International Journal of Theoretical Physics, Vol. 23, No. 8, 1984

Can Inflation Explain the Second Law of Thermodynamics?

Don N. Page

Center for Theoretical Physics, The University of Texas at Austin, Austin, Texas 78712
Department of Physics, The Pennsylvania State University, University Park, Pennsylvania 16802[1]

Received September 5, 1983

The inflationary model of the universe can explain several of the cosmological conundra that are mysteries in the standard hot big bang model. Paul Davies has suggested that inflation can also explain the second law of thermodynamics, which describes the time asymmetry of the universe. Here I note several difficulties with this suggestion, showing how the present inflationary models must assume the arrow of time rather than explaining it. If the second law is formulated as a consequence of the hypothesis that there were no long-range spatial correlations in the initial state of the universe, it is shown how some of the cosmological conundra might be explained even without inflation. But if the ultimate explanation is to include inflation, three essential elements remain to be demonstrated which I list.

The standard hot big bang model of the universe has been very successful in explaining the recession of distant galaxies, the cosmic microwave background radiation, and the abundances of helium and other light elements. However, there are a number of mysteries it has not explained:

(1) The flatness problem (Dicke and Peebles, 1979; Guth, 1981). (Why is the energy density of the universe still so near the critical density for the spatially flat $k=0$ Friedmann model? Alternatively, why is gravity a dynamically significant force despite the expansion of the universe to at least 10^{183} Planck volumes?)

(2) The homogeneity problem. (Why does the distribution of superclusters of galaxies appear to be fairly uniform in space?)

(3) The isotropy problem. (Why do distant galaxies show no statistically preferred directions, and why is the microwave background radiation isotropic to at least a few parts in 10^4?)

[1] Present address.

(4) The horizon problem (Rindler, 1956). (How could the universe possibly become highly homogeneous and isotropic if the separate regions we can now see were never in previous causal contact according to the standard model?)

(5) The galaxy-formation problem. (Why were the perturbations from precise homogeneity and isotropy of the right form for galaxies to form?)

(6) The monopole problem (Zeldovich and Khlopov, 1978; Preskill, 1979). (Why are monopoles now much less numerous than baryons, though grand unified theories (GUTs) tend to have about as many produced?)

(7) The problem of the second law of the thermodynamics. (Why did the universe start out highly ordered, enabling the entropy to increase to give an arrow of time despite the fact that all known microscopic dynamical laws are time reversible in the sense of being CPT invariant?)

As a suggestion for solving the horizon and flatness problems, Alan Guth (1981) proposed the inflationary universe model in which an early "false vacuum" phase with negative pressure equal to energy density produced a gravitational repulsion and led to an exponential expansion. All of the observed universe would have expanded from one very small region and would have had time for causal contact, thereby eliminating the horizon problem. If the exponential expansion occurred sufficiently long, the resulting spatial hypersurface would be very large and nearly flat, thereby solving the flatness problem if a transition to a radiation phase (positive pressure) could occur on such a hypersurface. As a bonus, the inflation could greatly dilute the monopoles before baryons are produced, so the monopole problem would also be solved.

The original inflationary model (Guth, 1981) had the difficulty that the phase transition ending the inflation occurred in bubbles rather than on a constant-time spatial hypersurface, so the resulting universe would not be at all homogeneous and isotropic. A new inflationary scenario (Linde, 1982; Hawking and Moss, 1982a; Albrecht and Steinhardt, 1982) allowed the possibility of a more homogeneous phase transition, though calculations (Guth and Pi, 1982; Hawking, 1982a; Bardeen et al., 1983; Hawking and Moss, 1982b) indicated that fluctuations would lead to perturbations from homogeneity and isotropy much larger than those observed unless the effective potential for the matter fields had a rather special form. More work is needed to show which models, if any, can solve problems (2), (3), and (5) above.

Paul Davies (1983) has argued that the inflationary universe scenario also explains the time asymmetry described by the second law of thermodynamics. In his description, the universe starts in an arbitrary state with spacetime irregular on all length scales at the Planck time (10^{-43} s). Expansion cools it below the GUT temperature, whereupon the stress–energy tensor is dominated by the false-vacuum contribution proportional to the metric (i.e., a perfect fluid with pressure $p = -\rho$, the negative of a constant energy density $\rho = \rho_0$). The situation corresponds to a positive effective cosmological constant which leads to gravitational repulsion. The resulting exponential expansion of the universe during this inflationary de Sitter phase smooths out the initial inhomogeneities and anisotropies, thereby greatly decreasing the gravitational entropy density. However, the false-vacuum quantum state is unstable and eventually undergoes a phase transition to a thermal state. This dumps an enormous amount of entropy into the matter fields and also makes the pressure positive ($p = \rho/3$) so that the universe no longer expands exponentially but by a power law in time. The gravitational entropy is much less than that of the matter but can now try to catch up by the clumping of mass to form galaxies, stars, and black holes. The continuing expansion also allows the out-of-equilibrium matter entropy to increase by nucleosynthesis as hydrogen is converted into helium and heavier elements. These two processes cause the present universe to have a distinct arrow of time.

