Rsug

Bull. Astr. Soc. India (1988) **16**, 1–9

Fifty years in cosmology*

Fred Hoyle *Cockley Moor, Dockray Penrith, Cumbria, England*

I graduated at Cambridge university in 1936, the year in which Hubble and Humason published their famous paper on the redshifts of galaxies. I had studied mathematics as an undergraduate, taking part III of the mathematical Tripos in my final year, my main topics in that examination being quantum mechanics, statistical mechanics, and relativity (both special and general). So from a student point of view I already knew a little about cosmology at the time that Hubble and Humason published their paper. In particular, I knew about de Sitter's cosmological model which was to play an important role in subsequent years.

The worrying situation at that time in cosmology, as it seemed, turned out to be relatively minor matter, namely the choice of suitable coordinates. Even the best-known cosmologists—de Sitter, Eddington, and Lemaitre—had chosen coordinates appropriate to localities in the universe rather than to the whole. This produced a sense of mystery that was more apparent than real as to what happened at the boundary of a locality. It is one of the features of Einstein's general relativity that when you choose coordinate systems with special properties you can mistakenly come to think of the properties as physical instead of as mathematical artefacts. Early workers on gravitational waves thought they were investigating physical waves when in fact the waves were in their coordinate system, and a similar situation existed in cosmology.

It was also in 1935–36 that this situation was put right, by H.P. Robertson in the United States and A.E. Walker in Britain, and the resulting choice of coordinates later became known as the Robertson-Walker line element. Then in 1937 Robertson published an important article on cosmology in the *Reviews of Modern Physics*, which unfortunately I did not read at that time because my research interests were in quantum mechanics and nuclear physics.

During the second world war it happened that Hermann Bondi and I worked closely together, and we continued to do so for a year or two after the war when we both returned to Cambridge as junior lecturers in mathematics. My interests were now in astrophysics, and when Bondi decided he was going to make something of a speciality in general relativity I joined him in that study. So it came about that at last in 1945–46 Bondi and I went in great detail through Robertson's

*First B. M. Birla memorial lecture delivered 1987 February 27 at B. M. Birla planetarium, Hyderabad.

article in the *Reviews of Modern Physics*, looking carefully into its fine points as well as into the broader arguments.

The cosmological models favoured by Robertson were of the so-called Friedmann type, which is to say what today we would call big-bang models, the idea being that the universe originated suddenly all in a moment. This view had a fairly wide constituency at that time in the United States, due in considerable measure to Robertson himself and also to George Gamow, but not in Europe. Many Europeans felt the theoretical conclusion of a big-bang origin, arrived at in Robertson's analysis, was a product of simplifying assumptions in the analysis. Notably, it was felt that the assumptions of isotropy and homogeneity were constraining influences on the problem. Lifshitz in the USSR, the collaborator of Landau in the famous Landau-and-Lifshitz textbooks, published an extremely long and complicated paper, in which he claimed that because of departures from homogeneity the big bang was an invalid concept. And at every conference on cosmology and relativity held in Europe in the 1950s, Otto Heckmann, the director of the Hamburg observatory, made a similar claim, at first with respect to departures from isotropy and then with respect to inhomogeneities. I personally found these claims both disturbing and irritating, because coming from people of high standing I felt I ought to understand them and I couldn't. Heckmann in particular was always telling me that some especially clever student of his had demonstrated the matter beyond dispute.

Eventually in the early 1960s I had an especially clever student of my own in the person of Professor Jayant Narlikar, now of the Tata Institute of Fundamental Research. The first issues I asked Narlikar to investigate were the claims of Lifshitz and Heckmann. Rather as I had expected he found them to be wrong. Narlikar was soon able to offer a simple proof that departures from homogeneity and isotropy cannot in themselves prevent the phenomenon of the big bang. A more ambitious proof of the same result was given a few years later by Hawking and Ellis. To complete this aspect of my story, it must have been in 1970 or thereabouts that Narlikar and I published a very different idea for invalidating the big bang, namely that the effects of quantum mechanics would need to be considered at the earliest moments of the universe, and as such might make the concept of the big bang meaningless. Since then, Professor Narlikar and his students have proved this to be the case. Because of quantum mechanics there can be no big bang in the sense the concept is widely used by those many commentators in the media and even in scientific journals—commentators who are unfortunately all too often ignorant of quantum mechanics or at least of its subtler aspects. To put the matter a little more technically, because of quantum mechanical uncertainities in the line element, spacetime singularities do not occur, a result that is also applicable to so-called black holes, vitiating many of the things which are commonly said about black holes.

