1
Introduction

We begin this chapter with an overview in section 1 of how the scalar–tensor theory was conceived, how it has evolved, and also what issues we are going to discuss from the point of view of such cosmological subjects as the cosmological constant and time-variability of coupling constants. In section 2 we provide a simplified view of fundamental theories which are supposed to lie behind the scalar–tensor theory. Section 3 includes comments expected to be useful for a better understanding of the whole subject. This section will also summarize briefly what we have achieved.

In section 1 we emphasize that the scalar field in what is qualified to be called the scalar–tensor theory is not simply added to the tensor gravitational field, but comes into play through the nonminimal coupling term, which was invented by P. Jordan. Subsequently, however, a version that we call the prototype Brans–Dicke (BD) model has played the most influential role up to the present time. We also explain the notation and the system of units to be used in this book.

The list of the fundamental ideas sketched in section 2 includes the Kaluza–Klein (KK) theory, string theory, brane theory as the latest outgrowth of string theory, and a conjecture on two-sheeted space-time. Particular emphasis is placed on showing how closely the scalar field can be related to the "dilaton" as a partner of the graviton in string theory that emerged from an entirely different point of view.

Section 3 will be a collection of comments. We wish to answer potential questions that might be asked by readers who have not yet entered the main body of the book but have nonetheless heard something about the scalar–tensor theory. We also embed certain abstracts or advertisements of the related subjects which will be discussed later in detail. As a result of our doing so we expect that readers may be acquainted beforehand with our achievements from a wider perspective of the whole development. The topics will cover the weak equivalence principle (WEP), parameters of the prototype BD model, conformal transformation, Mach's principle and variable G, and a question about whether there is any advantage to be gained from sticking to a complicated scalar–tensor theory instead of less constrained theories of scalar fields, like the quintessence model.

1.1 What is the scalar–tensor theory of gravitation?

Einstein's general theory of relativity is a geometrical theory of space-time. The fundamental building block is a metric tensor field. For this reason the theory may be called a "tensor theory." A "scalar theory" of gravity had earlier been attempted by G. Nodström by promoting the Newtonian potential function to a Lorentz scalar. Owing to the lack of a geometrical nature, however, the equivalence principle (EP), one of the two pillars supporting the entire structure of general relativity, was left outside the aim of the theory in the early 1910s. This did not satisfy Einstein, who eventually arrived at a dynamical theory of space-time geometry. His theory must have appeared highly speculative at first, but proved later to be truly realistic, since it was supported by observations of diverse physical phenomena, including those in modern cosmology. It also served as an excellent textbook showing how a new way of thinking develops to reality.

In spite of the widely recognized success of general relativity, now called the standard theory of gravitation, the theory has also nurtured many "alternative theories" for one reason or another. Among them we focus particularly on the "scalar–tensor theory." It might appear as if the old idea of scalar gravity were being resurrected. In fact, however, this type of theory does not merely combine the two kinds of fields. It is built on the solid foundation of general relativity, and the scalar field comes into play in a highly nontrivial manner, specifically through a "nonminimal coupling term," as will be explained shortly.

The scalar–tensor theory was conceived originally by Jordan, who started to embed a four-dimensional curved manifold in five-dimensional flat space-time [1]. He showed that a constraint in formulating projective geometry can be a four-dimensional scalar field, which enables one to describe a space-time-dependent gravitational "constant," in accordance with P. A. M. Dirac's argument that the gravitational constant should be time-dependent [2], which is obviously beyond what can be understood within the scope of the standard theory. He also discussed the possible connection of his theory with another five-dimensional theory, which had been offered by Th. Kaluza and O. Klein [3]. On the basis of these considerations he presented a general Lagrangian for the scalar field living in four-dimensional curved space-time:

$$\mathcal{L}_{\rm J} = \sqrt{-g}\left[\varphi_{\rm J}^{\gamma}\left(R - \omega_{\rm J}\frac{1}{\varphi_{\rm J}^{2}} g^{\mu\nu}\,\partial_\mu\varphi_{\rm J}\,\partial_\nu\varphi_{\rm J}\right) + L_{\rm matter}(\varphi_{\rm J},\Psi)\right], \tag{1.1}$$

where $\varphi_{\rm J}(x)$ is Jordan's scalar field, while γ and $\omega_{\rm J}$ are constants, also with Ψ representing matter fields collectively. The introduction of the *nonminimal coupling term*, $\varphi_{\rm J}^{\gamma}R$, the first term on the right-hand side, marked the birth of the scalar–tensor theory. The term $L_{\rm matter}(\varphi_{\rm J},\Psi)$ was for the matter Lagrangian, which depends generally on the scalar field, as well.

