General Relativity and Cosmology: Unsolved Questions and Future Directions

Ivan Debono $^{1,*,†}$ and George F. Smoot $^{1,2,3,*}$

$^1$ Paris Centre for Cosmological Physics, APC, AstroParticule et Cosmologie, Université Paris Diderot, CNRS/IN2P3, CEA/Irfu, Observatoire de Paris, Sorbonne Paris Cité, 10, rue Alice Domon et Léonie Duquet, 75205 Paris CEDEX 13, France; gfsmoot@lbl.gov

$^2$ Physics Department and Lawrence Berkeley National Laboratory, University of California, Berkeley, 94720 CA, USA

$^3$ Helmut and Anna Pao Sohmen Professor-at-Large, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, 999077 Hong Kong, China

* Correspondence: ivan.debono@apc.univ-paris7.fr; Tel.: +33-1-57276991

† These authors contributed equally to this work.

Academic Editors: Lorenzo Iorio and Elias C. Vagenas

Received: 21 August 2016; Accepted: 14 September 2016; Published: 28 September 2016

Abstract: For the last 100 years, General Relativity (GR) has taken over the gravitational theory mantle held by Newtonian Gravity for the previous 200 years. This article reviews the status of GR in terms of its self-consistency, completeness, and the evidence provided by observations, which have allowed GR to remain the champion of gravitational theories against several other classes of competing theories. We pay particular attention to the role of GR and gravity in cosmology, one of the areas in which one gravity dominates and new phenomena and effects challenge the orthodoxy. We also review other areas where there are likely conflicts pointing to the need to replace or revise GR to represent correctly observations and consistent theoretical framework. Observations have long been key both to the theoretical liveliness and viability of GR. We conclude with a discussion of the likely developments over the next 100 years.

Keywords: General Relativity; gravitation; cosmology; Concordance Model; dark energy; dark matter; inflation; large-scale structure

1. Perspective

Scientists have been fascinated by General Relativity ever since it was developed. It has been described as poetic, beautiful, elegant, and, at times, as impossible to understand.

General Relativity is often described as a simple theory. It is hard to define simplicity in science. One can always construct an entire theory encapsulated in one equation. Richard Feynman famously demonstrated this in a thought experiment where he rewrote all the laws of physics as $\vec{U} = 0$, where each element of $\vec{U}$ contained the hidden structure [1]. His point was that simplicity does not automatically bring truth.

An examination of the mathematical structure of General Relativity gives us a more sober definition of “simplicity”. Under certain assumptions about the structure of physical theories, and of the properties of the gravitational field, General Relativity is the only theory that describes gravity. Alternative theories introduce additional interactions and fields.

General Relativity is also unique among theories of fundamental interactions in the Standard Model. Like electromagnetism, but unlike the strong and weak interactions, its domain of validity covers the entire range of length scales from zero to infinity. However, unlike the other forces, gravity as described by General Relativity acts on all particles. This implies that the theory does not fail
below the Planck scale. All gravitational phenomena, from infinitesimal scales to distances beyond the observable universe, may be modelled by General Relativity. We may therefore formulate a mathematically rigorous description of General Relativity: it is the most complete theory of gravity ever developed.

All gravitational phenomena that have ever been observed can be modelled by General Relativity. It describes everything from falling apples, to the orbit of planets, the bending of light, the dynamics of galaxy clusters, and even black holes and gravitational waves. The domain of validity of the theory covers a wide range of energy levels and scales. That is why it has survived so long, and that is why it survives today, one hundred years after it was formulated, in an age in which the amount of data and knowledge increases by orders of magnitude every few years.

Figure 1. How we observe the universe. The lookback time is the difference between the age of the universe now, and the age of the universe when photons from an object were emitted. The more distant an object, the farther in its past we are observing its light. This distance in both space and time is expressed by the cosmological redshift \( z \). We obtain most of our astrophysical information from the surface of our past light cone, because it is carried by photons. The only information from within the cone come from local experiments and observations, such as geological records. The green dotted line is the world-line of the atoms and nuclei providing the material for our geological data. Local experiments are carried out along this bundle of world-lines. They provide a useful test of physical constants. One example is the observation of the Oklo phenomenon [2]. The earliest information we have collected so far comes from the cosmic microwave background (CMB). Earlier than the CMB time-like slice is the cosmic neutrino background. We observe Big Bang nucleosynthesis (BBN) indirectly, through observations of the abundances of chemical elements.

