag{5.90}$$

This gives a spectral index slightly less than 1, in agreement with observations.

5.3.3 Viability of non-inflationary explanations of structure formation

In the present section, I have looked at the viability of non-inflationary explanations of structure formation. Despite the fact that density perturbations can also emerge classically, it is naive to trace the observed density spectrum backwards in time, and take those as the initial conditions of our universe. At sufficiently early times, these perturbations would be incoherent and smaller than standard thermal fluctuations. Fortunately, more sophisticated non-inflationary models of structure formation are available.

In the model by Hollands and Wald, modes are assumed to pop into existence in their ground state when their physical wavelength is equal to some fundamental scale. As it turns out, this law-like initial condition for modes also gives the desired scale-invariant spectrum of density perturbations as observed from the CMB. The proposal by Melia refines this idea. In his model, the modes are born at the Planck scale, which is empirically supported by the observation of a cut-off value in the CMB. This removes the *ad hoc* and unrealistic fundamental scale l_0 from the model of Hollands and Wald. Nevertheless, to account for scale invariance, Melia needs to invoke a scalar field dominating the early universe so that $a(t) \propto t$. He claims that observational evidence for such a field abounds, although this is disputable.

As always in cosmology, a new model of the universe is expected to give a unifying account of many interconnected phenomena (the universe's expansion history, the statistical properties of the CMB, etc.). In many ways, such a model can then fail, either by predicting something that is unphysical or at odds with observed values. As is shown in the debate between Kofman et al. and Hollands and Wald, the non-inflationary models of structure formation harbour weird features which are allegedly unrealistic. I share this opinion. But, I want to place the side note that non-inflationary models are still in their infancy. Therefore, given the successes of such models described in earlier chapters, I believe it is fair to allow some room for their development, rather than to discard them altogether for containing a bug.

Indeed, some of the purported flaws of the model of Hollands and Wald are actually common to inflationary models. Specifically, I am referring to the trans-Planckian problem here. Both inflationary and non-inflationary models of structure formation are required to make assumptions about super-Planckian energy scales for their explanatory adequacy. As these energy scales are—and probably forever will be—empirically out of reach, I think we should not judge explanations too harshly for making such assumptions.

In the previous section, we have seen some indications from semi-classical gravity that the inflationary explanation of structure formation critically relies on its initial conditions. The goal of the present section was to deliberate the converse: can we have an initial-conditions explanation of structure formation without inflation? Interestingly, the model by Melia strongly suggests the need for an additional dynamical mechanism: a homogeneous scalar field satisfying a specific equation of state. I am gradually building up to my claim that the dichotomy between initial conditions and dynamical explanations fades. Either explanation tends to rely on core elements of the other. In the next chapter this idea will be developed.

Chapter 6

Discussion

6.1 A second critical look

It has been the primary goal of this thesis to present an updated survey of the issues faced by standard cosmology, and to explore the adequacy of the inflationary solution. Taking as a starting point the criticism by Earman and Mosterín, presented in the introduction to this thesis, I will advance here an updated criticism in light of the developments of the debate over the past 25 years. I do not claim that this effort is exhaustive, but I hope to have picked out the most notable, controversial, and philosophically interesting issues.

For the sake of completeness, let me briefly reiterate the verdict of Earman and Mosterín on the monopole problem. The lack of experimental verification of GUTs and the absence of the monopole problem within the GUT framework suggest that there is no monopole problem. As much as the inflationary cosmologist would be happy to see a connection between particle physics and cosmology, such a link is not instantiated here. Whether newer inflationary models, such as the Higgs model, can revive this connection remains doubtful at best.

The flatness problem suffers the same fate. There exists a largely neglected literature on the resolution of this problem (Helbig 2021). Particularly explicit is the dynamical-systems resolution of the problem. It shows that flatness is in reality a generic property of singular FLRW universes. In other words, it would be special to observe the relevant density parameters to be such that they indicate a non-flat universe. Notice again that the flatness problem can only be formulated for FLRW universes: it is chiefly a problem resulting from the dynamics of the Friedmann equations. From these equations it is also immediately apparent, however, that flatness is not special: today's observed flatness can be traced back into the past, where the universe is constrained to be incredibly flat.

Similar considerations defuse the classical horizon problem. Stated in an FLRW context, there is strictly no problem because homogeneity and isotropy are ingrained in the FLRW model. Stepping outside of this context, however, there could be a more general problem lurking. If 1) we assume that the initial conditions of the universe were non-uniform, and if 2) there is no causal mechanism smoothing out these non-uniformities, then there is a problem. This syllogism also indicates the way forward: one must reject one of its two premises. Rejection of the first corresponds to the initial-conditions explanation, rejection of the latter corresponds to a dynamical explanation. The two explanations will be discussed more elaborately in the next section.

Earman and Mosterín make several critical remarks pertaining to the inflationary solution to the uniformity problem. These mostly concern the applicability of the no-hair theorem to inflationary dynamics. I have given substantial evidence that these issues have mostly been resolved over the last 25 years. One last hurdle for the inflationary explanation of uniformity, however, is its generalisation to non-homogeneous universes. The past decade, a rich literature on this topic has developed, culminating in the remarkably adversarial debate between Ijjas et al. and Guth et al. I have shown that either side of the debate is swayed by different perspectives on which states of affairs are natural. More specifically, the naturalness of the chaotic pre-inflationary conditions and the complexity of the shape of the inflaton potential are under consideration here. The extent to which non-homogeneous pre-inflationary conditions obstruct inflation cannot be judged objectively without a measure giving a standard of naturalness. In light of the problems associated with constructing such a measure, one should suspend their judgement about who is right in this debate. Any claims regarding the unnaturalness of an initially uniform universe are similarly unwarranted.

Although the entropy problem is only briefly mentioned by Earman and Mosterín, they do touch on the heart of the issue: is inflation a measure-preserving process or not? To account for the extremely low entropy of the nascent universe, the inflationary advocate holds that only a small homogeneous initial patch is required, which then greatly expands, setting the stage for a uniform reheating process, corresponding to a very low entropy. This is not without controversy: if inflation is a measure-preserving process, then it cannot take a fairly generic state and transform it into a highly special state. In other words, if inflation is reversible, then it only kicks the can down the road, shifting the entropy problem to the pre-inflationary era. I propose that one's conclusion regarding entropy-preservation during inflation hinges on one's interpretation of entropy and one's views on the foundations of quantum mechanics. Additionally, I have argued that positions in the debate on reversibility are influenced by perspectives on what constitutes a satisfactory explanation. I will turn to the role of explanation in the next section.

Lastly, there is the problem of structure formation. I believe that the mechanism for structure formation is the cornerstone of the inflationary explanation. Attributing the origin of large-scale density perturbations to primordial quantum fluctuations gives a beautiful account of structure formation. Any critical remark on the proof of scale invariance in an inflationary universe is minor. A lack of empirical verification of the statistical properties of the CMB perturbations led Earman and Mosterín to express themselves conservatively on this topic. In light of the data of major observational satellite missions from the last decade, however, one cannot conclude otherwise than that the prediction of scale invariance is a stellar achievement of the theory of inflation. The disconfirmation of alternative explanations would then substantiate the role of inflation in the early universe. Let me not postpone the discussion of alternative explanations any longer.

6.2 Two explanatory camps

Throughout my analysis of the explanations that inflation provides, there has been a recurring theme: the possibility of an alternative, empirically adequate explanation grounded in law-like initial conditions. In the present section I juxtapose the two explanations. I try to give the best reading of each explanation in light of the results from previous chapters.