For later convenience, we here explain the unit system we are going to use throughout the book. Since we always encounter relativity and quantum mechanics in the area of particle physics, it is convenient to choose a unit system in which c and $\hbar$ are set equal to unity. By doing so we can express all three fundamental dimensions only in terms of one remaining dimension, which may be chosen as length, time, mass, or energy. In particular, the gravitational constant, or Newton's constant, G turns out to have a mass dimension -2, or length squared. We then write

$$8\pi G = c\hbar M_{\rm P}^{-2}, \tag{1.2}$$

with $M_{\rm P}$ called the *Planck mass*, which is estimated to be 2.44×10^{18} GeV, which is quite heavy compared with other ordinary particles. We hereafter choose $M_{\rm P}=1$. In this way we can express every quantity as if it were dimensionless. This unit system is called the *reduced Planckian unit system*, though $G=1$ is often chosen in the *plain* Planckian unit system. We prefer the former system with the difference of $\sqrt{8\pi}$. We show units of length, time, and energy in this system expressed in conventional units:

$$8.07\times 10^{-33}\,\text{cm},\quad 2.71\times 10^{-43}\,\text{s},\quad 2.44\times 10^{18}\,\text{GeV}. \tag{1.3}$$

Sometimes, however, it is convenient to leave one of the dimensions still "floating," not necessarily set fixed. We choose it to be mass, for example, as was shown in (1.2). In the same context, the Lagrangian is found to have a mass dimension 4, while a derivative contributes a mass dimension 1. The metric tensor is dimensionless. If a scalar field has a conventional canonical kinetic term, then we conclude that the field has a mass dimension 1.

Now the second term on the right-hand side of (1.1) resembles a kinetic term of $\varphi_{\rm J}$. Requiring this term to have a correct mass dimension 4 yields the result that $\varphi_{\rm J}$ has mass dimension $2/\gamma$. It then follows that $\varphi_{\rm J}^{\gamma}$, which multiplies R in the first term on the right-hand side of (1.1), has mass dimension 2, the same as G^{-1}. In this way we re-assure ourselves that the first two terms on the right-hand side of (1.1) contain no dimensional constant. This remains true for any value of γ, although this "invariance" under a change of γ need not be respected if $\varphi_{\rm J}$ enters the matter Lagrangian, in general.

Jordan's effort was taken over particularly by C. Brans and R. H. Dicke. They defined their scalar field φ by

$$\varphi = \varphi_{\rm J}^{\gamma}, \tag{1.4}$$

which simplifies (1.1) by making use of the fact that the specific choice of the value of γ is irrelevant, as explained above. This process is justified only because they demanded that the matter part of the Lagrangian $\sqrt{-g}L_{\rm matter}$ be decoupled from $\varphi(x)$ as an implementation of their requirement that the WEP be respected, in contrast to Jordan's model. The reason for this crucial decision, after the critical argument by Fierz [4] and others, will be made clear soon.

In this way they proposed the basic Lagrangian

$$\mathcal{L}_{\rm BD} = \sqrt{-g}\left(\varphi R - \omega\frac{1}{\varphi}g^{\mu\nu}\,\partial_\mu\varphi\,\partial_\nu\varphi + L_{\rm matter}(\Psi)\right). \tag{1.5}$$

We call the model described by (1.5) the *prototype BD model* throughout this book [5]. The adjective "prototype" emphasizes the unique features that characterize the original model compared with many extended versions. The dimensionless constant ω is the only parameter of the theory.

Note that we left out the factor 16π that multiplied the whole expression on the right-hand side in the original paper, to make the result appear in a more standard fashion. For this reason our φ is related to their original scalar field, denoted here by $\phi_{\rm BD}$, by the relation

$$\varphi = 16\pi\phi_{\rm BD}. \tag{1.6}$$

We now take a special look at the nonminimal coupling term, the first term on the right-hand side of (1.5). This replaces the Einstein–Hilbert term,

$$\mathcal{L}_{\rm EH} = \sqrt{-g}\frac{1}{16\pi G}R, \tag{1.7}$$

in the standard theory, in which R is multiplied by a constant G^{-1}. By comparing (1.5) and (1.7) we find that this model has no gravitational "constant," but is characterized by an *effective gravitational constant* $G_{\rm eff}$ defined by

$$\frac{1}{16\pi G_{\rm eff}} = \varphi, \tag{1.8}$$

as long as the dynamical field φ varies sufficiently slowly. In particular we may expect that φ is spatially uniform, but varies slowly with cosmic time, as suggested by Dirac. We should be careful, however, to distinguish $G_{\rm eff}$, the gravitational constant for the tensor force only, from the one including the possible contribution from the scalar field, as will be discussed later.

