Mon. Not. R. astr. Soc. (1986) **223,** 539–555

Alternatives to dark matter

R. H. Sanders *European Southern Observatory, D-8046 Garching, Federal Republic of Germany; and Kapteyn Astronomical Institute, 9700 AV Groningen, The Netherlands*

Accepted 1986 June 26. Received 1986 June 26; in original form 1986 April 21

Summary. The required matching of falling disc rotation curves with rising halo rotation curves in order to produce flat rotation curves in the outer parts of spiral galaxies covering a wide range in mass and size places strong constraints on dark matter hypotheses. This 'disc–halo conspiracy' is explained naturally by radical suggestions that the law of gravity is not the usual inverse square law in the limit of large distances from mass concentrations. The two recently suggested modifications of Newtonian gravity are considered in the light of general constraints on classical gravity theories and observations of mass discrepancies in galaxies and groups of galaxies. The relativistic theory of Bekenstein and Milgrom can satisfy most requirements on a gravity theory, but relativistic versions of finite length scale anti-gravity suggested by the present author are either inconsistent with geodesic motion of particles locally or contain a ghost field. Observational contradictions are apparent for both suggested hypotheses. A third theory, derived by a trivial modification of the Bekenstein–Milgrom Lagrangian, can account for the present observed systematics of mass discrepancies in the Universe and can predict from the observed distribution of visible matter the detailed rotation curves of several well-observed spiral galaxies ranging in size from 8 to 80 kpc.

1 Introduction

There is no evidence for significant mass discrepancies within the bright optical discs of spiral galaxies. This has been demonstrated by several studies in which the rotation curve calculated from the radial light distribution is compared with the observed rotation curve (Kalnajs 1983; Athanassoula, Bosma & Papaioannou 1985; Kent 1986). In general, within some standard photometric radius (R_{25}) the calculated curve matches the observed curve in detail. It is primarily those cases in which the rotation curve as measured in the 21-cm line of neutral hydrogen is seen to extend well beyond the optical disc where significant mass discrepancies are apparent (van Albada *et al.* 1985; Carignan & Freeman 1985). In such cases the observed rotation curve can be accounted for (in the context of Newtonian gravity) only by the addition of a massive dark

component, or, equivalently, by a mass-to-light ratio in the disc which rapidly increases in the outer regions to local values exceeding 1000. Of course the decomposition of an observed rotation curve into disc and dark halo contributions is not unique but there are several arguments, summarized by van Albada & Sancisi (1986), which support the suggestion that the maximum disc solution is generally correct. That is, the rotation curve in the very inner regions is entirely dominated by the visible disc and the ratio of unseen to visible matter within the optical disc is considerably less than unity. The detailed correspondence of photometrically calculated rotation curves to the observed curve strongly supports this suggestion (Kent 1986).

Combined with the fact that the rotation curves in the outer regions of galaxies are observed to be flat, the maximum disc hypothesis has two very dramatic consequences.

(1) It is the luminous matter that sets the constant asymptotic value of the rotation curve. This is already evident from the existence of a well-defined luminosity–rotation velocity relationship (the Tully–Fisher law).

(2) The flat rotation curve in the outer regions is due to the careful matching of a falling disc rotation curve with a rising halo rotation curve. This requires a very careful conspiracy between the luminous and dark matter in galaxies differing by factors of 100 in luminosity and 10 in size (Bahcall & Casertano 1985; van Albada & Sancisi 1986).

This 'disc–halo conspiracy' places severe constraints on the dark matter hypothesis. van Albada & Sancisi (1986) have demonstrated that simple dynamical coupling between a disc and halo cannot explain the conspiracy, and that a fine tuning of halo parameters is required to produce a flat rotation curve in any given galaxy. This might suggest that the conspiracy results from an important aspect of galaxy formation. Perhaps discs form in pre-existing haloes (Fall & Efstathiou 1980; Blumenthal *et al.* 1986), but then why should luminous matter determine the constant value of the rotation curves? Perhaps discs are the seeds for formation of dark haloes (Fabian 1985), but then why should the haloes have precisely the correct core radius and velocity dispersion to produce flat rotation curves in combination with the discs?

A very simple and natural solution of the conspiracy problem is provided if we are willing to consider the possibility that there is no dark matter – that it is the inverse square law of gravity that breaks down on scales comparable with galaxies. In this regard the two most recently suggested hypotheses are the modified Newtonian dynamics (MOND) of Milgrom (1983a, b, c) and the finite length scale anti-gravity (FLAG) of the present author (Sanders 1984, 1986, hereafter referred to as Papers 1 and 2). In the context of these radical suggestions, the law of gravity assures a flat rotation law either at low accelerations (MOND) or beyond a certain length scale (FLAG). Since there is no dark matter the value of the constant rotational velocity is of course set by the visible matter; i.e. Tully–Fisher laws are inherent to both hypotheses (albeit different laws).

It should be emphasized that these radical hypotheses are not, in original form, theories of gravity. They are empirically motivated *ad hoc* suggestions. Moreover, they modify the Newtonian inverse square law while the more complete theory of gravity is general relativity. There have been many attempts to modify general relativity (Will 1979), but none of these attempts has been experimentally motivated. This is because there are no local experimental contradictions to general relativity. But it is being suggested here that experimental contradictions do appear in the limit of very weak fields or large distance from mass concentrations where astronomical observations are now becoming increasingly precise. If this is true the theory of gravity should now be enlarged to include this new class of phenomena.

What then is necessary for a more complete theory of classical gravity? The basic theoretical requirements are as follows.

(1) General covariance: the form of the field equations should be independent of a general coordinate transformation and in particular should reduce to the dynamics of special relativity in flat space.

(2) The field equations and equation of motion should be derivable from a least-action principle. This assures compatibility of the equations and conservation of energy–momentum.

(3) The theory should avoid unphysical artefacts such as tachyons or ghost fields.

The experimental constraints are the following.

(1) The theory should predict to high precision geodesic motion of particles locally.

(2) It should be consistent with the classical tests of general relativity.

(3) It should predict the observed rate of decay for the binary pulsar.

(4) The implied cosmology should preserve the experimental successes of the standard hot big bang: primarily the nucleosynthesis of the light elements.

Taking the point of view that there is no dark matter, a fifth experimental constraint must be added.

(5) The theory should correctly predict the observed rotation law in the outer part of spiral galaxies from the observed distribution of visible matter.

In this paper I consider the two suggested modifications of Newtonian gravity in the light of the general constraints. Bekenstein & Milgrom (1984) have written a relativistic theory of MOND which satisfies most, if not all, of the above constraints (the cosmological constraint has not yet been investigated). However, relativistic theories of FLAG are either inconsistent with the concept of geodesic motion of particles or contain a ghost field which is usually considered to be an unphysical attribute. An interesting third theory is obtained by a trivial revision of the relativistic Lagrangian of Bekenstein & Milgrom (1984). This is a generalized Brans–Dicke theory which contradicts no local experiment and which in the weak field limit combines certain aspects of MOND and FLAG.