Let me reiterate for each of the surviving problems (uniformity, entropy, structure formation) how the two explanations contrast with each other. I will first consider a naive initial-conditions explanation. The clearest characterisation of the two explanations comes from the solution to the uniformity problem. Rejecting a uniform initial state of the universe implies that one has to introduce an additional causal smoothing mechanism. Alternatively, one can assume a uniform initial state and be done with it.[1] Postulating a uniform initial state immediately also solves the entropy problem, because we know from the considerations about gravitational entropy by Penrose that such a state has the required low entropy. The inflationary explanation can deal with the entropy problem if inflation is irreversible and if there exists a sufficiently large initial homogeneous patch. Lastly, there is the structure-formation problem. I have shown the beautiful explanation that inflation gives for the scale invariance of the perturbations. Alternatively, the naive initial-conditions explanation can simply take the initial fluctuations of the universe to be those observed in the CMB evolved backwards in time, and claim that this just happens to be the universe we live in. More economically, it could also assume only the required statistical properties, such as scale invariance and Gaussianity. We have seen, however, that a naive extrapolation of the density-perturbation spectrum back in time leads to strange conclusions (notably the coherence of modes at non-causal scales). To remain viable, the initial-conditions explanation will need to do better than this naive model.

Fortunately, more sophisticated models of the initial-conditions explanation have been developed. Most urgently, an update was needed to give a viable explanation of structure formation. The model of Hollands and Wald—which crucially imposes the condition that perturbation modes are born in their ground state when their characteristic wavelength equals some fundamental scale—addresses this issue. The elaboration of the model by Melia arguably makes it more realistic. Further sophistications of the initial conditions model exist in relation to the uniformity and entropy problem. First, the Weyl-curvature hypothesis covers the entropy problem, and thereby also the uniformity problem, although it is not clear how it is really different from the assumption of low initial entropy. Second, uniform universes will be picked out by requiring a finite-valued action in the modification of gravity which includes quadratic curvature terms. Third, imposing CPT symmetry on the universe is claimed to explain low entropy, uniformity, and structure formation. Of course, these sophistications also come with additional assumptions that may or may not be warranted.

Thus I have come to distinguish two explanatory camps. For reasons of clarity, I will sketch a typical 'story of the early universe' for an advocate from either camp. The dynamical account goes as follows.

> In the beginning, there was a generic hodgepodge of energy. After some small region spontaneously fluctuates into a homogeneous state, inflation sparks off. By the cosmic no-hair theorem, this region isotropises by means of the accelerated expansion it undergoes. This creates a large, uniform volume of extremely low entropy. The inflaton field driving this process should be treated as a quantum entity. Its inherent quantum fluctuations seed the formation of structure in the post-recombination universe. The accelerated expansion comes to an end, and the inflaton field decays into the ordinary matter that we know.

On the contrary, the non-naive initial-conditions story goes like this:

[1] Of course, one could also assume both a uniform initial state and an inflation-like mechanism. This could be the position of someone who thinks that the initial-conditions problem cannot be resolved within the context of inflation, but that we do require inflation to explain structure formation. Although I have yet to encounter this position in the literature, it could actually be a reasonable one, as I will show later in this chapter.

> In the beginning, there was a uniform (and therefore low-entropy) universe. Some law-like initial condition (past hypothesis, Weyl-curvature hypothesis, CPT symmetry, finite action in quadratic gravity) dictates this. Another law-like initial condition accounts for the birth of perturbations on this uniform background. This condition is such that the density-perturbation spectrum emerging from these possesses the required statistical properties, such as scale invariance.

This portrays the two explanatory camps encountered in this thesis. Let me now address the question of what it means to explain.

6.3 Explanation

An objection that one could make against the initial-conditions explanation is that it is not even an explanation. If the explanation is simply a restatement of the data, then what is really being explained? Such an objection stems from a dissatisfaction with a claim like ‘we just happen to live in a universe with property X’. The goal of science, after all, is to provide explanations of such properties, and postulating them as brute facts would be tantamount to giving up.

Note that this objection primarily pertains to the naive initial-conditions explanation. It merits the predicate ‘naive’ precisely for this reason: it takes the CMB data as a brute fact, and postulates the initial state of the universe to be such that it evolves into the appropriate state by the time of recombination. Allegedly, this does not do any explanatory work. In spite of the best efforts of the non-naive initial-conditions explanation to go beyond a mere tautology of observations, there are similar arguments lurking that could undermine its explanatory credentials. Indeed, as I have pointed out before, the Weyl-curvature hypothesis is simply a relabeling of the assumption of an extremely-low entropy initial state. As such, the non-naive explanation is susceptible to the allegation of being a non-explanation, too.

To get a grip on this criticism, I will first contemplate what it means to be an explanation. An influential account of explanation is the deductive-nomological (DN) model, due to Hempel (1965). The DN model posits that (scientific) explanations can be articulated as logical deductions, where the phenomenon to be explained, or explanandum (E), is derived from a set of general laws (L_i) and particular conditions (C_i), collectively known as the explanans. The presence of a law in such a deduction is essential. I will elaborate on the definition of a law later in this chapter.

Let’s see how the initial-conditions explanation conforms to the DN model. First, consider the uniformity problem. A DN explanation would look as follows.

(L_1)	The initial state of the universe is uniform.
(L_2)	The Einstein field equations dictate the evolution of the universe.
(C_1)	-
(E)	The universe is uniform at the time of recombination in spite of the non-existence of a smoothing mechanism.

This seems to confirm the suspicion that there is something tautological about initial-conditions explanations. Note that I cannot put the initial uniformity as a particular condition (C_1) because the explanatory work must be done by laws in the DN model.

The schematisation of the initial-conditions explanation for low entropy is identical in form. But, as I noted in Chapter 4, if the explanandum under consideration is the second law of thermodynamics, the situation changes. The observation that entropy increases in any thermodynamic process can be explained as follows.

(L_1) The initial state of the universe has extremely low entropy.

(L_2) Entropy always increases in thermodynamic processes, both towards the past and towards the future (by statistical arguments about microstates).

(C_1) X is a thermodynamic process.

(E) The entropy in X increases.

This is *not* a tautological explanation. Substituting L_1 for the law that the initial state of the universe is uniform (to a very high degree) makes the situation even better for initial-conditions explanations. Just this one law is crucial to both explaining uniformity *and* the second law of thermodynamics.

Let me finally try to schematise the argument for scale invariance. Naively, the initial-conditions explanation of scale invariance is the following.

(L_1) The universe has an initial perturbation spectrum that is the time-reversed evolution of the CMB perturbation spectrum.

(L_2) The Einstein field equations dictate the evolution of the universe.

(C_1) -

(E) The perturbation spectrum of the CMB is scale-invariant.

Apart from the physical problems associated with L_1, it may be objected that this is a non-explanation. Strictly, it does confirm to the DN model, but due to its trivial nature it can hardly be defended as a proper explanation, for it does not furnish any understanding. Yet understanding is arguably the main motivation for the DN model (Hempel 1965). Therefore, the naive initial-conditions explanation is problematic. Nevertheless, the updated (non-naive) initial-conditions explanation circumvents these problems.

(L_1) Perturbation modes are born in their ground state when their corresponding wavelength is equal to some fundamental length, quite possibly the Planck length.

(L_2) The Einstein field equations dictate the evolution of the universe.

(C_1) -

(E) The perturbation spectrum of the CMB is scale-invariant.

As is clear from the efforts of Hollands, Wald, and Melia, this explanation is not a tautology. It conforms to the DN model, and should therefore be appreciated as a proper explanation. Note that the particular condition slot is empty. The presence of such a condition, however, is not essential to DN explanations.