As another point, we note that the second term on the right-hand side of (1.5) appears to be a kinetic term of the scalar field φ, but looks slightly different. First, the presence of φ^{-1} seems to indicate a singularity. Secondly, there is a multiplying coefficient ω. These are, however, superficial differences. The whole term can be cast into the standard *canonical* form by redefining the scalar field.

For this purpose we introduce a new field ϕ and a new dimensionless constant ξ, chosen to be positive, by putting

$$\varphi = \tfrac{1}{2}\xi\phi^2 \tag{1.9}$$

and

$$\epsilon\xi^{-1} = 4\omega, \tag{1.10}$$

in terms of which the second term on the right-hand side of (1.5) is re-expressed in the desired form;

$$\sqrt{-g}\left(-\tfrac{1}{2}\epsilon g^{\mu\nu}\,\partial_\mu\phi\,\partial_\nu\phi\right), \tag{1.11}$$

with $\epsilon = \pm 1 = \mathrm{Sign}\,\omega$.

No singularity appears, as suggested. $\epsilon = +1$ corresponds to a normal field having a positive energy, in other words, not a *ghost*. Note that (1.11) becomes $\dot{\phi}^2/2$ for $\epsilon = +1$ in the limit of flat space-time where $g^{00} \sim \eta^{00} = -1$. The choice $\epsilon = -1$ seems to indicate a negative energy, which is unacceptable physically. As will be shown later in detail, however, this need not be an immediate difficulty owing to the presence of the nonminimal coupling. We will show in fact that some of the models do require $\epsilon = -1$. Even the extreme choice $\epsilon = 0$, corresponding to choosing $\omega = 0$ in the original formulation, according to (1.10), leaving ξ arbitrary, may not be excluded immediately. Note also that (1.9) shows that ϕ has a mass dimension 1, as in the usual formulation.

In this way (1.5) is cast into the new form

$$\mathcal{L}_{\rm BD} = \sqrt{-g}\left(\tfrac{1}{2}\xi\phi^2 R - \tfrac{1}{2}\epsilon g^{\mu\nu}\,\partial_\mu\phi\,\partial_\nu\phi + L_{\rm matter}\right). \tag{1.12}$$

Discussing consequences of (1.12) will be the main purpose of Chapter 2. We briefly outline here subjects of particular interest.

Obviously (1.12) describes something beyond what one would obtain simply by adding the kinetic term of the scalar field to the Lagrangian with the Einstein–Hilbert term. We reserve the term "scalar–tensor theory" specifically for a class of theories featuring a nonminimal coupling term or its certain extension.

As we explain in the subsequent section, there are theoretical models to be categorized in this class. More general models have also been discussed. The prototype BD model still deserves detailed study, from which we may learn many lessons useful in analyzing other models.

Deriving field equations from (1.12) is a somewhat nontrivial task, as we elaborate in Chapter 2. In particular, we obtain

$$\Box\varphi = \zeta^2 T, \tag{1.13}$$

where T is the trace of the matter energy–momentum tensor $T_{\mu\nu}$, while ζ is a constant defined by

$$\zeta^{-2} = 6 + \epsilon\xi^{-1} = 6 + 4\omega. \tag{1.14}$$

Notice that φ in (1.13) is the BD scalar field, now given by the combination of ϕ as given by (1.9), though the field equation itself was derived by considering ϕ as an independent field. The fact that the right-hand side of (1.13) is given in terms of the matter energy–momentum tensor guarantees that the force mediated by the scalar field respects the WEP, or universal free-fall (UFF). This is because, in the limit of zero momentum transferred, the source of the scalar field is given by the integrated T_{00}, which is the total energy of the system independent of what the content is.

One might be puzzled to see how the scalar field decoupled from the matter at the level of the Lagrangian comes to have a coupling at the level of the field equations. The underlying mechanism is provided by the nonminimal coupling term which acts as a mixing interaction between the scalar field and the spinless component of the tensor field, as will be elaborated toward the end of Chapter 2.

From (1.13) we expect that the scalar field mediates a long-range force between massive objects in the same way as the Newtonian force does in the weak-field limit of Einstein's gravity. The coupling strength is essentially of the same order of magnitude as that of the Newtonian force as long as ξ or ω is roughly of the order of unity, as we can see by restoring $8\pi G$ in the conventional unit system. Equation (1.14) also shows that the coupling vanishes as $\omega \to \infty$, or $\xi \to 0$. It is often stated that the theory reduces to Einstein's theory in this limit.