Meanwhile as early as 1947–48 a few of us in Cambridge were investigating a new physical idea in cosmology, namely that matter might be subject to a continuous form of creation. At first, Hermann Bondi would have none of it, although his close friend Tommy Gold was rather in favour of it. I was myself neutral to the idea. I realised in 1947, when Bondi and Gold turned to other ideas, that if continuous creation were to have any hope of acceptance it would have to be given a mathematical expression. In the latter part of 1947, I came to the conclusion that a new form of field would be needed, and that a scalar field was not only the simplest possibility but also the most promising. I wrote the field on paper as a capital C, and from then on it became known as the C-field. In January 1948 I found how to use the C-field in a modification of Einstein's equations with the result that the equations had as a particular solution what became later known as the steady-state model. This, let me emphasise, was not a *static* model but one in which the main features of the universe are steady like a steadily flowing river The universe expands but it does not become increasingly empty because new matter is constantly being created to make up the deficit produced by the expansion. By the end of February 1948 I had written my paper 'A new model for the expanding universe' in the form in which it was eventually published in the *Monthly Notices of the Royal Astronomical Society*, after being rejected by the *Proceedings of the Physical Society* and by the *Physical Reviews*. On 1948 March 1 I gave a colloquium on the new model at the Cavendish laboratory, at which both the great pioneers of quantum mechanics were present, Paul Dirac and Werner Heisenberg. Heisenberg had been invited to Cambridge for a six month period which he had spent at my own college, St. John's, and because of this I had got to know him quite well. Possibly this was the reason why I heard later that, after his return to Germany, he had said that the concept of a steady-state model was the most interesting thing he had heard during his stay in Britain.

Naturally I showed my paper to Bondi and Gold. Bondi saw immediately that his difficulty about the conservation of energy had been answered. The remarkable thing about the C-field was that its energy density was negative. As matter with positive energy was created the energy of the C-field became more negative. In the flat spacetime of special relativity this would have led to a creation catastrophe, with matter being created at an ever increasing rate as the C-field became more and more negative. But in general relativity, which is to say with gravitation present, this did not happen. The C-field was gravitationally self-repellant, so that as matter was created it was forced apart by the C-field, thereby maintaining a steady balance. Thus at a stroke two crucial features of the universe were explained, its matter content and its expansion. Neither had to be arbitrarily assumed, as in the big-bang models which had been discussed by H. P. Robertson. The universe expanded because it was forced to expand, not because it had arbitrarily been created in a state of explosion.

In March and April of 1948, Bondi and Gold then conceived of a remarkable point of view. Instead of regarding a steady state universe as a deductive consequence of a set of mathematical equations, as one normally does in theoretical physics, they conceived of it as a philosophical axiom, which they referred to as the 'perfect cosmological principle'. From their perfect cosmological principle they were then able to obtain the same geometrical structure for the universe as I had obtained from the mathematical equations, namely what is usually known as the de Sitter line element, and thereafter the discussion became similar in its astrophysical consequences to mine. By about May I think it was that they had written a paper to this effect which they sent to the *Monthly Notices of the Royal Astronomical Society*. In the ensuing months there occurred the first of the circumstances which have sometimes caused me to regret that I ever had anything to do with cosmology, for owing to my paper being rejected by two journals the time delay involved in its successive rejections led to its eventually being printed several weeks later than the Bondi-Gold paper. The lesser aspect of this inversion in printing of the order in which the papers had actually been written led to a mix-up over priority, which was compounded by the fact that Bondi and Gold had actually discussed the steady-state idea verbally, but without doing anything definite about it, as early as 1946, ahead of my C-field idea of 1947. So in the general confusion I was never able, even to myself, to make up my mind as to exactly how the history should be fairly started, and in the event I decided to say nothing at the time, leaving the situation to come out as it would.

Of much more scientific relevance than priority, was an immense difference of emphasis between my paper and the Bondi-Gold paper. My paper simply said: 'Here is a new cosmological model to be discussed along with other models'. Because of their central philosophical axiom Bondi and Gold could not take this guarded position. They had to come out and say assertively that of necessity the steady-state model must be the correct model. No question about it from their point. This had two main effects : It caused the theory to be attacked more ferociously than my point of view would have done, and it provoked far more discussions than mine would have done.