I critically examine these three theories in the light of observations of mass discrepancies in galaxies and groups of galaxies. While contradictions are apparent for both MOND and FLAG, the modified Bekenstein–Milgrom theory is consistent with the presently observed aspects of mass discrepancies in the Universe.

2 Theoretical and local experimental constraints on modified gravity theories

2.1 MODIFIED DYNAMICS

In original form (Milgrom 1983a, b, c), this was a modification of Newton's second law of motion in the limit of weak accelerations; i.e.

$$F=ma\mu(a/a_0) \tag{1}$$

where $\mu=1$ if $a\gg a_0$ and $\mu(x)=x$ if $a\ll a_0$. Here a_0 is an acceleration characteristic of the outer parts of galaxies ($2\times10^{-8}h_{50}^2$ cm s^{-2}). In the limit of very low accelerations, the circular velocity about a point mass (Kepler's law) is given by

$$V=(Ga_0m)^{1/4}. \tag{2}$$

This very simple modification, *ad hoc* though it is, contains immediately two very dramatic consequences.

(1) The rotation curves in the outer parts of galaxies should be flat independently of the size or mass of the galaxy.

(2) The value of the asymptotic velocity is set by the visible mass. Assuming that the mass-to-light ratio in spiral galaxies is approximately constant, there should exist a luminosity–rotation velocity relationship (Tully–Fisher) of the form $L \propto v^4$.

In other words the disc–halo conspiracy problem is solved.

This dynamical modification has been criticized on the basis that the linear momentum of an isolated system is not conserved (Felton 1984). This criticism was answered by Bekenstein & Milgrom (1984) who reframed Milgrom's original suggestion as a modification of gravity rather than dynamics. The modified gravitational field equation (in the Newtonian limit) may be written as

$$\nabla \cdot \left[\mu \left(\frac{\nabla \phi}{a_0} \right) \nabla \phi \right] = 4\pi G \varrho \tag{3}$$

which is shown to be equivalent to the assumptions of MOND in axisymmetric galaxies.

Bekenstein & Milgrom also describe a toy relativistic theory which is a generalization of Brans–Dicke theory. In Brans–Dicke theory (Brans & Dicke 1961) a universal scalar field ϕ is added to the usual tensor field $g_{\mu\nu}$ of general relativity. This postulated scalar field couples to all particles in the same way and can be absorbed into the metric to form a new metric $\tilde{g}_{\mu\nu}$ which satisfies revised field equations. Particles follow geodesics of this new metric. The theory is described by the action

$$S = \int \left(\phi R + \frac{16\pi G_0}{c^4} L_{\mathrm{m}} - \omega \frac{\phi_{,\alpha} \phi^{,\alpha}}{\phi} \right) \sqrt{-g}\, d^4x \tag{4}$$

where, as usual, R is the scalar curvature, L_{m} is the matter Lagrangian and the other symbols and conventions are as in Brans & Dicke (1961). Here ω is the free parameter of the theory and is assumed to be constant in the original Brans–Dicke theory. In terms of a scalar coupling constant α_{s},

$$\omega = \frac{1}{\alpha_{\mathrm{s}}} - \frac{3}{2} \tag{5}$$

(Weinberg 1972). Therefore, $\omega \to -3/2$ corresponds to a very strong coupling of the scalar (relative to gravity), and $\omega \to \infty$ corresponds to very weak coupling. The observed orbital decay of the binary pulsar (Taylor & Weisberg 1982) constrains ω, by plausibility arguments, to be quite large (>1000). If $\omega < 1000$ then the masses of the two neutron stars must be identical to within about $0.06\, M_{\odot}$ to prevent prohibitively rapid decay of the orbit by gravitational dipole radiation (Eardley 1975). With such weak coupling of the scalar field the predictions of Brans–Dicke are essentially identical with those of general relativity and the theory becomes uninteresting.

In the Bekenstein–Milgrom generalization, the parameter ω is not a constant but is a function of the magnitude of the scalar gradient; specifically,

$$\omega = \frac{1}{2} (2\omega_0 + 3) \frac{F(x)}{x} - \frac{3}{2} \tag{6}$$

where

$$x = \frac{c^4}{4} \frac{\phi_{,\alpha} \phi^{,\alpha}}{(2\omega_0 + 4)^2 a_0^2 \phi^3}. \tag{7}$$

Here a_0 is the Milgrom acceleration parameter and F is a function which is not specified but which is required to have certain asymptotic behaviour:

$$F(x)\approx x\,(x\gg 1) \qquad F(x)\approx \tfrac{2}{3}x^{3/2}\,(x\ll 1). \tag{8}$$

Near mass concentrations $(x>1)\,\omega\to\omega_0$ which may be very large (weak coupling), but far from mass concentrations $(x<1)\,\omega\to -3/2$ and the scalar coupling dominates.

In the weak field limit this theory reduces to a two-potential theory with the field equations given by

$$\nabla^2\phi_1=4\pi G(1-\lambda)\varrho \tag{9a}$$

$$\nabla\cdot\left[\mu\left(\frac{\lambda\nabla\phi_2}{a_0}\right)\nabla\phi_2\right]=4\pi G\lambda\varrho \tag{9b}$$

where

$$\lambda=\frac{1}{2\omega_0+4} \tag{10a}$$

$$\mu(x)=\frac{dF}{dx}. \tag{10b}$$

This is equivalent to the single-potential Newtonian theory described above (equation 3) and reduces to the phenomenology of MOND in cases of high symmetry.

Bekenstein & Milgrom point out that scalar waves can propagate faster than the speed of light but that this cannot induce acausal effects in the behaviour of particles or electromagnetic fields. Geodesic motion is built into the theory so there are no problems with violations of the weak equivalence principle (WEP). The local tests of general relativity place lower limits upon ω_0 but here this parameter may be taken to be arbitrarily large.

Thus this theory appears to satisfy most of the general requirements for a relativistic theory, although the cosmological connection remains unclear (see Appendix). A fundamental criticism is that the theory is ugly because it contains two parameters (a_0 and ω_0) and one arbitrary function which are not specified naturally from first principles.