Why, then, are we still testing General Relativity? Why do we still develop, discuss, test and fine-tune alternative theories? Because there are some very fundamental open questions in physics, particularly in cosmology. Moreover, the big questions in cosmology happen to be the ones that are not answered by General Relativity: the accelerated expansion of the universe, the presence of a mysterious form of matter which cannot be observed directly, and the initial conditions in the early universe.
The theoretical completeness described above is both a necessary and aesthetic feature of a fundamental theory. However, it creates experimental difficulties, for it compels us to test the theory at extreme scales, where experimental errors may be large enough to allow several alternative theories.

At extremely small scales below the Planck length, classical mechanics should break down. This compels us to question whether General Relativity is still accurate at these scales, whether it needs to be modified, and whether a quantum description of gravity can be formulated. At the other end of the scale, at cosmological distances, we may question whether General Relativity is valid, given that the universe cannot be modelled sufficiently accurately by General Relativity without invoking either a cosmological constant, or some additional, unknown component of the universe. Finally, we may question the accuracy of our solutions to the equations of General Relativity, which depend on some approximation scheme. These approximations provided analytical solutions which enabled most of the early progress in General Relativistic cosmology and astrophysics. However, one century after the formulation of the theory, we now have a flood of data from increasingly accurate observations (as shown in Figure 1), coupled with computing power which was hitherto unheard of. Tests of the higher-order effects predicted by General Relativity and some of its competitors are now within reach.

The purpose of this review is to examine the motivation for the development of alternative theories throughout the history of GR, to give an overview of the state of the art in General Relativistic cosmology, and to look ahead. In the next few decades, some of the open questions in cosmology may well be answered by a new generation of experiments, and GR may be challenged by alternative theories.

2. A Brief History

Let us start this review by breaking our own rule about unscientific adjectives. General Relativity is a beautiful theory of gravity. It has not only thrilled us, but has survived 100 years of challenges, both by experimental tests and by alternative theories. The beauty of the theory was clear at the beginning, but the initial focus was on whether it was right. When General Relativity provided an explanation for the 43 seconds of arc per century discrepancy in the advance of the perihelion of Mercury [3], it got the attention of the scientific community. However, it was the prediction and the observation of the bending of light by the Sun [4] that confirmed GR's place as the new reigning theory of gravity [5].

The setting at the Royal Society under the portrait of Newton for the report of the eclipse light bending observations led by Arthur Eddington, and reported by the great writer Aldous Huxley, was perfect to describe to the world the ascendancy of a new theory replacing Newton's gravity (see, e.g., [6]). From this point onwards, the scientific community started to take General Relativity seriously, and theorists worked hard to understand this new theory, beguilingly simple but hard to apply, and to advance its predictions.

Shortly after its publication, GR quickly became the framework for astrophysics, and for the Standard or Concordance Model of cosmology. However, it was still challenged by alternative theories. Initially, the alternatives were motivated by theoretical considerations. This early period led to a fuller understanding of GR and its predictions. Some of the predictions, such as black holes and gravitational waves, divided the scientific community. Did they exist as physical objects, or just as mathematical artifacts of the theory?

By the time GR turned 50, the model of cosmology had been established, GR had been tested, and things had started to stagnate. However, advances in observations led to new discoveries, which in turn led to renewed challenges.

First came the missing mass in the universe. Could GR be modified to account for it? Then came the theories about the very early universe, and the behaviour of the quantum-scale, tiny initial universe. Finally, twenty years ago, came the confirmation of cosmic acceleration. This had a twofold effect. On one hand, it spurred the development of a whole range of alternative theories of gravity. On the other hand, it confirmed GR like never before, for General Relativity, with a cosmological constant, can account for the observations perfectly.
In 2015, on the 100th birthday of General Relativity, gravitational waves were observed for the first time. This had been the last major untested prediction of General Relativity. It was a remarkable achievement, and in many ways it heralds a new age of astrophysical observations. The experimental capabilities and the computing power have finally caught up with the theory. Cosmology and astrophysics have now entered the era of Big Data, and much of the theoretical effort is now driven by data. However, the foundation for almost the entire scientific endeavour is still this theory of chronogeometrodynamics, developed 100 years ago when today’s instruments and computers were still a distant dream.

From Aristotle to Einstein

General Relativity is the basis for the Standard Model of physical cosmology, and here we shall discuss the development of General Relativity (GR). The history of cosmology and GR are intertwined. We shall discuss why the theory has been so successful, and the criteria that must be satisfied by any alternative theory of physics, and by cosmological models.

Cosmology, in its broadest definition, is the study of the cosmos. It aims to provide an accurate description of the universe. Throughout much of the history of science, the development of cosmology was hampered by the lack of a universal physical theory. Observational tools were extremely limited, and there was no mathematical formulation for physical laws. The cosmos was described in metaphysical, rather than physical terms.