According to (1.13) the scalar field does not couple to the photon, for example, indicating that the light-deflection phenomenon will remain unaffected by the scalar force. This is an example displaying how the scalar force makes a difference from general relativity. On re-examining what had been done in general relativity, Brans and Dicke discovered room for the scalar component to be accommodated within the limit $\omega \gtrsim 6$ or $|\xi| \lesssim 0.042$ [5].

Shapiro time-delay measurements during the Viking Project in the 1970s, however, yielded the constraint $\omega \gtrsim 1000$, or $\xi \lesssim 2.5 \times 10^{-4}$ [6]. Two decades later, the latest bound from the VLBI experiments, basically concerning the light-deflection phenomenon involving light from extragalactic radio sources, is even stronger [7]:

$$\omega \gtrsim 3.6 \times 10^3, \quad \text{or} \quad \xi \lesssim 7.0 \times 10^{-5}, \tag{1.15}$$

severer than that expected initially by nearly three orders of magnitude. Notice that $\epsilon = +1$ is also implied. There has been no unambiguous evidence for the presence of the additional scalar field, only certain bounds of the parameter having been obtained.

It seems as if Brans and Dicke wished naturally to find positive evidence right at the very beginning. At one time, Dicke, as an experimentalist, wondered whether there was a flaw in comparing theory and observation. He specifically suspected that the Sun is not completely spherically symmetric, which property was used extensively to derive the Schwarzschild solution. He performed his own experiment to re-measure the Sun's oblateness. If the quadrupole moment J_2 turned out to be as large as $\sim 10^{-5}$, as he and Goldenberg reported [8], it would have allowed more deviation from general relativity, resulting in $\omega \sim 5$ or ~ 0.2 for $\epsilon = 1$ or -1, respectively.

Unfortunately, however, subsequent re-measurements by other groups yielded values mostly as small as $\sim 10^{-6}$ for J_2, including the latest even smaller value [7]. In this sense, there seems to be little hope that ω or ξ is close to anywhere around unity. The smallness of ξ has affected considerably the development of the theory during the years that followed. It appeared as if the theory were destined to grow only to occupy an ever smaller territory without an obvious reason, although Dicke himself worked actively on initiating a new era of "experimental relativity." In some sense, the scalar–tensor theory served as an explicit model illustrating what the world could be like if Einstein were not entirely right.

In spite of all these circumstances surrounding the scalar–tensor theory, however, there have been some people who were deeply impressed by the idea that nature's simplest phenomenon, a scalar field, plays a major role, and tried to modify the prototype BD model in such a way that the constraint (1.15) could be evaded. V. Wagoner suggested extending the original model by introducing arbitrary functions of the scalar field, including a mass term as well [9], though without well-defined physical principles to determine those functions.

One of the present authors (Y. F.) proposed that a dilaton, a Nambu–Goldstone (NG) boson of broken scale invariance, might mediate a finite-range gravity (non-Newtonian force) based on an idea in particle physics [10]. O'Hanlon [11], and Acharia and Hogan [12] showed immediately that the dilaton can be identified as the scalar field of a version of the prototype BD model, hence finding that the massive scalar field does have a place in the theory of gravity.

A crucial point in these approaches is that the scalar field is naturally not immune against acquiring a nonzero mass. If the corresponding force-range of the scalar force turns out to be smaller than the size of the solar system, it no longer affects the perihelion advance of Mercury, for example, thus leaving a constraint like (1.15) irrelevant. This will free us from a long-standing curse.

More recently, T. Damour and A. Polyakov showed that extending the prototype BD model in a way allowing the scalar field to enter in a more complicated manner is rather natural from the viewpoint of string theory [13]. They specifically proposed the "least-coupling principle" (LCP) according to which one might be able to understand why the deviation from general relativity is so small if there is any, though the idea is still short of being implemented from a realistic point of view.

In this book we will be interested also in the cosmological constant, which seems to be one of the hot topics at the present time [14, 15]. Although this subject has a long but widely known history, what we face today is quite new, and appears to be a challenge that probably requires something beyond the standard theory. Today might be a time when the discovery of an accelerating universe [14, 15] is in fact a crisis on which physics will thrive [16]. We may more specifically expect this issue to be a fresh ground to which the scalar–tensor theory applies. The situation might provide a chance to go beyond an "alternative theory," suggesting phenomena that had never been thought of in general relativity.