2.2 FINITE ANTI-GRAVITY

This idea is broadly based on supergravity theories which predict the existence of bosons that mediate a force with a coupling to matter roughly comparable with that of gravity (Gibbons & Whiting 1981). Such a hypothetical particle may have a non-zero mass; i.e. this additional force has a finite length scale. The total 'gravitational' potential of a point mass would then be written with a Yukawa component, i.e.

$$U=\frac{G_\infty M}{r}\left[1+\alpha\exp\left(-\frac{r}{r_0}\right)\right]. \tag{11}$$

If the mass of the mediating boson is very small (10^{-26} eV) then the corresponding length scale is comparable with the size of galaxies (20–30 kpc). Flat rotation curves of galaxies are then possible within $0.5\,r_0<r<2.5\,r_0$ if $\alpha=-0.92$; i.e. this additional force is repulsive, implying that the exchange boson is a spin 1 or vector particle (Paper 1). In this picture gravity is $1/r^2$ on scales much smaller than r_0 and $1/r^2$ on scales much larger than r_0 but with a larger effective constant of

gravity. Thus there is a maximum possible conventionally calculated mass discrepancy (12.5 if $\alpha=-0.92$). In the transition region ($r\sim r_0$) the rotational velocity has the constant value

$$V_p=0.57\left(\frac{GM}{r_0}\right)^{1/2} \tag{12}$$

termed the 'plateau velocity' in Paper 2. Therefore, this hypothesis also predicts that the value of the constant rotation velocity is simply related to the visible mass and that the Tully–Fisher law should be $L\propto V^2$.

In the non-relativistic limit, FLAG may be described by a two-potential theory with the Lagrangian

$$L=-\int d^3r\left\{\varrho(\phi_1+\phi_2)+(8\pi G_\infty)^{-1}\left[(\nabla\phi_1)^2+\frac{(\nabla\phi_2)^2}{\alpha}+\frac{\phi_2^2}{\alpha r_0^2}\right]\right\}. \tag{13}$$

The resulting field equations are

$$\nabla^2\phi_1=4\pi G_\infty\varrho \tag{14a}$$

$$\nabla^2\phi_2-\phi_2/r_0^2=4\pi G_\infty\varrho \tag{14b}$$

which give a total potential about a point mass of the form of equation (11).

Problems arise when one attempts to carry this over to a fully relativistic theory. One field is obviously to be identified with normal gravity (g_{00}) but the second field is problematic. Following the suggestion in Paper 1, the second field may be a vector field similar to electromagnetism (except massive). The difficulty here is that a vector field usually couples to an absolutely conserved quantity like charge which in this case would be something like baryon number or quark number (Lee & Yang 1955; Sherk 1979). But gravity couples to the inertial mass which implies that acceleration in a gravitational field should depend upon the composition of the falling material. In particular, the fractional difference in the acceleration of lead and copper should be of the order of

$$\frac{\delta g}{g}=\gamma\left[\left(\frac{A}{M}\right)_{\rm Pb}-\left(\frac{A}{M}\right)_{\rm Cu}\right]\simeq 4\times10^{-3}\gamma \tag{15}$$

where A/M is the 'charge' to mass ratio of the particular substance and γ is the relative coupling of the vector field which in the context of FLAG must be approximately 0.92. In other words there are violations of the WEP at levels which are overwhelmingly forbidden by the sensitive Eötvös experiments (Roll, Krotkov & Dicke 1964; Braginski & Panov 1972) which restrict $\delta g/g$ to less than 10^{-12}. It should be added that, unless a specific interaction of the proposed vector field with the electromagnetic field is written explicitly into the action, light follows null geodesics of the metric implying a deflection about the Sun 10 times greater than observed.

A second possibility for a relativistic theory would be to add two additional fields to the tensor field of gravity: a massless scalar (attractive) and a massive vector (repulsive) which couple to matter with equal strength but 10 times stronger than gravity. On scales smaller than the galaxies the scalar and vector exactly cancel one another but on large scales the repulsive force vanishes leaving normal gravity and a long range scalar which is 10 times stronger than normal gravity. Such a theory (apart from the aspect of a finite range vector) has been suggested (Macrae & Riegert 1984) to account for the precession of the orbit of Mercury in the presence of a large solar oblateness. The difficulty here is that the vector has an additional velocity-dependent coupling to matter (due to the 'gravi-magnetic field') which again leads to violations of the WEP at a level of

$$\delta g/g\sim 2\times10^{-3}\gamma\beta_{\rm E}^2 \tag{16}$$

between lead and copper as measured by solar Eötvös experiments. Here β_E is the orbital velocity of the Earth in units of the speed of light (10^{-4}). Thus the Eötvös experiments constrain the additional fields to be 10 times weaker (not stronger) than gravity.

The basic problem of many-field theories of gravity has been well discussed (Dicke 1965). If the additional fields couple as strongly as gravity then such theories inevitably imply non-geodesic motion of particles and violation of the WEP at a detectable level. A toy relativistic theory which avoids this difficulty can be constructed. The two additional fields are a massless attractive scalar and a massive repulsive scalar which both couple to matter with a strength 10 times greater than gravity. The action would then be

$$S=\int d^4x\left[R-\phi_{,\alpha}\phi^{,\alpha}+\psi_{,\alpha}\psi^{,\alpha}+\frac{\psi^2}{r_0^2}-\frac{1}{\sqrt{-g}}(1+q\phi+q\psi)\int mc^2\,d\tau\right]\sqrt{-g} \tag{17}$$

where q is a scalar charge-to-mass ratio for a given substance. Again locally the two scalar fields almost exactly cancel one another leaving pure general relativity on those scales where it has been tested. Violations of the WEP would still be present but only at a level of $\delta g/g\sim10^{-22}$ in solar Eötvös experiments. On large scales the attractive scalar remains. The difficulty here is that a 'repulsive scalar' is a ghost field, i.e. the energy density is negative (indeed both positive- and negative-energy scalar waves would be emitted by the binary pulsar, leaving the orbital decay due only to gravitational quadrupole radiation). While ghost fields should perhaps not be dismissed in the absence of specific experimental contradictions, there are numerous thought experiments (such as the decay of a particle to a more massive particle with the emission of negative-energy radiation) which make such fields appear to be very unphysical artefacts which a viable theory should avoid. Thus, there is no obvious relativistic theory which contains in its Newtonian limit a repulsive component with a definite length scale and which satisfies the constraints listed in Section 1.

2.3 A REVISED BEKENSTEIN–MILGROM THEORY

A generalized scalar–tensor theory similar to that suggested by Bekenstein & Milgrom would seem to be the most viable candidate for a theory of 'stronger gravity' on a large scale. An interesting modification of the Bekenstein–Milgrom theory (revised MOND or REM) is obtained if one alters the asymptotic behaviour of the function $F(x)$ such that

$$F(x)\simeq x \qquad (x\gg1) \tag{18a}$$

and

$$F(x)=x/p(2\omega_0+4) \qquad (x\ll1) \tag{18b}$$

where p is a constant such that

$$p>1/(2\omega_0+4).$$

This is a generalized Brans–Dicke theory where near mass concentrations ($x\gg1$)

$$\omega\simeq\omega_0 \tag{19a}$$

but far from mass concentrations ($x\ll1$)

$$\omega\simeq-3/2+\varepsilon \tag{19b}$$

where

$$\varepsilon=\frac{1}{2p}\frac{2\omega_0+3}{2\omega_0+4}. \tag{19c}$$

As in the Bekenstein–Milgrom theory the scalar couples much more strongly to matter in regions of very weak fields, but unlike Bekenstein–Milgrom theory there is an upper limit to the strength of the scalar coupling.