Thus, at least per one prominent account of explanation, the initial-conditions camp is safe. Of course, one could criticise the DN model for not being an adequate characterisation of explanation. Such criticisms have indeed been made (Woodward and Ross 2021). Most prominently, it has been argued that the DN model does not give sufficient conditions for explanation when it overlooks relevant causal mechanisms. A famous example of this is the DN explanation of the height of a flag pole by the length of its shadow and the angle of the sun above the horizon. Such considerations are not applicable to the deductions above. Therefore, I take the DN model to be a satisfactory account of explanation for the case at hand.

There is another objection, due to Ellis (2007), which states that there is no clear distinction between initial conditions and laws in cosmology. Since we cannot alter the initial conditions, they appear necessary rather than contingent. That is to say, if the present state of the universe is determined by its initial conditions and the laws of physics, then these initial conditions could not have been different. Thus, the fundamental distinction between initial conditions and laws is absent, as the former are characterised by their contingency and the latter by their necessity. This is a problem for the DN explanations considered so far, because they require this distinction. Indeed, the explanatory work must be done by laws in these deductions, and not by initial conditions.

To save this distinction, Antoniou (2024) argues for a theory-relative interpretation of physical modality. All physical theories have a constitutive structure, which specifies the minimum necessary background structure to describe physical systems of the theory and its laws, and a nomic structure, which specifies the constraints on the possible configurations of a physical system, stated as laws. These laws can be dynamical laws, describing how states change over time; scaling laws, dictating what instantaneous configurations of parameters are allowed; or law-like fundamental assumptions, which is any other statement constraining the possible configurations of physical systems, such as the principle of locality or the Weyl-curvature hypothesis. In this framework, physical possibility refers to the range of initial configurations and subsequent dynamic evolutions of physical systems permitted by the theory's nomic structure. Physical necessity, on the other hand, refers to the required constraints on these initial configurations and their ensuing changes to maintain consistency between the laws of the theory. To illustrate these concepts, consider the theory of Newtonion mechanics. In this theory, it is possible to have particles travelling faster than the speed of light, and it is necessary that an accelerated particle will eventually attain an infinite speed. A more relevant example is the theory of standard cosmology. In this theory, the initial conditions of the universe correspond to different possible configurations that leave the Friedmann equations invariant, reflecting their contingency. These configurations represent counterfactual possibilities that do not describe our universe but maintain the theory's self-consistency. Whether these counterfactual models could actually occur in nature is irrelevant.

Using this theory-dependent conception of physical modality, the distinction between initial conditions and laws can be upheld. This is required for the cogency of the DN model. The deductions above should now be understood as taking place in the context of a relevant theory, in this case general relativity conjoined with thermodynamics. Each of the laws L_i considered so far can indeed be construed as part of the nomic structure of a theory, as they provide constraints on the possible configurations of physical systems. On this view, the DN explanations presented in this section make sense in spite of the criticism due to Ellis. I now turn to another definition of lawhood that accommodates initial-conditions explanations.

6.3.1 Can laws be puzzling?

One may have the intuition that laws themselves cannot be puzzling. If a law cries out for explanation, then it cannot, ultimately, be a law. An example is the law that the universe is governed by the Einstein field equations plus a cosmological constant. To some, this cannot be the final picture, because there needs to be something physical (dark energy) corresponding to this cosmological constant, like vacuum energy. See Bianchi and Rovelli (2010) for a critique of the belief that the cosmological constant is puzzling. Another example is the law-like initial condition that the universe has extremely low entropy. In the present subsection, I point out a defense for the initial-conditions camp against the accusation that such conditions are unlikely and therefore cry out for explanation. These ideas are due to Callender (2004).

Consider again Penrose's entropy problem, which I discussed in Chapter 4. Crucial to our understanding of thermodynamics is the idea that, under the standard microcanonical measure of statistical mechanics, microstates are more likely to evolve towards a state of equilibrium. Such an argument, however, is invariant under time reversal. Therefore, we should expect entropy to increase both towards the future *and* towards the past. Since we only every observe entropy-increasing processes, we need an explanation for this symmetry breaking. Thus we invoke the past hypothesis, which postulates that the universe occupied an initial macroscopic state of extremely low entropy. By the standard measure of statistical mechanics, this is extremely unlikely.

The question at hand is whether this past state stands in need of explanation. Price has argued that it does (Price 2002). To Price, not explaining the past state is tantamount to giving up. He recognises two possible positions that could be held by someone who believes that the past state does not stand in need of explanation. Either, one could hold that for any time the macroscopic state of the universe does not need explaining. This position amounts to global explanatory nihilism, answering every why question with *that's just the way it is.* Alternatively, one could hold that the past state is the only state that does not stand in need of explanation. On this view, one maintains a bias with respect to temporal direction. If we consider the time-reversed evolution of our universe, we would see all sorts of mysterious contra-thermodynamic phenomena as we approach the big-bang singularity. This would surely cry out for explanation. In essence, however, it is nothing other than the past state viewed from an different temporal direction. Assuming the laws of our universe are time-reversal invariant, it is unjustified to absolve only the past hypothesis of a demand for explanation. This prevents the defender of the non-explainability of the past hypothesis from making the argument that the past hypothesis is special because it cannot be *causally* explained.

Callender's reply is as follows. First, he questions what it is about the past state that makes it stand in need of explanation. Could it not be a brute fact? This raises the more general question what it is about facts in general that makes them brute or not. To Price, it is the unlikeliness of the state that lies at the heart of the problem. But Callender raises the objection that the demand for explanation of the past state does not follow from its unlikeliness. In fact, innumerable low-probability events occur every day that definitely do not require explanation. What incredibly narrow set of initial conditions of the universe would lead me to write the number 61 at this precise moment? And yet, here we are.

Callender submits a different view. There is nothing about facts that makes them brute or in need of explanation. In particular, the likeliness of a fact does not determine whether or not it is to be explained. Instead, Callender suggests to treat theoretical systems holistically, and see how they perform empirically. Barring antirealism, one could then compare empirically adequate systems on the basis of theoretical virtues, such as explanatory power and simplicity. On this account, the past hypothesis fares excellently. It is simple, and it provides a coherent explanation of all thermodynamical processes.

In Callender's view, this is what science is after: explaining as much as possible with as little as possible. We aim to find theoretical systems that capture myriads of empirical regularities with simple generalisations. The past hypothesis, complemented with the laws of statistical mechanics, gives an example of a system that strikes the right balance between simplicity and strength. If there would be no past state then all ordinary thermal processes would become unlikely, so the system would lose strength. If the laws of thermodynamics could somehow be adjusted to rid it of the past hypothesis, the system would lose simplicity. In the next section, I will apply this 'best-systems' framework to the two explanatory camps under discussion.

The best-systems definition of laws of nature—developed by Mill, Ramsey, and Lewis—offers a definition of what it means for a statement to be a law of nature (Carroll 2020). According to this definition, laws of nature are those statements that appear in the best systematization of our empirical observations. This best systematization of phenomena is achieved by striking an optimal balance between simplicity and strength. The best-systems approach is grounded in the concept of Humean supervenience, which holds that all facts about the world, including laws of nature, supervene on the spatiotemporal distribution of particular, local matters of fact. In other words, the best-systems definition assumes that the laws of nature do not 'govern the world from above' but rather emerge from the regularities present in the mosaic of particular events. Hence, the laws are those descriptive summaries that provide the most efficient and informative account of this mosaic, reflecting the underlying regularities without invoking metaphysically heavyweight notions like necessity, universality, or causality.