Ideas based on scalar fields have already been attempted, particularly under the name of "quintessence" [17]. Some of these theories are not necessarily related to the scalar–tensor theory. One has more flexibility, but to some extent they are more phenomenological. After giving a brief overview of the recent developments on these subjects, we will see how we reach an understanding of the accelerating universe in terms of the scalar–tensor theory, which has been extended minimally from the prototype BD model, from our point of view. As a further attempt, we apply the theory also to the reported time-dependence of α, the fine-structure constant [18].

As one of the conclusions that has emerged from our efforts, our classical solutions of the cosmological equations are partially chaotic, with very sensitive dependence on the initial values. Closely connected with this is the great likelihood that we are in a transient state before reaching a final, asymptotic state. All these things may alter our traditional view that the universe we see at the present time should have an attractor solution that depends on the initial values supposedly as little as possible. The universe after all might also be like many natural phenomena around us.

1.2 Where does the scalar field come from?

As we stated before, Brans and Dicke assumed that decoupling of the scalar field from the matter part of the Lagrangian occurs. As we see in the following, this is an assumption that hardly seems to be supported by any of the examples of more fundamental theories. They never made it clear how they could avoid this. It appears as if they had never been particularly concerned about whether there were any theories at a deeper level behind their model, which they viewed as an alternative theory in its own right.

It is nevertheless hard to deny that the scalar–tensor theory has attracted wide interest because it appears to provide a small window through which one can look into phenomenological aspects of more fundamental theories to which one is still denied any direct access otherwise.

It is truly remarkable and even surprising to find that a candidate scalar field of the desired nature is provided by the string theory of the late twentieth century, not to mention the KK theory of the 1920s. We will discuss briefly such candidates, starting from a reasonably detailed account of the KK approach. We then move on to the "dilaton" expected from string theory, and further to the recent development of the "brane," which has become a focus of attention even though it is still highly speculative. We also sketch another highly hypothetical idea that is closely related to "noncommutative geometry," which turns out to be yet another supplier of a scalar field.

1.2.1 The scalar field arising from the size of compactified internal space

Shortly after the advent of general relativity, the historic attempts at unification appeared, first due to H. Weyl [19], and then due to Kaluza. Weyl's theory eventually laid the foundation for what was later called gauge theory, the heart of the contemporary version of unification theories, whereas Kaluza's proposal, later known as the KK theory, played a decisive role in making clear the importance of higher-dimensional space-time in string theory, not to mention serving as an ancestor of the scalar–tensor theory.

Kaluza envisioned five-dimensional space-time to which general relativity was applied. One of the spatial dimensions was assumed to be "compactified" to a small circle leaving four-dimensional space-time extended infinitely as we see it. The size of the circle is so small that no phenomena of sufficiently low energies can detect it.

He started with the metric in five dimensions, of which the "off-diagonal" components connecting the four dimensions with the fifth dimension behave as a 4-vector that has been shown to play the role of the electromagnetic potential. In this way the theory offered the unified Einstein–Maxwell theory. The gauge transformation for the potential is interpreted as an isometry transformation along the circle.

The idea was re-discovered later from a more contemporary point of view [20], in particular in connection with the realization that string theory requires higher-dimensional space-time [21]. We outline briefly how the size of compactified internal space behaves as a four-dimensional scalar field precisely of the nature of the prototype BD model, with the parameter determined uniquely in terms of the dimensionality of space-time. See Appendix A for more details of derivations.

Let us assume the "*Ansatz*" for the $D = (4+n)$-dimensional metric with n-dimensional compactified space:

$$g_{\bar{\mu}\bar{\nu}} = \begin{pmatrix} g_{\mu\nu}(x) & 0 \\ 0 & A(x)^2 \tilde{g}_{\alpha\beta}(\theta) \end{pmatrix}, \tag{1.16}$$

with the radius A, while $\tilde{g}_{\alpha\beta}(\theta)$ means the purely geometrical portion described by the coordinates θ_α with $\alpha, \beta = 1, 2, \ldots, n$. We choose θ_α to be dimensionless, like angles. Notice that we omitted, for the moment, the off-diagonal components for the gauge fields, focusing on the scalar field.

We also have

$$\sqrt{-^{(D)}g} = \sqrt{-g}A^n\sqrt{\tilde{g}}, \tag{1.17}$$

where g is the four-dimensional determinant, while $\sqrt{\tilde{g}}$ is related to the volume V_n of compactified space by

$$V_n = A^n \tilde{\mathcal{V}}_n, \tag{1.18}$$

where

$$\tilde{\mathcal{V}}_n = \int \sqrt{\tilde{g}}\, d^n\theta. \tag{1.19}$$