Davies' account is a nice description of the thermodynamic processes that occur in the inflationary model, but I would dispute the claim that inflation *explains* the time asymmetry rather than merely assuming it (Page, 1983a). As an explanation rather than a description of the second law, the inflationary model as outlined by Davies appears to be open to at least three objections:

(1) For inflation to get started, it must be assumed that the universe cools below the GUT temperature. If the universe in the Planck era is in a state of thermal spacetime foam with maximum entropy as Davies assumes, it is by no means obvious that this state will evolve into something else or, if it does, that this something else will be an expanding universe which is described by a temperature which drops below the GUT value. One apparently has to invoke the second law or something similar to get into the inflationary phase. Alexander Vilenkin (Vilenkin, 1982, 1983) has emphasized this problem and has suggested that the universe was created from "nothing" in a quantum tunneling event via the Hawking–Moss instanton (Hawking and Moss, 1982a) from a state with no classical spacetime to a state with de Sitter space and the Higgs field at a maximum of its effective potential. This would eliminate the question of what initial conditions lead to inflation, but it raises the problem of showing that the tunneling is more likely to go to an inflationary state rather than some other state.

(2) It must be assumed that the inhomogeneities and anisotropies are decaying rather than growing during the inflationary de Sitter phase. It can indeed be shown (Boucher and Gibbons, 1983) that within the cosmological event horizon of any future inextendible timelike geodesic, the metric asymptotically approaches that of de Sitter space exponentially fast. However, this asymptotic analysis does not show what will happen during a

finite inflationary phase. If the irregularities are viewed as perturbations of a background de Sitter space, the temporal invariance of the latter implies that there is no *a priori* reason for supposing the perturbations get smaller with time. Nothing in the CPT-invariant dynamical equations implies this time asymmetry.

Of course, it is natural to assume that the initial perturbations spread out with time and get weaker rather than come together and get stronger, but this is a thermodynamic assumption. This unexplained time-asymmetric assumption is only "natural" because that is how we see nature behave. We do not naturally see highly focused incoming radiation, so we postulate a time-asymmetric second law of thermodynamics to exclude it. The inflationary scenario relies on this assumption and is consistent with it, but inflation does not explain it.

(3) In order that the present universe be as homogeneous and isotropic as it is observed to be, the phase transition from the inflationary phase must produce a large amount of matter entropy without a comparable amount of gravitational entropy (i.e., inhomogeneities and anisotropies). This has been a problem with the original (Guth, 1981) and with the new (Linde, 1982; Hawking and Moss, 1982a; Albrecht and Steinhardt, 1982; Guth and Pi, 1982; Hawking, 1982a; Bardeen et al., 1983; Hawking and Moss, 1982b) inflationary scenario, though it may be possible with a suitable effective potential for the matter fields having sufficiently restricted coupling constants so that the entropy is channeled almost entirely into the matter.

Thus it does not appear that the inflationary model by itself can *explain* the second law of thermodynamics. One needs an additional assumption to restrict the possible initial states for the universe. One such hypothesis, arising out of a quantum version of the law of conditional independence (Penrose and Percival, 1962) is that the universe began without long-range spatial correlations in its quantum state. Roughly speaking, this means that if the universe at some initial hypersurface $t=0$ were divided into n disjoint spatial regions, the density matrix $\rho(0)$ for gravitational and matter fields on the complete hypersurface would be the tensor product of the n density matrices $\rho_i(0)$ for the fields in each region. (Of course, this can only be approximately true, for if the fields on opposite sides of a boundary between regions were completely uncorrelated, the resulting discontinuities in the fluctuating fields would give infinite contributions to the energy. There is also the uncertainty in quantum gravity of what the initial hypersurface is and how to define the different spatial regions on it.)