To sum, the best-systems account of physical theories places the initial-conditions explanation on firmer footing in two ways. First, it helps to make sense of the concept of law in the DN-type explanations presented earlier. The initial state of the universe can be postulated as an axiom of the theory of cosmology (whether this is the *best* system is debatable of course), and as such enjoy the status of law. It is on this view that the term 'law-like initial condition' is justified. Second, the best-systems account helps to make sense of the criticism that laws should not be puzzling by corresponding to an unlikely state of affairs. Any law can be taken as a brute fact, not requiring any additional explanation, in virtue of their axiomatic position in a best system. I have noted that the best-systems approach stems from a Humean conception of empirical regularity.

6.3.2 Non-Humean definitions of lawhood

According to the Humean best-systems definition of lawhood, laws are metaphysically determined by the theoretical virtues of simplicity and strength once some part of the mosaic is charted. Laws are essentially summaries of facts within the world and they do not exist independently of this mosaic. In contrast, in non-Humean definitions of lawhood, laws are also fundamental. They do not supervene on the collection of all local matters of fact.

Chen and Goldstein (2022) argue that the past hypothesis can be construed as a law in a non-Humean framework, too. They defend a primitivist form of non-Humeanism. That is to say, fundamental laws are primitive facts about the world, that are not to be analysed in terms of other fundamental entities like universals or essences. They take this view because the concept of law is more intuitive than these other concepts. Chen and Goldstein develop an unrestricted definition of lawhood called minimal primitivism, in which any laws may have any form. They are not required to be dynamical, but may also take the form of boundary conditions. Laws *govern* our world by making metaphysical statements about what configurations of physical states are allowed, and why this is so. Given a state of the universe at some time, the laws explain why its history is constrained to some specific subset of all possible histories. Note that this idea of lawhood features implicitly in the theory-dependent conception of physical modality presented in the previous section. On the view of minimal primitivism, theoretical virtues like simplicity are not fundamental to lawhood, but merely serve as epistemic guides. Simplicity guides us towards the past hypothesis, which provides a restriction on the physically possible evolution of the state of the universe, making a reversal of the arrow of time with immensely improbable.

So, Humean or not, there is substantial support for the use of law-like initial conditions in physical theories. Many law-like initial conditions have been showcased in this thesis: the past hypothesis, the Weyl-curvature hypothesis, the finite-action demand in quadratic gravity, CPT symmetry, and the birth of perturbation modes at a fundamental length scale. Each of these conforms to the definitions of lawhood presented in this section. I conclude that the explanations of the initial-conditions camp are proper ones, that do not amount to some sort of disguised defeatism.

I wish to highlight two more arguments in favour of the law-like status of the past hypothesis. For a more detailed exposition of these arguments, see Chen (2023). First, if the past hypothesis is assumed to be a law, then many arrows of time, other than the thermodynamic one, are explained as well. One example is the records arrow of time. Why do we find so many records of the past (craters, fossils, photographs), but none of the future? Other examples of arrows of time are the epistemic arrow (knowledge about the past is more thorough and accessible than knowledge about the future) and the intervention arrow (we have the ability to shape future events, yet we lack the ability to change events that have already occurred). Each of these arrows derives from the past hypothesis *qua* law. Second, the law-like status of the past hypothesis may address the issue of complexity of the universal wave function, and consequently shine light on the nature of the quantum state. For details, see Chen (2023).

6.4 Explanatory preference

What compels one to favour one explanation over the other? What are reasons to join the dynamical camp rather than the initial-conditions camp, and vice versa? I will first collect the results towards answering this question from earlier chapters. Next, I investigate if there are certain explanatory virtues that drive physicists to one camp or the other. I conclude that there are many such extra-empirical considerations at play. Preference for one camp above the other can be reasonably motivated by these considerations, which implies the legitimacy of both explanations. Notwithstanding the interesting sociological analysis offered in Earman and Mosterin (1999, pp. 4–10), I mostly refrain from discussing the sociological dimension of this matter.

The debates on initial conditions for inflation, its reversibility, and on alternative mechanisms for scale invariance have proven to be interesting case studies of extra-empirical virtues in action. The initial conditions debate, discussed in Chapter 3, exposed the considerations of naturalness that play a crucial role in determining one's explanatory preference.[2] Taking both chaotic initial conditions and a featureless inflaton potential to be natural leads to the conclusion that inflation itself requires very special initial conditions. Since a major goal of inflation is to explain the special initial conditions (uniformity) of the post-recombination universe, it renders itself obsolete in this respect. Hedging inflation against this conclusion, Guth et al. argue that a featureless potential is actually unnatural. I have already argued that it is unfeasible to construct an objective standard of naturality based on empirical evidence. To add to that, there are not even—in view of the measure problem—rigorous arguments to show that initial uniformity of the universe is unnatural. Thus, the initial conditions debate shows that explanatory preferences are significantly influenced by the extra-empirical virtue of naturalness, although no objective standards are available evaluate this.

The reversibility debate, discussed in Chapter 4, revolves around extra-empirical virtues, too. As I have shown, whether one thinks of inflation as a reversible process depends on one's attitude towards what entropy is, and on one's preferred interpretation of quantum mechanics. Particularly the latter is strongly associated with extra-empirical virtues. Each major interpretation of quantum mechanics claims to be empirically adequate, and as a result there is underdetermination of theory by data. Still, one may strongly prefer one interpretation above another because of one's stance on locality, unitarity, scientific realism, etc. The reversibility debate inherits these considerations: if you prefer some interpretation of quantum mechanics because of this or that extra-empirical virtue, then this same virtue compels you to pick a side in the reversibility debate on account of this interpretation's perspective on unitarity.

The debate on alternative (non-inflationary) mechanisms to explain scale invariance, which I discussed in the previous chapter, again involves extra-empirical considerations. In this case, I use the term extra-empirical in a practical sense. Indeed, the trans-Planckian issues featuring in the debate of Hollands and Wald versus Kofman et al. are connected to assumptions about energy scales that are far beyond current experimental and observational reach, although they are testable in principle. I consider the main extra-empirical motivations of this debate to be assumptions about how and when perturbation modes come into existence and what initial energy state (with respect to some vacuum state) is natural for them to occupy. Note that some trans-Planckian assumptions can be rejected for physical reasons: coherency of perturbations outside of causal limits is at odds with relativity; suppressing thermal fluctuations to prevent these from dominating smaller-scale perturbations is at odds with thermodynamics. In principle, one could weasel out of such incompatibilities by introducing even more trans-Planckian assumptions, but modifying well-established theories is a high price to pay.

In another attempt to clarify the explanatory preference for one of the camps, I apply the best-systems approach, developed in the previous section. Recall that the 'better' system is the one that can explain more (or constrain more) while assuming less. I have shown that both the initial conditions and dynamical explanations are comparable in strength, as either can adequately solve the problems of uniformity, entropy, and structure formation. So the question becomes which of the two explanations is simpler. Needless to say, this is a vague notion. How could one reliably determine which explanation makes fewer assumptions? How does

[2] The two sides of this debate do not neatly conform to the two explanatory camps that I have introduced. The authors arguing against inflation (Ijjas, Steinhardt, and Loeb) seem to favour a cyclic-universe model, which does not conform to the demands of the initial-conditions camp. The debate itself revolves around initial conditions, however, and as such it is an applicable case study for the explanatory dichotomy under consideration.

one 'weigh' assumptions? In other words, is postulating extremely uniform initial conditions together with an assumption on mode generation somehow simpler than assuming a scalar field with a particular potential whose decay gives rise to ordinary matter particles? In absence of an unambiguous method of quantifying simplicity, I think one cannot make claims about which explanation corresponds to a better system.