Although through out the 1950s the three of us were thrown together in order to defend the theory, I must confess that I did not myself have much liking for the physical aspects of the 'perfect cosmological principle' and already in 1949 I wrote a paper critical of it. The perfect cosmological principle required the universe to be unchanging with respect to time. But quite evidently localities within the universe are indeed changing with time. What was it that decides, I asked, the scale at which there is a change from local change to universal invariances? This question has lain for almost forty years unanswered. As I shall indicate at the end of my talk, it may well be the most relevant question of all, unfortunately asked long before its time. My suggestion in 1949 as to its answer would have been that localities deviating from a steady-state condition might have dimensions of the order of thirty million light years. By the early 1960s, Narlikar and I had increased this estimate tenfold, to about three hundred million light years. Today I would increase it to at least the greatest distances at which galaxies and quasars are observed.

The 1950s were noteworthy for two quite different developments, one theoretical, the other observational. The theoretical challenge was to improve the mathematical elegance of the theory. I made the mistake myself of sticking to the physical equations, whereas a friend, M. H. L. Pryce, used a more abstract approach known as an action principle. This gave a better classical formulation of the steady-state theory. On the observational side, attempts were made to disprove the steady-state theory by showing that the astrophysical properties of galaxies change with time, which is to say with respect to redshift. Many disproofs were claimed. Some were withdrawn and others were maintained with increasing emphasis as the years passed by. Tommy Gold was the most outspoken of us in replying to these criticisms. He pointed out that the greater the distance of a galaxy the fainter it became and the more difficult it necessarily was to make observations accurately. Tommy's point was that progressively increasing errors with increasing distances were being falsely interpreted as physical changes—in other words the claims were artefacts arising from errors of observation. In all cases known to me, this riposte from Gold has turned out to be correct. With the much more sensitive observations available today, no astrophysical property shows evidence of evolution such as was claimed in the 1950s to disprove the steady-state theory. In particular, the strong claims of Martin Ryle have turned out to be wrong, as Gold always said they would. Technically speaking again, the luminosity function of radio galaxies has turned out to be invariant with respect to redshift, the opposite of what Ryle claimed. If all this had been known in 1960, the steady-state theory would then have been considered proven, and the development of cosmology following the discovery of the microwave background in 1965 would probably have been very different.

In 1963–64 I gave a course of lectures at Cambridge on relativity and cosmology in the preparation for which I went carefully over the work of George Gamow and his colleagues on the synthesis of light elements in a hot big-bang model of the universe. It seemed that a calculation of the helium/hydrogen ratio to be expected in such a model could be improved, and together with Roger Taylor I set out to make the necessary calculations. In such a model there is a present-day microwave background temperature. Taylor and I found that if we knew this temperature we could calculate a cosmic value for the hydrogen-helium ratio, and vice-versa. If we knew a cosmic value for the hydrogen-helium ratio we could infer what the present day microwave background temperature had to be. But when we examined the astronomical literature concerning hydrogen-helium ratios determined by astronomical means we found a wide range of values corresponding to helium abundances by mass ranging from a low of about 15% to a high of about 40%. This was vastly too broad a range for anything useful to be inferred about a possible microwave background temperature.

It must have been in 1964 that I was sitting beside lake Camo in Italy, with Bob Dicke from Princeton university. Dicke told me that his group at Princeton was setting up an experiment to look for a possible microwave background, and that they were expecting a temperature of about 20K. I said this was much too high, because a background, if there was one, could not have a temperature above 3K, the excitation temperature of molecular lines of CH and CN found by Mckellar in 1940. Shortly after that the background was found at the Bell telephone laboratories by Penzias and Wilson, and it had a temperature almost exactly on Mckellar's value. The big mistake Bob Dicke and I had made was not to realize we had it there beside lake Camo, in our coffee cups. However carefully one guards against it, opportunities like this come and then slip away through one's fingers.

The discovery of an actual microwave background made it profitable to calculate the light element synthesis problem more ambitiously than Taylor and I had done, and in 1966–67 Bob Wagoner, Willy Fowler and I set ourselves to do this. Interesting results were confined to just four light nuclei, ^{2}D, ^{3}He, ^{4}He, and ^{7}Li. From the results we were able to show how astrophysical measurements of cosmic values for these light nuclei could be used to infer the properties of a hot big-bang model, assuming the latter to be correct and assuming the astrophysical measurements to be truly cosmic. Since then, Bob Wagoner has periodically updated these calculations, obtaining slightly differing results as the physics of the problem has changed somewhat over the years, for example, by there being three types of neutrino instead of the two types used in the first calculations.