This can also be viewed as a theory permitting two extreme values for the constant of gravity. Near mass concentrations

$$G=G_0\frac{2\omega_0+4}{2\omega_0+3} \tag{20a}$$

but far from mass concentrations

$$G_\infty\simeq G(1+p). \tag{20b}$$

In the Newtonian limit (equations 9 and 10) the potential of a point mass would be

$$\phi=Gm/r \tag{21a}$$

where

$$x=G_0M\lambda/r^2a_0\gg 1 \tag{21b}$$

and

$$\phi=G_\infty M/r \qquad x\ll 1. \tag{21c}$$

Thus in the limit of very weak fields the gravitational force is again $1/r^2$ with a larger effective constant of gravity. Moreover, there is a maximum conventionally detected mass discrepancy in the universe (G_∞/G).

The parameters of REM may be chosen such that Kepler's law (the rotational velocity about a point mass as a function of distance) has a form similar to that of FLAG. Taking $\omega_0=4.5\times10^4$ (which is large enough to avoid difficulties with local tests), $p=11.5$ (or $G_\infty/G=12.5$ as in FLAG) and

$$\mu(x)=1-\left(1-\frac{\lambda}{p}\right)\exp(-x^3) \tag{22}$$

which has the desired asymptotic behaviour, we find Kepler's law as shown in Fig. 1. Thus, as in FLAG, there is a transition region between two asymptotic values of the gravitational constant where the circular velocity is constant. The Kepler's law corresponding to FLAG in original form is shown by the broken line in Fig. 1.

Unlike FLAG, however, the effective scale length of the transition region now depends upon the value of the central mass

$$r_0=14.5\left(\frac{GM\lambda}{a_0}\right)^{1/2} \tag{23}$$

as does the value of the plateau velocity,

$$v_p=0.54\left(\frac{Gma_0}{\lambda}\right)^{1/4} \tag{24}$$

which is the same dependence as that of the asymptotic velocity in MOND. Thus, the predicted Tully–Fisher law is $L\propto v^4$. As in pure MOND, REM can account for mass discrepancies in galaxies on all scales, but as in FLAG there is a return on a large scale to $1/r^2$ gravity and a maximum conventionally calculated mass discrepancy. Again I stress that there is nothing special

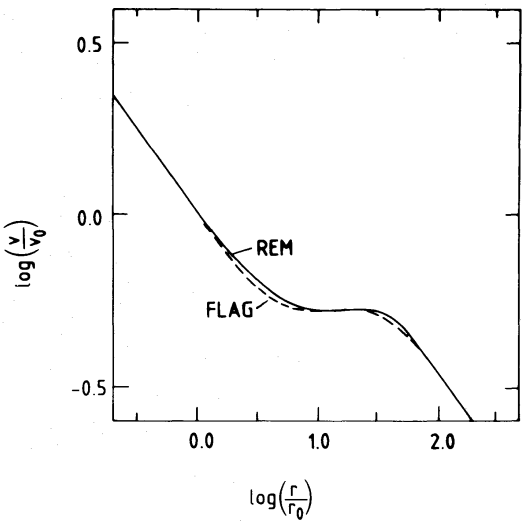

Figure 1. The circular rotational velocity of a test particle orbiting a point mass as a function of distance from the point mass (Kepler's law) for REM (full line) with the parameters described in the text and FLAG (broken line). The unit of distance is $r_0=(GM\lambda/a_0)^{1/2}$ and the unit of velocity is $v_0=(GMa_0/\lambda)^{1/4}$.

about $G_\infty/G=12.5$ or the form of μ taken above; these were chosen to be roughly consistent with FLAG in original form, but larger values of G_∞/G with correspondingly more extensive rotation curve plateaux are also possible.

This revised Bekenstein–Milgrom theory also satisfies the general constraints for a relativistic theory, but can also be criticized for the number (one more than MOND) of unspecified parameters.

3 Confrontation of modified gravity theories with observational aspects of mass discrepancies

3.1 MODIFIED DYNAMICS

Although MOND generally removes mass discrepancies in galaxies and groups of galaxies, there appear to be some cases where it overcorrects. The first example is the dwarf irregular galaxy NGC 3109. The maximum rotation velocity in this galaxy is about 60 km s^{-1} (Carignan 1985). By equation (2) the MOND mass of the galaxy is about $5\times10^8\,M_\odot$ using Milgrom's suggested value of a_0. Distance estimates to this small nearby galaxy range from 1.7 to 2.6 Mpc (Carignan 1985). Taking the lower value we find that the total mass of neutral hydrogen associated with NGC 3109 is $1\times10^9\,M_\odot$ (Huchtmeier, Seiradakis & Materne 1980). Therefore the dynamically calculated mass is less than the observed gas mass not even considering the mass of the stellar disc. The only possible escape from this contradiction is that the distance has been overestimated. The distance estimated from the blue Tully–Fisher relationship as calibrated by Aaronson & Mould (1983) is only 0.78 Mpc and would reduce the combined mass of an $M/L=1$ stellar disc and the observed gas to about $4\times10^8\,M_\odot$. However, this distance estimate is not really independent because MOND is constructed to be consistent with the steeper Tully-Fisher law of Aaronson & Mould.

A second example of overcorrection of an apparent mass discrepancy is provided by the Local Group of galaxies. The Local Group is dominated by M31 and the Galaxy which are separated by about 700 kpc and moving together with a velocity of 90 km s^{-1}. The mass of the Galaxy and M31 may be estimated by assuming that the two galaxies were initially close together separating with the Hubble flow but by now have turned around and are falling back towards one another with the observed velocity and separation (see Peebles 1971 and references therein). The combined mass

turns out to be around $2\times10^{12}\,M_\odot$ which implies a mass discrepancy of 10–20 (assuming normal mass-to-light ratios of the two spiral galaxies). In the context of MOND the assumptions are the same but now the equation of motion to be solved is

$$\frac{d^2r}{dt^2}=-\left(\frac{Gma_0}{r^2}\right)^{1/2}. \tag{25}$$

In this case the combined mass of the galaxies is found to be $7\times10^9\,M_\odot$ which would imply extraordinarily low values of mass-to-light ratios of the two galaxies ($M/L\sim0.1$). There are possible escapes from this argument: perhaps the Galaxy and M31 have normal masses ($\sim5\times10^{10}\,M_\odot$) but form a binary system and have had at least one close encounter in the past, or perhaps the MOND parameter a_0 varies significantly over the time since galaxy formation. The important point, however, is that there is no evidence from the conventional Local Group dynamics of a mass discrepancy of much more than a factor of 10 while Milgrom dynamics considerably overcorrects for this discrepancy.

A third example is provided by several of the clusters considered by Milgrom (1983c) which have extraordinarily low MOND mass-to-light ratios considering the likely contribution of hot gas to the total mass. In particular, the MOND mass of the cluster A2197 is estimated to be $8\times10^{12}\,M_\odot$ (correcting to a radial velocity dispersion of 396 km s^{-1}) whereas the total mass of the X-ray emitting gas is estimated to be $1.8\times10^{13}\,M_\odot$ (Abramopoulos & Ku 1983); i.e. again the actual observed mass apparently exceeds the MOND dynamical mass.