What is it that each of the two explanations assumes? For the initial-conditions camp, it is, of course, the initial state of the universe. But need this explanation assume the precise configuration of the CMB evolved backwards in time to some initial spacelike slice? This would amount to a lot of assumptions, namely, the initial positions and velocities of all particles. But this reasoning overlooks the fact that the explanandum under consideration is not the precise configuration of the CMB, but merely its statistical properties, in particular its approximate uniformity and the scale-free spectrum. In order to explain these, one would merely have to assume an initial state that is also sufficiently uniform, and which leads to a scale-free spectrum by the time of recombination. Now, considering all the ways in which the initial state could be configured, such an assumption would still be special or unlikely. Callender argued that this does not demand explanation, but that we should evaluate the system as a whole based on the ratio of how much it explains to how much it assumes. But is the assumption of near-uniformity not a very big assumption, in the sense that it excludes innumerable initial configurations? Unlikeliness returns to the stage, not as a feature that requires explanation, but as a counterpart to simplicity.

As for the dynamical-explanation camp, the assumptions are: an inflationary mechanism, a corresponding potential-energy curve, a reheating mechanism, and chaotic initial conditions. In light of the debate on special initial conditions for inflation, the shape of the potential must be complicated, or the initial homogeneous patch from which inflation starts must be large. In addition, an entity causing inflation must be postulated. There is currently no confirmed particle-physics mechanism with the properties required to drive inflation. Even the Higgs model makes assumptions about the properties of the particle that have not been observed in experiments. But this means that the inflationary advocate has to postulate a new entity with properties unlike anything we have ever seen before. In the vocabulary of the literature on Ockham's razor, this means giving up qualitative parsimony (Baker 2022). That is to say, compared to a non-inflationary explanation, the inflationary story is less ontologically parsimonious. It assumes a new *type* of entity. Of course, there is no *ceteris paribus* comparison to be made with the initial-conditions explanation. For one, the initial-conditions explanation could be accused of being less ontologically parsimonious in a quantitative sense. In assuming a large number of patches each with the same uniform conditions, it has to give up quantitative parsimony. I conclude that the best-systems approach—and related concerns about simplicity—does not provide any meaningful suggestions for which explanatory camp is to be preferred. Nevertheless, different conceptions of simplicity may motivate an explanatory preference. On the hand, it can be taken to signify quantitative parsimony, being a counterpart to unlikeliness. On the other hand, it can be taken to signify qualitative parsimony, expressing a conservative attitude towards introducing new *kinds* of entities.

Finally, I discuss explanatory preference based on considerations about fine-tuning. Let me first make this notion precise. In line with Wolf and Thébault (2023) I define fine-tuning to mean sensitivity to counterfactual perturbations. That is to say, a fine-tuned explanation would break down if the fine-tuned property had been slightly different. For the initial-conditions explanation, the main fine-tuning issue is that gravity will cause non-uniformities to grow over time, as per the result of Collin and Hawking discussed in Chapter 3. The law-like initial condition of a uniform state is not robust to small perturbations, as these would grow and result in a non-uniform CMB spectrum. Note, however, that some more sophisticated initial-conditions explanations can deal with this sensitivity. For example, in quadratic gravity any non-uniform models of the universe are excluded because they are unphysical (the action blows up).

The dynamical explanation, in turn, also suffers from fine-tuning. Indeed, this is what the debate on special initial conditions for inflation largely revolves around. If inflation requires a very large homogeneous patch to spark off, then any small perturbation on this patch may block inflation, resulting in a fine-tuned explanation. The debate is ongoing, however, and we have seen that the conclusion can possibly be avoided if the shape of the inflaton potential is a complicated one. But this could introduce new fine-tuned features into the potential. More significantly, Wolf and Thébault have argued that solving the trans-Planckian problem comes at the expense of additional fine-tuning (Wolf and Thébault 2023). Once again, I conclude that an extra-empirical virtue is at play here. Fine-tuning does not strongly favour one explanation over the other, but it may be a strong intuition lingering in the background. Like the other extra-empirical virtues under consideration here, it operates quietly behind the scenes. Careful scrutiny of one's dispositions may show an explanatory preference for either camp. Both can be justified.

I conclude that accepting or rejecting the inflationary explanation is strongly dependent on implicit extra-empirical considerations. I have shown that there are many such considerations at play: conceptions of what physical states are natural or likely; interpretations of the foundations of quantum mechanics, in particular regarding unitary evolution; attitudes towards inaccessible energy scales; opinions on what basic assumptions of a theory are simple; judgements on fine-tuning in explanations. Importantly, one may justify one's allegiance to either explanatory camp by appealing to these virtues. This places the two explanations on par.

6.5 A fading dichotomy

The parity of the two explanations with respect to extra-empirical virtues suggests something more. Perhaps the fact that both explanations possibly suffer from fine-tuning, trans-Planckian problems, or convoluted assumptions is an indication that a stalemate has been reached. The fact that no extra-empirical virtue picks out one explanation or the other suggests that we should try a new angle. I believe that this new angle looks at hybrid explanations. Throughout this thesis, I have already hinted at the idea that the dynamical explanation may rely on law-like initial conditions, and that the initial-conditions explanation may rely on additional dynamics. In this section, I develop this idea. I draw inspiration from Martens and Lehmkuhl (2020).

Let me revisit the instances when the dynamical explanation capitalised on non-dynamical law-like statements or special initial conditions. First, the inflationary explanation depends significantly on the dominant energy condition. I have presented the no-hair theorem and some generalisations in great detail. It is clear from the proof that the DEC is a crucial assumption. Without it, it is not at all clear whether inflationary universes (homogeneous or not) will converge to isotropy. Thus, there must be a law-like condition in place which restricts all matter to be of a particular kind. Second, inflation may need special initial conditions to start off in some region. Exactly how special depends on one's stance in the initial conditions debate and the reversibility debate. Third, I have pointed out in Chapter 5 that in the framework of semi-classical gravity it can be shown that inflation requires special initial conditions to give a non-pathological description of the universe when the scale factor is small. In short, there are suggestions that the dynamical explanation must rely on non-dynamical conditions and special initial conditions for its cogency.

The converse is also true. The first example of this is the quadratic-gravity resolution of the entropy problem discussed in Chapter 4. By excluding finite values of the gravitational action, it gives a more elaborate initial-conditions explanation of the low initial entropy of the universe. But in doing so, it makes additional assumptions about (gravitational) dynamics by including quadratic terms in the action. These modified dynamics also lead to an accelerated increase of the scale factor in the early universe, providing a basis for an inflationary phase. Second, in Chapter 5 I have presented Melia's model of structure formation. In making the model by Hollands and Wald more realistic, an additional scalar field with a specific equation of state had to be introduced. So, this model has to bring in an additional dynamical mechanism. Thus, there are clear indications that the initial-conditions explanation requires modifications or additions of dynamical mechanisms.

This goes to show that the best explanation is maybe not a pure initial-conditions explanation or an explanation purely in terms of dynamical mechanisms. The above examples suggest that it could be better to look for a hybrid explanation. The existence of these hybrids implies the fading of the dichotomy of the two camps. The two explanations borrow from one another, so conceivably it is the ground between the two camps that proves most fertile. See Figure 6.1 for a visual overview of the main ideas presented in this thesis.