This work with Wagoner and Fowler was my last essay in cosmology by orthodox methods. For the following eight years, up to 1975, I was heavily occupied in administration, and since 1975, my thoughts have run in other directions. This does not mean that I have lost interest in cosmology, but rather that I have sat around waiting for something significantly new to happen. Despite immense numbers of people swarming into cosmology in the United States, in Europe, and in the Soviet Union, nothing very profound seems to me to have happened over the past twenty years. Ask one of the younger generation what evidence they would offer for the correctness of a hot big-bang model and the chances are that they would say, first, the existence of the microwave background, and, second, the synthesis of D, ^{3}He, ^{4}He and ^{7}Li. After that there wouldn't be much to offer. So the situation remains essentially as it was in 1965–66, which I regard as a distinctly bad omen for the theory. Always in the past, whenever a correct theory has been established, a decade or more of rapid progress has been forthcoming, not a state of stagnation more or less. Indeed the one interesting thing to emerge in the early 1980s was a partial reversion to the steady-state model, which came about in the following way.

From a properly based scientific point of view the discovery of the microwave background did not come as an unmixed blessing to protagonists of the hot big-bang model. It was soon found that the background had a remarkable large-scale isotropy. It was almost the same coming from regions of the sky diametrically opposed to each other, despite such regions never having been in communication with each other at any time in the past in such a model. Attempts made to explain this large-scale uniformity with respect to direction in terms of physical processes failed, leaving proponents of the model to fall back on the arbitrary supposition that the background was isotropic because it was created that way at the origin of the universe. Indeed the model required every important observable aspect of the universe to be derived from the manner of its creation, essentially making it just as impossible to understand anything on rational scientific grounds as if one were to believe in the first page of the Christian Bible, which actually, I had to suspect was playing an important role in the minds of those who supported the model.

Narlikar and I, had given already in the early 1960s an explanation of isotropy and homogeneity in terms of the steadystate model, which after a sufficient number of generations simply expands away initial irregularities, just as irregular motions in a gas disappear if the gas expands adiabatically to a sufficient extent. From about 1980 onwards this idea was taken over in what were called inflationary scenarios. An inflationary scenario has three parts to it. There is an initial big bang, then a steady-state phase even in some scenarios down to the operation of a scalar field to that which I had postulated at the end of 1947, and finally there is freely expanding phase like one of the Friedmann models studied by H. P. Robertson. The initial big bang has no observable function in such a scenario, for nothing in it lives through the steady-state phase to come down to us today as an observable entity. The microwave background, the creation of matter, and the irregularities which become the galaxies, all belong to the second phase the steady-state phase. Thus the initial big bang is superfluous like the attempts made in the 1930s to give quantum mechanical systems unobservable internal variables in the hope of restoring determinism. An inflationary scenario functions just as well if the initial big bang is omitted, in which case the universe is steady-state followed by a freely expanding Friedmann model, and is the same so far as astrophysics is concerned as a model studied by Narlikar and me in 1966, a model which we referred to as a 'bubble universe'. Our perceptions differed from an inflationary scenario only, so far as I can see, in them being dominantly mathematical rather than physical. The equations which relate the geometrical behaviour of the universe to its physical content and to the creation of matter are nonlinear, and it is a mathematical feature of nonlinear equations that as well as possessing non-unique ordinary solutions they can possess a unique singular solution. What we found was that our equation had a unique singular solution and it was this solution that yielded the steady-state model with creation of matter. The ordinary solutions on the other hand were analogous to models of the Friedmann type without creation of matter. Our idea in the bubble universe was that there might be switches in particular localities out of a universal singular solution into a regional ordinary solution, yielding a Friedmann type bubble embedded in a steady-state universal ocean. Or of course many such bubbles. The difference between this model and an inflationary scenario is that our switches in the mathematical solutions were conjectural, whereas it is now claimed in an inflationary scenario that the switches can be understood in terms of modern supersymmetry theories in particle physics. If people generally and cosmologists in particular would get away from this fixation with a mock-Biblical big bang and think more about the relation of a bubble universe with particle physics in it than I think there would be the best chance yet of relating cosmology to astrophysics, a relation which to this point has been almost non-existent.