This absence of a maximum predicted discrepancy and resulting overcorrection may be the essential observational problem for MOND. The reason for this is that MOND gravity on large scales is effectively a $1/r$ force and becomes stronger without limit compared with the usual Newtonian inverse square force. This contains possible serious implications for cosmology (Felton 1984), since in the quasi-Newtonian derivation of the expansion equation the physical length scale cannot be cancelled and an isotropic, homogeneous universe is not possible. None of the above arguments taken singly necessarily eliminates MOND, but they do suggest that a possible improvement would be a theory which returns to inverse square gravity on large scales and thereby predicts an upper limit to the conventionally calculated mass discrepancy.

3.2 FLAG

It is not clear that any such revision of Newtonian gravity which is connected to a definite length scale can solve the 'disc–halo conspiracy' problem because galaxies with apparent mass discrepancies span such a large range of size. In Paper 2 it was shown that the rotation curves of several galaxies could be well fitted by such a revision of the gravitational potential if the mass is distributed like the light and if the length scale of the repulsive component is $24\,h_{75}^{-1}$ kpc. For all of the galaxies considered, however, the radial scale length of the equivalent exponential disc varied only between 2.5 and 3.5 kpc. Now there exist kinematic and photometric data on galaxies which are significantly larger and smaller.

UGC 2259 is a small, low surface brightness Sc galaxy which has a very symmetric structure and a measured 21-cm line rotation curve which extends well beyond the optical disc ($2.2\,R_{25}$) (Carignan, Sancisi & van Albada 1986). The kinematic structure is also very symmetrical which suggests that warps or other distortions do not complicate the velocity field. The smoothly rising rotation curve beyond the optical disc and the large conventionally calculated mass-to-light ratio ($14\,h_{75}^{-1}$) imply a large mass discrepancy in this galaxy. There is no published photometry of UGC 2259 but the central surface brightness (22.65 mag arcsec^{-2}) and the optical size ($R_{25}=3.3\,h_{75}$ kpc) imply that the scale length of an equivalent exponential disc is about 1.7 kpc (with the distance

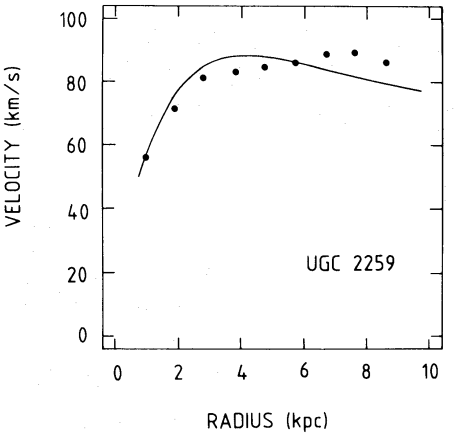

Figure 2. The FLAG rotation curve for UGC 2259 (full line) calculated from the estimated light distribution with a mass-to-light ratio of 7. The points show the observed rotation curve (Carignan *et al.* 1986).

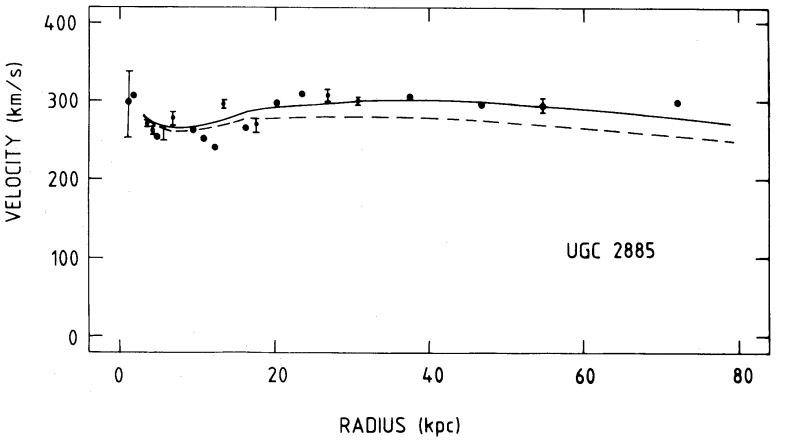

Figure 3. The FLAG rotation curve for UGC 2885 (full line) calculated from the observed light distribution (Kent 1986). The FLAG length scale is 24 kpc as in Paper 2. The points are the observed curve (Rubin *et al.* 1980; Roelfsema & Allen 1985). The broken curve is the rotation curve for the bulge component alone demonstrating that the disc can make only a very small contribution to the total mass in the context of FLAG. This implies an implausibly low M/L of 0.1 for the disc.

equal to $9.8\,h_{75}$ Mpc). The observed 21-cm line rotation curve is shown by the points in Fig. 2 together with the FLAG rotation curve for an exponential disc with the estimated length scale. We see that the general form of the rotation curve is not reproduced by FLAG because the galaxy lies well inside 24 kpc and yet exhibits a significant mass discrepancy.

UGC 2885 is a very large ScI galaxy with a rotation curve which extends to $80\,h_{75}$ kpc (Rubin, Ford & Thonnard 1980; Roelfsema & Allen 1985). Surface photometry of this galaxy (Kent 1986) now shows that the rotation curve inside 45 kpc can be well fitted by the observed light distribution and a reasonable mass-to-light ratio. The observed rotation curve is shown in Fig. 3 together with the FLAG rotation curve determined from the two-component (bulge and disc) light distribution. The fit appears to be reasonable but there is a very serious problem. In this model fit the bulge component must be given a mass of $1.1\times10^{11}\,M_\odot$ in order to account for the high rotational velocities observed in the inner 5 kpc. However, since the plateau velocity (equation 12) of the massive bulge is already 270 km s^{-1} and is attained well within the bright disc, the disc cannot be very massive ($\sim1.8\times10^{10}\,M_\odot$). In Kent's decomposition of the light distribution the luminosity of the bulge and disc respectively are 2.3×10^{10} and 1.9×10^{11}. Thus in

the FLAG model for this galaxy the bulge has all the mass but the disc has all the light and therefore an implausibly low mass-to-light ratio of 0.1. The basic problem is that this galaxy is so large that the entire maximum mass discrepancy (12.5) should be present within the optical disc and yet the global mass-to-light ratio calculated from conventional dynamics (approximately 2) does not imply a large mass discrepancy.

It appears that FLAG cannot solve the disc–halo conspiracy problem in very small or very large galaxies. More generally this is also apparent from the Tully–Fisher relationship. The infrared luminosity – rotation velocity relationship based upon a large sample of data is of the form $L \propto V^4$ (Aaronson *et al.* 1982). In Paper 2, I argued that this did not necessarily contradict the prediction of FLAG because the relation $L \propto V^2$ should apply to those galaxies which are large enough to extend into the regime of modified gravity (see Paper 2, fig. 11). While this argument may be valid it does leave the excellent $L \propto V^4$ correlation for the smaller galaxies (90 per cent of the Aaronson *et al.* sample) unexplained. These arguments taken together make it very unlikely that FLAG, or any such modification attached to a definite length scale, provides the explanation for missing mass in the Universe.