6.6 Outlook

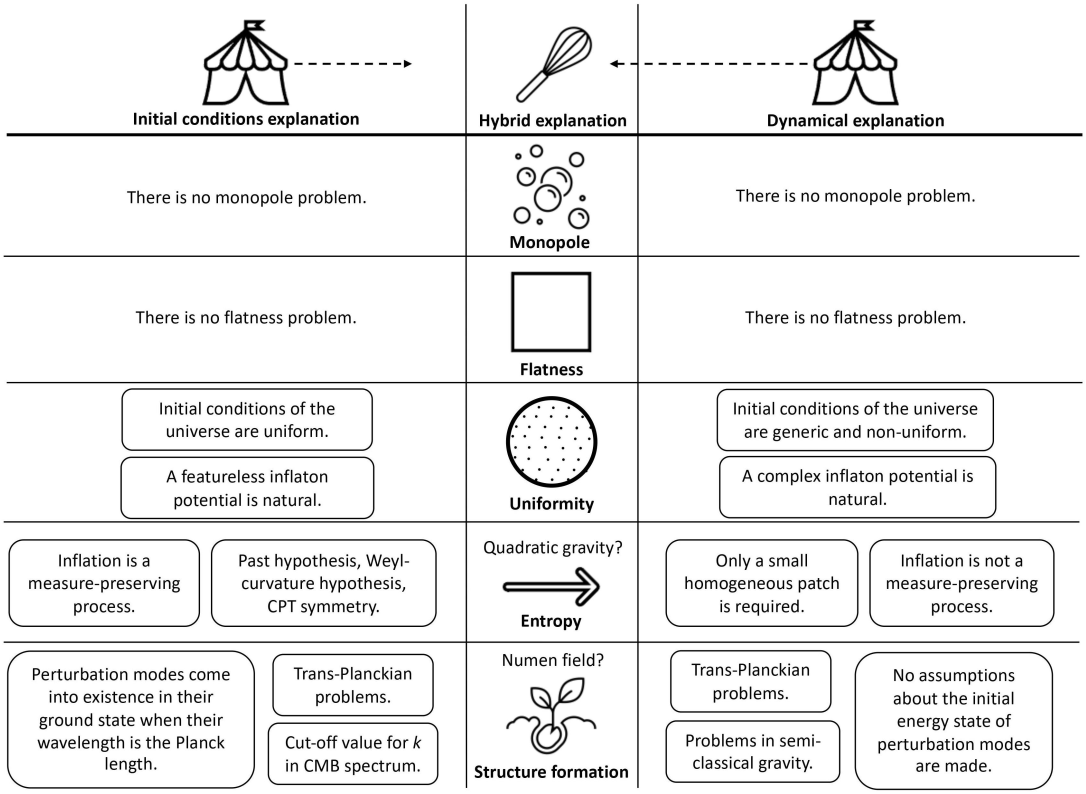
The principal objective of this thesis is to provide an update of the criticism against inflationary cosmology published by Earman and Mosterín 25 years ago. Along the way, I picked up an alternative explanation of the problems of standard cosmology in terms of law-like initial conditions. I have tried to vindicate this explanation of the allegation that it is a non-explanation. In so doing, I hope to have justified a tenable alternative to inflation. Although this alternative can account for the problems of uniformity, entropy, and structure formation, this does not imply that it is to be preferred above inflation. In fact, explanatory preference is entirely a matter of non-empirical considerations.

Despite my best efforts, I should note that my overview is not complete. For example, a statistical feature of the perturbation spectrum of the CMB that has only recently come to my attention is that of phase coherence. This is another prediction of inflation on structure formation that has been empirically confirmed. Future research should reveal whether an initial-conditions explanation can account for this feature. Does the model by Hollands and Wald predict phase coherence? Another example is the recent revival of the special initial conditions debate in the field of numerical GR. I suspect that new research will show that extra-empirical virtues are still at play here.

I believe the future lies in the middle. On the one hand, inflationary cosmologists may learn from the initial-conditions explanation that a conservative approach can still be a fruitful one. Some scepticism could be good for the zoo of inflationary models, and by pointing out problems for inflation it could even reveal the way forward. On the other hand, those favouring the initial-conditions explanation should not neglect novel ideas that are speculative. This explanation stands in need of development, and so it can benefit from additional scalar fields, semi-classical treatments of gravity, or modifications of the gravitational action. I do wish to remark that the inflationary explanation has received substantially more attention in the literature over the past decades. As such, the under-developed initial-conditions camp may merit some more research effort.

The intention of my considerations, however, is not to tell the working physicist what's right and wrong. Indeed, I should be careful not to forego an important message of Earman and Mosterín.

> *It is creative physicists, not philosophers of science, who must place the bets that count on which avenues of research will prove to be fruitful.* (Earman and Mosterin 1999, p. 46)



Initial conditions explanation	Hybrid explanation	Dynamical explanation
There is no monopole problem.	Monopole	There is no monopole problem.
There is no flatness problem.	Flatness	There is no flatness problem.
Initial conditions of the universe are uniform. A featureless inflaton potential is natural.	Uniformity	Initial conditions of the universe are generic and non-uniform. A complex inflaton potential is natural.
Inflation is a measure-preserving process. Past hypothesis, Weyl-curvature hypothesis, CPT symmetry.	Quadratic gravity? Entropy	Only a small homogeneous patch is required. Inflation is not a measure-preserving process.
Perturbation modes come into existence in their ground state when their wavelength is the Planck length. Trans-Planckian problems. Cut-off value for *k* in CMB spectrum.	Numen field? Structure formation	Trans-Planckian problems. Problems in semi-classical gravity. No assumptions about the initial energy state of perturbation modes are made.

Figure 6.1: An infographic summarising the main ideas of this thesis.

Bibliography

Albrecht, Andreas (Oct. 2002). “Cosmic inflation and the arrow of time”. In: *Science and Ultimate Reality*. Ed. by John D. Barrow, P. C. W. Davies, and C. L. Harper. _eprint: astro-ph/0210527. Cambridge: Cambridge University Press, pp. 363–401.

Antoniou, Antonis (Jan. 30, 2024). *Laws, Initial Conditions and Physical Modality: Lessons from Cosmology*. Forthcoming.

Ashtekar, Abhay, Alejandro Corichi, and Aruna Kesavan (July 7, 2020). “Emergence of classical behavior in the early Universe”. In: *Physical Review D* 102.2. Publisher: American Physical Society, p. 023512. DOI: 10.1103/PhysRevD.102.023512. URL: https://link.aps.org/doi/10.1103/PhysRevD.102.023512.

Aurrekoetxea, Josu C. et al. (May 2020). “The effects of potential shape on inhomogeneous inflation”. In: *Journal of Cosmology and Astroparticle Physics* 2020.5, p. 030. ISSN: 1475-7516. DOI: 10.1088/1475-7516/2020/05/030. URL: https://dx.doi.org/10.1088/1475-7516/2020/05/030.

Baker, Alan (2022). “Simplicity”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2022. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/sum2022/entries/simplicity/.

Baumann, Daniel (Nov. 30, 2012). “TASI Lectures on Inflation”. arXiv: 0907.5424[astro-ph,physics:gr-qc,physics:hep-ph,physics:hep-th]. URL: http://arxiv.org/abs/0907.5424.

Bianchi, Eugenio and Carlo Rovelli (Apr. 11, 2010). *Why all these prejudices against a constant?* DOI: 10.48550/arXiv.1002.3966. arXiv: 1002.3966[astro-ph,physics:gr-qc,physics:hep-th,physics:physics]. URL: http://arxiv.org/abs/1002.3966.

Blau, S. K. and A. H. Guth (Jan. 1, 1987). “Inflationary cosmology.” In: *Three Hundred Years of Gravitation*, pp. 524–603. URL: https://ui.adsabs.harvard.edu/abs/1987thyg.book..524B.