The first big issue to get straight is that the mean density existing in the basic steady-state cannot be superdense. Superdense conditions with super-symmetries in particle physics playing a dominant role, exist in local objects not smoothly everywhere. The typical mean density is that which we observe in galaxies. Indeed the galaxies are aggregates of material left over in our particular bubble from its former steady-state condition, a mean density typically in the range 10^{-24} to 10^{-21} gm cm^{-3}. Stars condense everywhere in the steady-state, producing a radiation background with a temperature that can be calculated to be necessarily around 300k. In the expansion of our bubble the mean density has fallen by about a million and the temperature of the background by about a hundred. The origin of the radiation background was simply stellar radiation thermalised by immense quantities of dust.

This picture is close to being proven. Indeed I would say that it is proven by the observed fine-scale isotropy of the microwave background. If the background came from a universal superdense state, where it had uniformity and isotropy, the later propagation of the radiation in non-uniform gravitational field which must have happened in such a theory when, clusters of galaxies form should have produced measurable non-uniformity in the background on the scale of clusters of galaxies. This is not observed. Production of radiation by stars rather uniformly distributed, and by thermalization due to dust, must produce an exceedingly uniform background on the other hand. This is because the dust, being rather insubstantial in its mass, would be subject to immense forces due to radiation pressure, were there any appreciable fluctuations in the intensity of the radiation background, causing the dust to adjust itself, quickly so as to produce a very uniform situation.

The conventional notion that life originates on the earth is the greatest running force in scientific history. Anyone with a little physical sense should be able to see from the complexity and function of proteins that not even a single enzyme could ever be produced by random processes here on the earth. Only if the whole universe is typically at a steady-state temperature of 300K over a very long span of time can the origin of life be understood in rational terms. The ideas currently held by biologists appeal just as much to irrational miracles as did the creationists who preceded them.

In the old steady-state theory of the 1950s and 1960s the balance between creation of matter and the expansion of the universe was thought to be stable. This I now think was a mistake, caused by regarding the creation process as being spatially uniform. If, however, creation occurs in localized objects, of which the quasars are perhaps the still recognizible remnants in our region of the universe, then quite likely the balance between creation and expansion is unstable, in which case every locality in the universe would be oscillatory, approximated to by a closed Friedmann model of finite volume except near minimum phase when the ambient C-field becomes strong enough to produce an intense burst of creation at a multitude of quasarlike centres of activity. The resulting sharp increase of the C-field then blows the locality back into an expanding phase. On this view the important question then is to decide how big are the oscillating localities, thereby determining the period of the oscillations. On this view our locality, at present in an expanding phase, will eventually fall back on itself, contracting until the mean density rises to 10^{-24}–10^{-21} gm cm^{-3}, the temperature of the microwave background rises to 300K, and the C-field becomes sufficient to produce another intense round of creation and of the birth of a multitude of stars, preceding yet another expanding phase. The picture is of a multitude of expanding and contracting bubbles with an immense flash of creation occurring as each bubble reverses from contraction to expansion, and with the whole ensemble of bubbles forming a kind of dynamic steady-state universe.

There is a furter line of agreement to which I attach great weight, that demands a universe of this type, and which rules out the purely Friedmann models. Many of our basic physical equations are time symmetric; for instance the equations which determine the generation and propagation of light are time symmetric. The usual assumption that only the past-to-future solutions of such equations exist in the universe seems artificial and unsatisfactory. A theory in which both past-to-future and future-to-past solutions are generated equally in every local radiation process seems dictated by considerations of completeness and elegance. In such a time-symmetric theory what we observe locally is a sum of radiation generated locally, of radiation received from the past and of radiation received from the future. The latter is an addition to usual considerations, and by including it the possibility exists that the sum of the three contributions for a time-symmetric theory turns out to be the same as for the usual time-symmetric theory. For this to be possible, the universe itself must have an overall expansion which is of the steady-state type. The Friedmann models will not do, they give a wrong summation. This matter was first discussed some 40 years ago, so far as classical theory is concerned, by Wheeler and Feyman. Narlikar and I showed in 1968 that the same result holds in non-relativistic quantum mechanics, and in 1970 we extended our proof to relativistic quantum electro dynamics. The fact that an overall steady-state structure for the universe permits local radiation processes to be time-symmetric and yet leads to the imposition of normal cause and effect on the flow of events has always been to my mind a guarantee that this form of cosmological theory will turn out to be basically correct.