3.3 REVISED BEKENSTEIN–MILGROM THEORY

We have seen that there are observational difficulties with both of the above-suggested modifications of Newtonian gravity. In MOND, gravity on the largest scale is a $1/r$ force which tends, in some cases, to overcorrect the apparent mass discrepancies. In FLAG the modification of Newtonian gravity is attached to a definite length scale, whereas mass discrepancies in galaxies appear over a wide range of size scales. However, REM overcomes both these observational problems.

The parameters of the theory (essentially a_0/λ) may be set by fitting to a well-determined rotation curve like that of NGC 3198 (van Albada *et al.* 1985). In Fig. 4 the rotation curve has been calculated from the visible light distribution (Wevers 1984) by numerical solution of the field

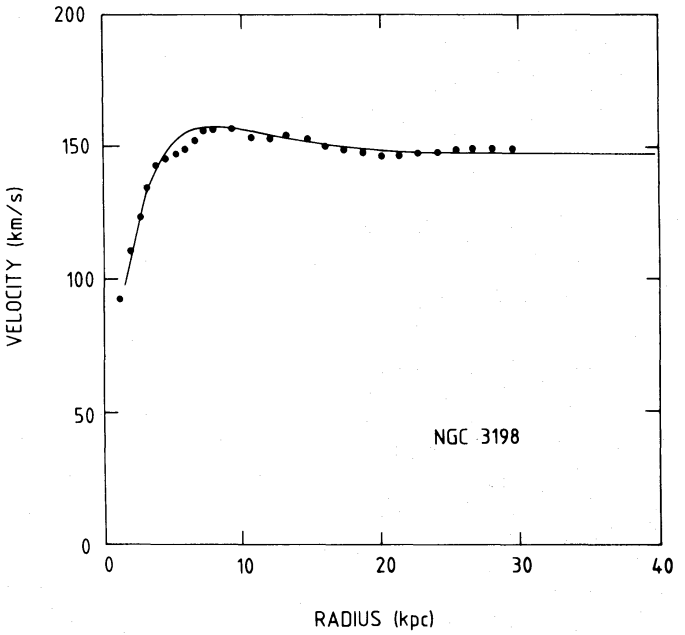

Figure 4. The REM rotation curve (full line) for NGC 3198 calculated from the observed light distribution (Wevers 1984) with $M/L=2.9$. The points are the observed curve (van Albada *et al.* 1985). The REM function μ is assumed to be of the form of equation (22) and the fit to the observed curve sets $a_0/\lambda = 1.8\times10^{-7}\,\mathrm{cm\,s^{-2}}$.

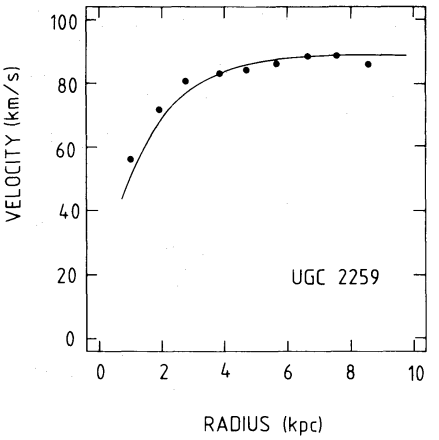

Figure 5. The REM rotation curve for UGC 2259 (full line) calculated from the estimated light distribution with a mass-to-light ratio of 3. The points are the observed curve.

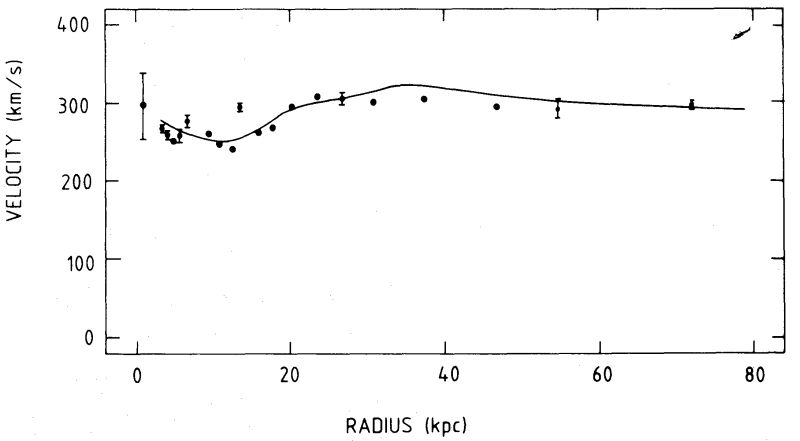

Figure 6. The REM rotation curve for UGC 2885 calculated from the observed light distribution. Here the bulge and disc mass-to-light ratios are 3.7 and 1.6 respectively.

equations (equation 9) with the values of G_∞/G and the form of μ (equation 22) chosen above and $a_0/\lambda = 1.8\times10^{-7}\,\mathrm{cm\,s^{-2}}$. With this same value the rotation curves of the small and large galaxies, UGC 2259 and 2885, may also be calculated from the light distribution (estimated as before in the case of NGC 2259; the results are shown in Figs 5 and 6. For UGC 2259 the mass of the model galaxy is $3\times10^9 M_\odot$ and the corresponding mass-to-light ratio is 3. We see that the calculated rotation curve for UGC 2885 agrees in the inner regions quite well in detail with the observed curve. The total mass of the model galaxy is $3.9\times10^{11} M_\odot$ implying a global mass-to-light ratio of 1.9 with M/L for the bulge and disc components separately being 3.7 and 1.6 respectively. Thus, unlike the original FLAG model the component mass-to-light ratios are plausible.

With respect to mass discrepancies in the individual galaxies described here, the predictions are essentially identical with those of pure MOND. Therefore the overcorrection of the mass discrepancy in NGC 3109 noted above remains problematic for REM. Here an accurate determination of the distance (independent of the Aaronson & Mould calibration of the Tully–Fisher relation) may falsify (but cannot prove) MOND or REM.

Differences between MOND and REM should appear on larger scales. For example, REM would predict decreasing rotation curves beyond some point whereas MOND rotation curves are asymptotically flat. In the context of REM the Local Group analysis would proceed in the conventional way except with a larger effective constant of gravity. If $G_\infty/G = 12.5$, then the

combined mass of the two galaxies would be $1\times10^{11}\,M_\odot$ which would be consistent with reasonable mass-to-light ratios in the two galaxies.