If the quantum state evolves unitarily, the microscopic entropy on a later hypersurface $t>0$,

$$S(t) \equiv -\operatorname{tr}\rho(t)\ln\rho(t) = S(0) \tag{1}$$

retains its original value, but the coarse-grained entropy, obtained by adding up the individual entropies of the separate regions, increases (at least initially) as the correlations that develop between the regions are ignored:

$$S_{\text{coarse}}(t) \equiv \sum_{i=1}^{n} -\operatorname{tr}\rho_i(t)\ln\rho_i(t) > S_{\text{coarse}}(0) \qquad (2)$$

That is, information is not really lost globally but goes into spatial correlations which become inaccessible to local observations, so it appears that information is lost and that the universe becomes more disordered as the second law describes. The absence of initial long-range correlations also means it is statistically improbable for fluctuations to be focused so as to come together and get much stronger (e.g., as advanced radiation would do). Thus this hypothesis would also imply that during an exponentially expanding phase the irregularities would tend to spread out and get weaker as the inflationary scenario assumes.

However, the hypothesis of no initial long-range spatial correlations might explain the observed homogeneity and isotropy of the universe even without an inflationary phase (Page, 1983b). If the expectation values of the metric and matter fields are homogeneous and isotropic, which seems most natural and presumably can be assumed without loss of generality as a consequence of a superselection rule associated with coordinate invariance, then inhomogeneities and anisotropies represent quantum fluctuations from the mean. These are statistically unlikely to have macroscopically coherent values over large distances unless there are long-range spatial correlations in the quantum state. If such correlations were absent initially, the large-scale inhomogeneities and anisotropies could only have developed casually during the expansion of the universe. Note that this modification of Hawking's "principle of ignorance" (Hawking, 1976) is not subject to Penrose's criticism (Penrose, 1979, 1981) that inhomogeneous and anisotropic gravitational fields (e.g., black holes) have higher entropy, for the quantum state is not assumed here to maximize the entropy. Indeed, Penrose's suggestion (Penrose, 1979, 1981) of an initially vanishing Weyl tensor appears to be a consequence (at least on large scales) of the hypothesis of no initial long-range correlations.

If the homogeneity and isotropy of the universe as well as the second law of thermodynamics can be explained by this one hypothesis even without inflation, one might ask whether the other cosmological conundra can likewise be explained without it. Indeed, the flatness problem (Dicke and Peebles, 1979; Guth, 1981) might be explained by the weak anthropic principle (Carter, 1974) applied to the quantum state of the universe. That is, life might be possible only in those components of the universal wave function in which space grows large enough and gravitational forces remain strong enough for galaxies, stars, and planets to form. Then our observation of the very close balance between expansion and gravity would be merely a selection effect conditioned by our existence.

The horizon problem (Rindler, 1956) is not really an independent problem but rather one of the difficulties of solving the large-scale homogeneity and isotropy if the distant regions that appear remarkably similar were never in causal contact during their past evolution. If the homogeneity and isotropy are explained by the hypothesis of no initial long-range correlations, then particle horizons do not necessarily have to be eliminated. Indeed, in classical general relativity or even in quantum field theory in a classical spacetime metric, the initial conditions are given over an acausal hypersurface, so particle horizons are always present in some form or other. The situation is less clear in quantum gravity, since the quantum uncertainties in the light cones make the whole concept of horizons rather fuzzy.

The galaxy-formation problem has as yet no satisfactory solution in any model. The inflationary models are very beautiful for being among the first to give concrete predictions for the perturbations. Unfortunately, even the new inflationary scenario (Linde, 1982; Hawking and Moss, 1982a; Albrecht and Steinhardt, 1982) predicts perturbations far larger than those observed unless the effective potential for the matter fields takes a special form (Guth and Pi, 1982; Hawking, 1982a; Bardeen et al., 1983; Hawking and Moss, 1982b). It is not yet known whether an initial state without long-range spatial correlations can lead to the right perturbations in a noninflationary model.

The monopole problem (Zeldovich and Khlopov, 1978; Preskill, 1979) is one problem for which inflation has so far appeared to be the most attractive solution. This problem arises *because* of the assumption that the Higgs field was spatially uncorrelated initially, which leads to knots in the field when the internal symmetry was spontaneously broken in an independent manner in each causally disjoint region of the universe (Einhorn et al., 1980; Guth and Tye, 1980). Except for some rather contrived models in which monopole–antimonopole pairs become confined by flux tubes (Lazarides and Shafi, 1980; Linde, 1980; Langacker and Pi, 1980; Bais and Langacker, 1982), it appears difficult to get enough such pairs to annihilate to be consistent with the upper bounds today (Zeldovich and Khlopov, 1978; Preskill, 1979; Goldman et al., 1981; Fry, 1981; Dicus et al., 1982; Page, 1983b), unless the universe is inflated by a very large factor to dilute the monopoles below the current upper limits. Of course, it might simply be that the grand unified theories are wrong and that there are no monopoles to be produced, but then we would have given up the present explanation for the quantization of electric charge and for the excess of baryons over antibaryons.