Boyle, Latham, Kieran Finn, and Neil Turok (Dec. 20, 2018). “CPT-Symmetric Universe”. In: *Physical Review Letters* 121.25. Publisher: American Physical Society, p. 251301. DOI: 10.1103/PhysRevLett.121.251301. URL: https://link.aps.org/doi/10.1103/PhysRevLett.121.251301.

— (Mar. 1, 2022). “The Big Bang, CPT, and neutrino dark matter”. In: *Annals of Physics* 438, p. 168767. ISSN: 0003-4916. DOI: 10.1016/j.aop.2022.168767. URL: https://www.sciencedirect.com/science/article/pii/S0003491622000070.

Brandenberger, Robert H. and Jérôme Martin (Apr. 2013). “Trans-Planckian issues for inflationary cosmology”. In: *Classical and Quantum Gravity* 30.11. Publisher: IOP Publishing, p. 113001. ISSN: 0264-9381. DOI: 10.1088/0264-9381/30/11/113001. URL: https://dx.doi.org/10.1088/0264-9381/30/11/113001.

Callender, Craig (Jan. 2004). "There is No Puzzle About the Low Entropy Past". In: *Contemporary Debates in Philosophy of Science*. Malden, MA: Blackwell Publishing Ltd, pp. 240–255. ISBN: 978-1-4051-0151-6.

Carroll, John W. (2020). "Laws of Nature". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2020. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/win2020/entries/laws-of-nature/.

Carroll, Sean M. (Aug. 8, 2019). *Spacetime and Geometry: An Introduction to General Relativity*. Cambridge: Cambridge University Press. ISBN: 978-1-108-77038-5. URL: https://www.cambridge.org/highereducation/books/spacetime-and-geometry/38EDABF9E2BADCE6FBCF2B22DC12BFFE.

Carroll, Sean M. and Heywood Tam (July 8, 2010). *Unitary Evolution and Cosmological Fine-Tuning*. DOI: 10.48550/arXiv.1007.1417. arXiv: 1007.1417[astro-ph,physics:gr-qc,physics:hep-th]. URL: http://arxiv.org/abs/1007.1417.

Chen, Eddy Keming (2023). "The Past Hypothesis and the Nature of Physical Laws". In: *The Probability Map of the Universe: Essays on David Albert's Time and Chance*. Ed. by Barry Loewer, Brad Weslake, and Eric B. Winsberg. Harvard University Press, pp. 204–248. URL: https://philarchive.org/rec/CHETPH.

Chen, Eddy Keming and Sheldon Goldstein (2022). "Governing Without a Fundamental Direction of Time: Minimal Primitivism About Laws of Nature". In: *Rethinking the Concept of Law of Nature: Natural Order in the Light of Contemporary Science*. Ed. by Yemima Ben-Menahem. Cham: Springer International Publishing, pp. 21–64. ISBN: 978-3-030-96775-8. DOI: 10.1007/978-3-030-96775-8_2. URL: https://doi.org/10.1007/978-3-030-96775-8_2.

Clough, Katy et al. (Sept. 2017). "Robustness of inflation to inhomogeneous initial conditions". In: *Journal of Cosmology and Astroparticle Physics* 2017.9, p. 025. ISSN: 1475-7516. DOI: 10.1088/1475-7516/2017/09/025. URL: https://dx.doi.org/10.1088/1475-7516/2017/09/025.

Collins, C. B. and S. W. Hawking (Sept. 1973). "Why is the Universe isotropic?" In: *Astrophys. J.* 180, pp. 317–334. DOI: 10.1086/151965.

Dawid, Richard and Casey McCoy (Nov. 7, 2023). "Testability and viability: is inflationary cosmology "Scientific"?" In: *European Journal for Philosophy of Science* 13.4, p. 51. ISSN: 1879-4920. DOI: 10.1007/s13194-023-00556-3. URL: https://doi.org/10.1007/s13194-023-00556-3.

Di Tucci, Alice et al. (Sept. 16, 2019). "Quantum incompleteness of inflation". In: *Physical Review D* 100.6. Publisher: American Physical Society, p. 063517. DOI: 10.1103/PhysRevD.100.063517. URL: https://link.aps.org/doi/10.1103/PhysRevD.100.063517.

Earman, John and Jesus Mosterin (1999). "A Critical Look at Inflationary Cosmology". In: *Philosophy of Science* 66.1. Publisher: [The University of Chicago Press, Philosophy of Science Association], pp. 1–49. ISSN: 0031-8248. URL: https://www.jstor.org/stable/188736.

Elley, Matthew et al. (May 6, 2024). *Robustness of inflation to kinetic inhomogeneities*. DOI: 10.48550/arXiv.2405.03490. arXiv: 2405.03490[astro-ph,physics:gr-qc,physics:hep-th]. URL: http://arxiv.org/abs/2405.03490.

Ellis, George F R (Jan. 1, 2007). "Issues in the Philosophy of Cosmology". In: *Philosophy of Physics*. Ed. by Jeremy Butterfield and John Earman. Handbook of the Philosophy of Science. Amsterdam: North-Holland, pp. 1183–1285. DOI: 10.1016/B978-044451560-5/50014-2. URL: https://www.sciencedirect.com/science/article/pii/B9780444515605500142.

Garfinkle, David, Anna Ijjas, and Paul J. Steinhardt (Aug. 10, 2023). "Initial conditions problem in cosmological inflation revisited". In: *Physics Letters B* 843, p. 138028. ISSN: 0370-2693.

DOI: 10.1016/j.physletb.2023.138028. URL: https://www.sciencedirect.com/science/article/pii/S0370269323003623.

Guth, Alan H. (Jan. 15, 1981). “Inflationary universe: A possible solution to the horizon and flatness problems”. In: *Physical Review D* 23.2. Publisher: American Physical Society, pp. 347–356. DOI: 10.1103/PhysRevD.23.347. URL: https://link.aps.org/doi/10.1103/PhysRevD.23.347.

Guth, Alan H., David I. Kaiser, and Yasunori Nomura (June 2, 2014). “Inflationary paradigm after Planck 2013”. In: *Physics Letters B* 733, pp. 112–119. ISSN: 0370-2693. DOI: 10.1016/j.physletb.2014.03.020. URL: https://www.sciencedirect.com/science/article/pii/S0370269314001762.

Helbig, Phillip (Mar. 1, 2012). “Is there a flatness problem in classical cosmology?” In: *Monthly Notices of the Royal Astronomical Society* 421.1, pp. 561–569. ISSN: 0035-8711. DOI: 10.1111/j.1365-2966.2011.20334.x. URL: https://doi.org/10.1111/j.1365-2966.2011.20334.x.

— (July 11, 2020). “The flatness problem and the age of the Universe”. In: *Monthly Notices of the Royal Astronomical Society* 495.4, pp. 3571–3575. ISSN: 0035-8711. DOI: 10.1093/mnras/staa1082. URL: https://doi.org/10.1093/mnras/staa1082.

— (Dec. 1, 2021). “Arguments against the flatness problem in classical cosmology: a review”. In: *The European Physical Journal H* 46.1. Number: 1 Publisher: Springer Berlin Heidelberg, p. 10. ISSN: 2102-6459, 2102-6467. DOI: 10.1140/epjh/s13129-021-00006-9. URL: https://epjh.epj.org/articles/epjh/abs/2021/01/13129_2021_Article_6/13129_2021_Article_6.html.

Hempel, Carl Gustav (1965). *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. New York: The Free Press.

Hollands, Stefan and Robert M. Wald (Dec. 2002a). “An Alternative to Inflation”. In: *General Relativity and Gravitation* 34.12, pp. 2043–2055. ISSN: 0001-7701, 1572-9532. DOI: 10.1023/A:1021175216055.