Milgrom's analysis would generally apply to clusters and groups of galaxies, except that the maximum predicted mass discrepancy would be set by G_∞/G. For the cluster considered above (A2197) the virial mass is estimated to be $1.5\times10^{14}\,M_\odot$ (see Oemler 1974). The true mass can only be smaller by a factor of G_∞/G and therefore is $(1.0\text{–}1.5)\times10^{13}\,M_\odot$. This is comparable with (but not less than) the estimated mass of hot gas. The REM dynamics of elliptical galaxies and clusters of galaxies will be discussed in a later paper on polytropic spheres, but as in MOND there is a mass–velocity relation (Faber–Jackson relation) of the form $M\propto\sigma^4$.

In general, such a modification of Newtonian gravity appearing at some critical acceleration (or potential gradient) but returning on larger scales to pure inverse-square gravity would seem to be the most viable alternative to dark matter. It can account for comparable mass discrepancies in small and large galaxies but cannot overcorrect the mass discrepancies in bound groups of galaxies. Moreover, on the largest scale, it is consistent with an isotropic homogeneous universe (see Appendix). But such a theory must predict in detail the rotation curves of galaxies from the observed distribution of visible matter. In the three examples shown the agreement between the predicted and observed rotation curves is impressive, but it must be so for every galaxy. In this sense, such hypotheses are eminently falsifiable.

4 Conclusions

Newtonian gravity was constructed to explain planetary motions and it works very well on the scale of the Solar System. In the context of Newtonian gravity the outstanding success of the dark matter hypothesis was the prediction of the existence of an outer planet perturbing the motion of Uranus and the subsequent discovery of Neptune (see Weinberg 1971 for a discussion of these developments). The pre-relativity attempts to explain the anomalous precession of the perihelion of Mercury by retaining Newtonian gravity and adding dark matter (inner planets or small particles) were less successful. This is because, as we now know, in the limit of a very strong gravitational field new effects appear which must be explained by a more complete theory of gravity – and that is general relativity.

But there is another limit and that is the limit of very weak fields where precise astronomical measurements now probe for the first time. One should not be shocked to find new physical phenomena in this limit. If, for example, there were a universal scalar field which significantly couples to matter only in regions of low scalar gradient – far from large mass concentrations – its effects would not have been noticed except in astronomical observations of large systems.

How could any such idea be proved? It is very easy, in fact, to falsify any specific theory which purports to explain the rotation curves of galaxies from the observed distribution of visible matter. Simply find one galaxy where the predicted rotation curve differs significantly from the observed curve and the theory is eliminated. It is rather more difficult to falsify the dark matter hypothesis where it is the unobserved distribution of dark matter which determines the shape of galaxy rotation curves. Of greater significance, however, are the systematics of mass discrepancies in galaxies which are just now being fully elucidated by accurate observations. If dark matter is a viable explanation for mass discrepancies in the Universe, then it must, in a natural way, solve the disc–halo conspiracy problem. A hypothesis in which at least two dark halo parameters must be fine tuned in every galaxy in order to explain flat rotation curves over a wide range of galaxy mass and size is far less 'economical' than a hypothesis in which, having set one or two universal parameters, every galaxy rotation curve is flat in the outer regions – indeed, in which the observed form of the rotation curve follows in detail from the observed distribution of visible mass. Clearly it is time to take such ideas seriously.

In this paper I have emphasized that there are viable fully relativistic theories which do lead to stronger gravity on larger scales. The revised version of Bekenstein–Milgrom theory suggested here does appear to be consistent with all the presently observed systematics of mass discrepancies in galaxies and groups of galaxies. The problem with this theory is that there is an embarrassing number of parameters which are not yet set by first principles. It should be remembered, however, that even Newton's law of gravity was not a theory of gravity but a phenomenological description, and that Einstein's field equations were designed to be consistent with Newton's law in the limit of weak fields. In this sense, pure intellectual achievement that it was, general relativity was also strongly observationally based, but perhaps upon a range of observations which was still too limited.

Acknowledgments

I thank G. Setti for inviting me to visit the European Southern Observatory where most of this work was done. I am very grateful to M. de Roo and J. Winicour for their patient explanations of aspects of field theory and general relativity. I thank R. Sancisi and L. Lucy for critical comments and discussions and especially T. S. van Albada for a critical reading of the manuscript and many useful comments.

References

Aaronson, M., Huchra, H., Mould, J. R., Tully, R. B., Fisher, J. R., van Woerden, H., Goss, W. M., Chamaraux, P., Mebold, U., Siegeman, B., Berriman, G. & Persson, S. E., 1982. *Astrophys. J. Suppl.,* **50,** 241.
Aaronson, M. & Mould, J., 1983. *Astrophys. J.,* **265,** 1.
Abramopoulos, F. & Ku, W. H.-M., 1983. *Astrophys. J.,* **271,** 446.
van Albada, T. S., Bahcall, J. N., Begeman, K. & Sancisi, R., 1985. *Astrophys. J.,* **295,** 305.
van Albada, T. S. & Sancisi, R., 1986. *Phil. Trans. R. Soc. Lond.,* in press.
Athanassoula, E., Bosma, A. & Papaioannou, S., 1985. In: *Dark Matter in the Universe, IAU Symp. No. 117,* eds Knapp, G. & Kormendy, J., Reidel, Dordrecht, Holland.
Bahcall, J. N. & Casertano, S., 1985. *Astrophys. J.,* **293,** L7.
Bekenstein, J. & Milgrom, M., 1984. *Astrophys. J.,* **286,** 7.
Blumenthal, G. R., Faber, S. M., Flores, R., Primack, J. R., 1986. *Astrophys. J.,* **301,** 27.
Braginski, V. B. & Panov, V. I., 1972. *Sov. Phys. JETP,* **34,** 463.
Brans, C. & Dicke, R. H., 1961. *Phys. Rev.,* **124,** 925.
Carignan, C., 1985. *Astrophys. J.,* **299,** 59.
Carignan, C. & Freeman, K. C., 1985. *Astrophys. J.,* **294,** 494.
Carignan, C., Sancisi, R. & van Albada, T. S., 1986 (in preparation).
Dicke, R. H., 1965. *The Theoretical Significance of Experimental Relativity,* Gordon & Breach, New York.
Eardley, D. M., 1975. *Astrophys. J.,* **196,** L59.
Fabian, A., 1985. In: *Dark Matter in the Universe, IAU Symp. No. 117,* eds Knapp, G. & Kormendy, J., Reidel, Dordrecht, Holland.
Fall, S. M. & Efstathiou, G., 1980. *Mon. Not. R. astr. Soc.,* **186,** 133.
Felton, J. E., 1984. *Astrophys. J.,* **286,** 3.
Gibbons, G. W. & Whiting, B. F., 1981. *Nature,* **291,** 636.
Huchtmeier, W. K., Seiradakis, J. H. & Materne, J., 1980. *Astr. Astrophys.,* **91,** 341.
Kalnajs, A. J., 1983. In: *Internal Kinematics and Dynamics of Galaxies, IAU Symp. No. 100,* ed. Athanassoula, E., Reidel, Dordrecht, Holland.
Kent, S. M., 1986. *Astr. J.,* **91,** 1301.
Lee, T. D. & Yang, C. N., 1955. *Phys. Rev.,* **98,** 1501.
Macrae, K. I. & Riegert, R. J., 1984. *Nucl. Phys.,* **B244,** 513.
Milgrom, M., 1983a. *Astrophys. J.,* **270,** 365.
Milgrom, M., 1983b. *Astrophys. J.,* **270,** 371.
Milgrom, M., 1983c. *Astrophys. J.,* **270,** 384.
Oemler, A., 1974. *Astrophys. J.,* **194,** 1.
Peebles, P. J. E., 1971. *Physical Cosmology,* Princeton University Press, Princeton NJ.