If many (though perhaps not all) of the cosmological conundra can be solved by the hypothesis that the initial quantum state of the universe had no long-range spatial correlations, we still might want a simpler hypothesis to explain this one and pin down more precisely the initial state. One bold new approach to this is Hawking's idea (Hawking, 1982b; Hartle and Hawking, 1983; Hawking, 1983) that the state on any closed spatial hypersurface is given by a path integral over all compact Euclidean 4-geometries (i.e., positive-definite four-dimensional metrics) and matter fields bounding the three-dimensional hypersurface in question. This proposal predicts a unique quantum state for the universe, which, if correct, would solve the mystery of the second law of thermodynamics and the other cosmological problems, once the correct dynamical laws (i.e., the action in the path integral) are known. Of course, it is far beyond our present computational powers to calculate this universal wave function for any realistic candidate for the action, but it may be done approximately for certain simple minisuperspace models (Hartle and Hawking, 1983; Hawking, 1983). If these approximately soluble models can be made sophisticated enough, they may be able to test whether Hawking's proposal can give the qualitative behavior of no long-range spatial correlations at early times or other attributes ascribed to the second law.

It is conceivable that if some such proposal can give a unique state for the universe, it may include an inflationary phase which plays a role rather like what Davies envisages. However, there seem to be three essential elements to be demonstrated for this to work:

(1) The quantum state should make it highly probable for the universe to start out small (say in Planck volumes when the energy density, or some such variable representing time but not conjugate to the volume, has the Planck value) and then get very large. But if it is more probable for the universe to start off small, why is it not simply more probable for the universe never to get very large? We might invoke the weak anthropic principle (Carter, 1974) to say that we can only observe those components of the quantum state in which the universe does get very large, but then we have the alternative problem of explaining why in these components it was more probably once small (i.e., deflated) rather than always being large in Planck units (as for example is the standard hot big bang model if cut off at the Planck time). There appear to be far more configurations available in the classical phase space, at least, in which a space-time is always large as measured on Cauchy spatial hypersurfaces whose extrinsic curvature is everywhere bounded by the Planck value.

(2) It should be shown that when small, the universe necessarily has low entropy. A conjecture phrased in the language of classical general relativity but using Planck units would be that on any spatial hypersurface the entropy is bounded by the inequality

$$S \lesssim f^{1/4}(\max|R_{\alpha\beta\gamma\delta}|)^{3/4} V \tag{3}$$

where f is the number of quantum field species (including gravity), $R_{\alpha\beta\gamma\delta}$ are the Riemann curvature components in the set of orthonormal frames whose timelike unit vectors are normal to the hypersurface, and V is the 3-volume of the hypersurface. (This inequality was chosen so as to be approximately saturated for radiation-dominated Friedmann–Robertson–Walker models, in which the right-hand side stays approximately constant. Alternately, in an anisotropic collapse the right-hand side gets arbitrarily large if one lets the hypersurface get sufficiently near the singularity, thus allowing the bound to be satisfied even if the gravitational entropy grows indefinitely, provided it does not grow too fast.) However, a similar conjecture in quantum gravity would require a suitable replacement of the four-dimensional curvature components. One would also need a definition of entropy that can allow it to increase. A quantity which grows with some measure of the spatial correlations of the quantum state, such as the coarse-grained entropy in equation (2), might be suitable, but there is the question of how to define such a quantity precisely.

(3) The inflationary phase which expands the universe from a small to a large volume should be shown to end in a phase transition which increases the matter entropy enormously without increasing the gravitational entropy by a comparable amount. As discussed above, this may require the action to have a special form.

It is not yet obvious that these three elements will necessarily occur in the correct model for the universe, but if they can be shown to do so, then we might have an inflationary explanation for the second law of thermodynamics. Alternatively, a model giving the state of the universe might explain the second law and hence the arrow of time even without an inflationary phase.

ACKNOWLEDGMENTS

The ideas expressed here have arisen out of conversations and correspondence with many people, including Marek Demianski, Gary Gibbons, Alan Guth, Stephen Hawking, André Linde, Baidyanath Misra, Roger Penrose, Ilya Prigogine, Alexie Starobinsky, Alexander Vilenkin, Ethan Vishniac, Steven Weinberg, and John Wheeler. This work was supported in part by grants from the NSF (PHY-8117464 and PHY-8205717) and the Alfred P. Sloan Foundation.