— (Sept. 30, 2002b). *Comment on Inflation and Alternative Cosmology*. DOI: 10.48550/arXiv.hep-th/0210001. arXiv: hep-th/0210001. URL: http://arxiv.org/abs/hep-th/0210001.

Holman, M. (Nov. 1, 2018). “How Problematic is the Near-Euclidean Spatial Geometry of the Large-Scale Universe?” In: *Foundations of Physics* 48.11, pp. 1617–1647. ISSN: 1572-9516. DOI: 10.1007/s10701-018-0218-4. URL: https://doi.org/10.1007/s10701-018-0218-4.

Ijjas, Anna, Paul J. Steinhardt, and Abraham Loeb (June 25, 2013). “Inflationary paradigm in trouble after Planck2013”. In: *Physics Letters B* 723.4, pp. 261–266. ISSN: 0370-2693. DOI: 10.1016/j.physletb.2013.05.023. URL: https://www.sciencedirect.com/science/article/pii/S0370269313003924.

Jensen, Lars Gerhard and Jaime A. Stein-Schabes (Feb. 15, 1987). “Is inflation natural?” In: *Physical Review D* 35.4. Publisher: American Physical Society, pp. 1146–1150. DOI: 10.1103/PhysRevD.35.1146. URL: https://link.aps.org/doi/10.1103/PhysRevD.35.1146.

Kleban, Matthew and Leonardo Senatore (Oct. 2016). “Inhomogeneous anisotropic cosmology”. In: *Journal of Cosmology and Astroparticle Physics* 2016.10, p. 022. ISSN: 1475-7516. DOI: 10.1088/1475-7516/2016/10/022. URL: https://dx.doi.org/10.1088/1475-7516/2016/10/022.

Kofman, Lev, Andrei Linde, and Viatchesiav Mukhanov (Nov. 2002). “Inflationary Theory and Alternative Cosmology”. In: *Journal of High Energy Physics* 2002.10, p. 057. ISSN: 1126-

6708. DOI: 10.1088/1126-6708/2002/10/057. URL: https://dx.doi.org/10.1088/1126-6708/2002/10/057.

Lake, Kayll (May 27, 2005). "The Flatness Problem and Lambda". In: *Physical Review Letters* 94.20. Publisher: American Physical Society, p. 201102. DOI: 10.1103/PhysRevLett.94.201102. URL: https://link.aps.org/doi/10.1103/PhysRevLett.94.201102.

Lehners, Jean-Luc and K. S. Stelle (Oct. 29, 2019). "Safe beginning for the Universe?" In: *Physical Review D* 100.8. Publisher: American Physical Society, p. 083540. DOI: 10.1103/PhysRevD.100.083540. URL: https://link.aps.org/doi/10.1103/PhysRevD.100.083540.

Linde, A. D. (Sept. 22, 1983). "Chaotic inflation". In: *Physics Letters B* 129.3, pp. 177–181. ISSN: 0370-2693. DOI: 10.1016/0370-2693(83)90837-7. URL: https://www.sciencedirect.com/science/article/pii/0370269383908377.

— (Nov. 14, 1985). "Initial conditions for inflation". In: *Physics Letters B* 162.4, pp. 281–286. ISSN: 0370-2693. DOI: 10.1016/0370-2693(85)90923-2. URL: https://www.sciencedirect.com/science/article/pii/0370269385909232.

Maleknejad, A. and M. M. Sheikh-Jabbari (June 5, 2012). "Revisiting cosmic no-hair theorem for inflationary settings". In: *Physical Review D* 85.12. Publisher: American Physical Society, p. 123508. DOI: 10.1103/PhysRevD.85.123508. URL: https://link.aps.org/doi/10.1103/PhysRevD.85.123508.

Martens, Niels C. M. and Dennis Lehmkuhl (Nov. 1, 2020). "Cartography of the space of theories: An interpretational chart for fields that are both (dark) matter and spacetime". In: *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 72, pp. 217–236. ISSN: 1355-2198. DOI: 10.1016/j.shpsb.2020.08.004. URL: https://www.sciencedirect.com/science/article/pii/S1355219820301106.

McCabe, Gordon (Aug. 1, 2018). "Inflationary cosmology and the scale-invariant spectrum". In: *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 63, pp. 39–49. ISSN: 1355-2198. DOI: 10.1016/j.shpsb.2017.11.002. URL: https://www.sciencedirect.com/science/article/pii/S1355219817301272.

Melia, Fulvio (July 10, 2021). "Classicalization of quantum fluctuations at the Planck scale in the Rh=ct universe". In: *Physics Letters B* 818, p. 136362. ISSN: 0370-2693. DOI: 10.1016/j.physletb.2021.136362. URL: https://www.sciencedirect.com/science/article/pii/S0370269321003026.

Mukhanov, Viatcheslav and Sergei Winitzki (2007). *Introduction to Quantum Effects in Gravity*. Cambridge: Cambridge University Press. ISBN: 978-0-521-86834-1. DOI: 10.1017/CBO9780511809149. URL: https://www.cambridge.org/core/books/introduction-to-quantum-effects-in-gravity/23A1F0C929D8985A5A3DB319E9808AE8.

Penrose, R. (1989). "Difficulties with inflationary cosmology". In: *Annals N. Y. Acad. Sci.* 571. Ed. by E. J. Fenyves, pp. 249–264. DOI: 10.1111/j.1749-6632.1989.tb50513.x.

Piran, Tsvi (Dec. 4, 1986). "On general conditions for inflation". In: *Physics Letters B* 181.3, pp. 238–243. ISSN: 0370-2693. DOI: 10.1016/0370-2693(86)90039-0. URL: https://www.sciencedirect.com/science/article/pii/0370269386900390.

Price, Huw (2002). "Burbury's Last Case: The Mystery of the Entropic Arrow". In: *Royal Institute of Philosophy Supplement* 50, pp. 19–56. DOI: 10.1017/s1358246100010493.

Tod, Paul (May 2010). "Penrose's Weyl curvature hypothesis and conformally-cyclic cosmology". In: *Journal of Physics: Conference Series* 229.1, p. 012013. ISSN: 1742-6596. DOI: 10.1088/1742-6596/229/1/012013. URL: https://dx.doi.org/10.1088/1742-6596/229/1/012013.

Wald, Robert M. (Oct. 15, 1983). “Asymptotic behavior of homogeneous cosmological models in the presence of a positive cosmological constant”. In: *Physical Review D* 28.8. Publisher: American Physical Society, pp. 2118–2120. DOI: 10.1103/PhysRevD.28.2118. URL: https://link.aps.org/doi/10.1103/PhysRevD.28.2118.

— (1984). *General Relativity*. Chicago, USA: Chicago Univ. Pr. DOI: 10.7208/chicago/9780226870373.001.0001.

— (Sept. 1, 2006). “The arrow of time and the initial conditions of the universe”. In: *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*. The arrows of time, 2006 37.3, pp. 394–398. ISSN: 1355-2198. DOI: 10.1016/j.shpsb.2006.03.005. URL: https://www.sciencedirect.com/science/article/pii/S1355219806000396.

Wolf, William J. and Karim P. Y. Thébault (Mar. 17, 2023). “Explanatory Depth in Primordial Cosmology: A Comparative Study of Inflationary and Bouncing Paradigms”. In: *The British Journal for the Philosophy of Science*, p. 725096. ISSN: 0007-0882, 1464-3537. DOI: 10.1086/725096. arXiv: 2210.14625[astro-ph,physics:gr-qc,physics:physics]. URL: http://arxiv.org/abs/2210.14625.

Woodward, James and Lauren Ross (2021). “Scientific Explanation”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2021. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/sum2021/entries/scientific-explanation/.