Roelfsema, P. R. & Allen, R. J., 1985. *Astr. Astrophys.*, **146,** 213.
Roll, P. G., Krotkov, R. & Dicke, R. H., 1964. *Ann. Phys. (New York),* **26,** 442.
Rubin, V. C., Ford, W. K. & Thonnard, N., 1980. *Astrophys. J.,* **238,** 471.
Sanders, R. H., 1984. *Astr. Astrophys.,* **136,** L21 (Paper 1).
Sanders, R. H., 1986. *Astr. Astrophys.,* **154,** 135 (Paper 2).
Sherk, J., 1979. *Phys. Lett.,* **88B,** 265.
Taylor, J. H. & Weisberg, J. M., 1982. *Astrophys. J.,* **253,** 908.
Weinberg, S., 1972. *Gravitation and Cosmology,* Wiley, New York.
Wevers, B. M. H. R., 1984. *Doctoral Thesis,* University of Gröningen.
Will, C. M., 1979. In: *General Relativity, An Einstein Centenary,* eds Hawking, S. W. & Israel, W., Cambridge University Press, Cambridge.

Appendix: The Bekenstein–Milgrom field equations and cosmological equations

The action written by Bekenstein & Milgrom (1984) is

$$S=\int \exp\left(-\frac{2\psi}{c^2}\right)\left[R+\frac{6}{c^4}\,\psi_{,\alpha}\psi^{,\alpha}-\frac{2a_0^2\beta\exp(-2\psi/c^2)}{(1+\beta)^2c^4}\,F(x)-\frac{16\pi G_0}{c^4}\,L_{\rm m}\right]\sqrt{-g}\,d^4x \qquad \text{(A1)}$$

where $L_{\rm m}$ is the usual matter Lagrangian and the scalar field ψ is related to the field defined by Brans & Dicke as

$$\phi_{\rm BD}=\exp(-2\psi/c^2) \qquad \text{(A2)}$$

with

$$x=\frac{\exp(2\psi/c^2)\psi_{,\alpha}\psi^{,\alpha}}{a_0^2(1+\beta)^2} \qquad \text{(A3)}$$

$$\beta=2\omega_0+3. \qquad \text{(A4)}$$

By varying the action with respect to $g^{\mu\nu}$ and ψ and setting $\delta S=0$ in the usual way, the following field equations can be derived

$$R_{ij}-\tfrac{1}{2}g_{ij}R=\frac{8\pi G_0}{c^4}\,T_{ij}\exp\left(\frac{2\psi}{c^2}\right)-\frac{2}{c^4}(\psi_{,i}\psi_{,j}+\tfrac{1}{2}g_{ij}\psi_{,\alpha}\psi^{,\alpha})$$
$$+\frac{2}{c^4}\,\mu\beta\psi_{,i}\psi_{,j}-\frac{a_0^2}{c^4}(1+\beta)^2g_{ij}F(x)\exp\left(\frac{2\psi}{c^2}\right)-\frac{2}{c^2}(\psi_{,i;j}-g_{ij}\psi^{,k}{}_{;k}) \qquad \text{(A5)}$$

$$\beta c^2\left[\mu\exp\left(-\frac{2\psi}{c^2}\right)\psi^{,j}\right]_{;j}=-4\pi G_0T^{\alpha}_{\alpha}. \qquad \text{(A6)}$$

Taking the metric for an isotropic, homogeneous universe (Robertson–Walker), the cosmological equations become

$$\left(\frac{\dot R}{R}\right)^2+\frac{k}{R^2}=\frac{8\pi G_0\exp(-2\psi/c^2)}{3}\varrho+\frac{2}{3}\frac{\mu\beta\dot\psi^2}{c^4}-\frac{\psi^2}{c^4}$$
$$+\frac{1}{3}\frac{a_0^2}{c^2}\,\beta(1+\beta)^2F(x)\exp\left(-\frac{2\psi}{c^2}\right)+\frac{2}{c^2}\,\dot\psi\,\frac{\dot R}{R} \qquad \text{(A7)}$$

$$\frac{1}{R^3}\frac{d}{dt}\left[\mu\exp\left(-\frac{2\psi}{c^2}\right)\dot\psi R^3\right]=\frac{4\pi G_0}{\beta}(3p-\varrho) \qquad \text{(A8)}$$

where R is the dimensionless scale factor for the Universe and k is the curvature ($0.\pm1$). In the original Bekenstein–Milgrom theory where the function F has the asymptotic behaviour described in the text (equation 8), the expansion equation is rather difficult and includes terms on the right-hand side of the form

$$\dot{\psi}^3/c^5a_0.$$

Thus the parameter a_0 explicitly enters the expansion equation with c/a_0 appearing as a time-scale. It is not necessarily a disadvantage to have a natural time-scale in addition to the Planck time, particularly since c/a_0 is comparable with the Hubble time. It remains to be seen whether the cosmological solutions in pure Bekenstein–Milgrom theory are viable.

In the revised theory suggested here the equations take a rather simpler form. There are essentially two epochs of expansion, where

$$\frac{\dot{\psi}}{a_0(1+\beta)c}\gtrless 1.$$

In both epochs the cosmology is the usual Brans–Dicke with $\omega\simeq\omega_0$ at earlier times and $\omega=-3/2+\varepsilon$ at later times (see equation 19). In the second epoch the local value of G (i.e. near mass concentrations) may differ from G_∞ at large distances from mass concentrations. At a transition time, t_e, where

$$\dot{\psi}/a_0(1+\beta)c\sim 1$$

the equations become complicated as two extreme values of G become possible.

In REM cosmology the present Universe would be described by Brans–Dicke model with $\omega=-3/2+1/2p$ where $1/2p\simeq 0.05$ (see equation 19). With such strong scalar coupling the only physically viable solution ($\phi_{BD}>0$) would be low density negative-curvature models. The local constant of gravity would vary with time at a rate

$$\frac{\dot{G}}{G}=3p\Omega_0H_0 \tag{A9}$$

where

$$\Omega_0=\frac{8\pi G_0\varrho_0}{3H_0^2\phi_0}. \tag{A10}$$

Measurements of or upper limits on $\dot{G}/G$ could establish or severely constrain such a theory.