REFERENCES

Albrecht, A., and Steinhardt, P. J. (1982). *Phys. Rev. Lett.*, **48**, 1220–1223.
Bais, F. A., and Langacker, P. (1982). *Nucl. Phys.*, **B197**, 520–532.
Bardeen, J. M., Steinhardt, P. J., and Turner, M. S. (1983). *Phys. Rev. D*, **28**, 679–693.

Boucher, W., and Gibbons, G. W. (1983). In *The Very Early Universe*, G. W. Gibbons, S. W. Hawking, and S. T. C. Siklos, eds. Cambridge University Press, Cambridge.

Carter, B. (1974). In *IAU Symposium No. 63: Confrontation of Cosmological Theories with Observational Data*, M. S. Longair, ed. Reidel, Dordrecht.

Davies, P. C. W. (1983). *Nature*, **301**, 398–400.

Dicke, R. H. and Peebles, P. J. E. (1979). In *General Relativity: An Einstein Centenary Survey*, S. W. Hawking and W. Israel, eds. Cambridge University Press, Cambridge.

Dicus, D. A., Page, D. N., and Teplitz, V. L. (1982). *Phys. Rev. D*, **26**, 1306–1316.

Einhorn M. B., Stein, D. L., and Toussaint, D. *Phys. Rev. D*, **21**, 3295–3298.

Fry, J. N. (1981). *Astrophys. J. Lett.*, **246**, L93–L97.

Goldman, T., Kolb, E. W., and Toussaint, D. (1981). *Phys. Rev. D*, **23**, 867–875.

Guth, A. H. (1981). *Phys. Rev. D*, **23**, 347–356.

Guth, A. H. and Pi, S.-Y. (1982). *Phys. Rev. Lett.*, **49**, 1110–1113.

Guth, A. H., and Tye, S.-H. H. (1980). *Phys. Rev. Lett.* **44**, 631–635, 963.

Hartle, J. B., and Hawking, S. W. (1983). *Phys. Rev. D*, **28**, 2960–2975.

Hawking, S. W. (1976). *Phys. Rev. D*, **14**, 2460–2473.

Hawking, S. W., (1982a) *Phys. Lett.*, **115B**, 295–297.

Hawking, S. W. (1982b). In *Astrophysical Cosmology: Proceedings of the Study Week on Cosmology and Fundamental Physics*, H. A. Brück, G. V. Coyne, and M. S. Longair, eds. Pontifica Academia Scientiarum, Vatican.

Hawking, S. W. (1983). Lectures at the NATO Summer School on Relativity, Groups, and Topology in Les Houches, France, 21 June–4 August 1983, and *Nucl. Phys. B*, in press.

Hawking, S. W., and Moss, I. G. (1982a) *Phys. Lett.*, **110B**, 35–38.

Hawking, S. W., and Moss, I. G. (1982b). University of Cambridge, preprint.

Langacker, P., and Pi, S.-Y. (1980). *Phys. Rev. Lett.*, **45**, 1–4.

Lazarides, G., and Shafi, Q. (1980). *Phys. Lett.*, **94B**, 149–152.

Linde, A. D. (1980). *Phys. Lett.*, **96B**, 293–296.

Linde, A. D. (1982). *Phys. Lett.*, **108B**, 389–393.

Page, D. N. (1983a). *Nature*, **304**, 39–41.

Page, D. N. (1983b). In *The Very Early Universe*, G. W. Gibbons, S. W. Hawking, and S. T. C. Siklos, eds. Cambridge University Press, Cambridge.

Penrose, O., and Percival, I. C. (1961). *Proc. Phys. Soc.*, **79**, 605–616.

Penrose, R. (1979). In *General Relativity: An Einstein Centenary Survey*, S. W. Hawking and W. Israel, eds. Cambridge University Press, Cambridge.

Penrose, R. (1981). In *Quantum Gravity 2: A Second Oxford Symposium*, C. J. Isham, R. Penrose, and D. W. Sciama, eds. Clarendon Press, Oxford.

Preskill, J. P. (1979). *Phys. Rev. Lett.*, **43**, 1365–1368.

Rindler, W. (1956). *Mon. Not. R. Astron. Soc.*, **116**, 662–677.

Vilenkin, A. (1982). *Phys. Lett.*, **117B**, 25–28.

Vilenkin, A. (1983). *Phys. Rev.* **D27**, 2848–2855.

Zeldovich, Ya. B. and Khlopov, M. Yu. (1978). *Phys. Lett.*, **79B**, 239–241.