Discussions on the history of physics often refer to Karl Popper’s concept of ‘Falsifizierbarkeit’ (falsifiability) [7]. In this formalism, scientific discovery proceeds by successive falsifications of theories. A falsifiable theory that covers observations, and that has not yet been proven false can be regarded as provisionally acceptable. Yet we know that in reality it is not quite as straightforward. A theory that is considered to be correct acquires this status by accumulation of evidence rather than by a single falsification of a previous theory [8]. This is especially true in cosmology, where the selection of theoretical models often depends on the outcome of statistical calculations.

The scientific revolution which brought about the development of a precise mathematical language for physical theories heralded the scientific age of cosmology. Physical laws, tested here on Earth and later in the Solar System, could be applied to the ‘entire universe’, and could thus provide a precise physical description of the cosmos. Modern cosmology is based upon this epistemological framework. Cosmology depends upon a fundamental premise. As a science, it must deal strictly with what can be observed, but the observable universe forms only a fraction of the whole cosmos. One is forced to make the fundamental but unverifiable assumption that the portion of the universe which can be observed is representative of the whole, and that the laws of physics are the same throughout the whole universe [9]. Once we make this assumption, we can construct a model of the universe based on a description of its observable part.

Any cosmological model which assumes the universality of physical laws must be based upon some physical theory. Since cosmology aims to describe the universe on the largest possible scales, it must be based upon an application long-range physical interactions. Since the theory of gravitation is the physical theory at the basis of standard cosmology, and is also at the centre of the big questions facing modern cosmology, we shall give an overview of the development of theories of gravitation.

The development of physical theories of gravity was far from smooth, nor did it always conform to Popper’s scheme. Before the logical tools (mathematics) for the phenomenological description (physics) were invented, progress was rather haphazard.

According to Popper’s scheme, this development should be driven by the search for ever more general principles. Yet Aristotelian theory, to take one example, considered itself to be general enough—its claimed region of validity was the entire universe, except that rising smoke, floating feathers, falling apples and orbiting celestial spheres each had their own rules.
The real revolution came when it was realised that the behaviour of all bodies could be described by a single rule—a universal theory of gravitation. This theory is a description of the long range forces that electrically neutral bodies exert on one another because of their matter content.

Whether they choose to or not, scientists will always stand on the shoulders of giants. No theory is invented in a scientific vacuum. This goes all the way back to the cosmology of the Euro-Mediterranean Ancient World, codified in the Aristotelian teachings of the 4th century B.C. This Hellenic “natural philosophy” provided qualitative rather than quantitative descriptions for what we would call today the free parameters of the theory [10]. It stands to reason—the instruments had not yet been invented that could test the theory of gravity to within numerical accuracy. Without accurate timekeeping instruments, processes could at best be described as “slower than” or “faster than”. However, instruments to measure the movement of the celestial bodies, such as sundials, quadrants and astrolabes, were invented and improved upon, and measurements were carried out [11]. Astronomy flourished.

There is a certain logic to the development of physical theories from the Ancient World, to the Middle Ages, and right up to the Renaissance [12–14]. The basic tenet of the physics of Aristotle is that actions follow logically from causes. He distinguished between natural and violent motion. Natural motion implies falling at a speed proportional to the weight of the object and inversely proportional to the density of the medium. Violent motion happens whenever there is a force acting on an object, and the speed of the object is proportional to this force. Strato of Lampsacus replaced Aristotle’s explanation of ‘unnatural’ motion with one that is very close to the modern notion of inertia. He identified natural motion as a form of acceleration, and demonstrated experimentally that falling bodies accelerate. In the 14th century, Jean Buridan came up with the notion of impetus, where the initial force imparts motion to the object, which gradually diminishes as gravity and air resistance act against this initial force. Concurrently, Nicole Oresme was using a crude early form of graph to describe motion, and unwittingly showing the complicated notions of differentiation and integration in pictorial form[15,16].

The cosmological observations, limited to the innermost five planets of the Solar System (Mercury, Venus, Mars, Jupiter, and Saturn) and the sphere of stars, seemed to confirm the Aristotelian-Ptolemaic theory. Celestial bodies moved in regular patterns made up of repeating circles. Small discrepancies were explained by circles within circles.

The fact that the theories were based on these regular patterns is no accident. Patterns are the keyword in all of physics. Human beings are wired to recognise patterns. We can only build theories because we recognise patterns in the universe. This characteristic of valid theories has been called sloppiness. The patterns fall within some hyper-ribbon of stability in the theory [17].

The revolution in physics came with the development of mathematical, quantitative, models to describe physical reality. Starting in the 1580 Galileo carried out a series of observations in which he subjected kinematics to rigorous experiment, and showed that naturally-falling objects really do accelerate. Crucially, he showed that the composition of the body has no effect whatsoever on this acceleration. He also realised that for violent motion, the speed is constant in the absence of friction. Galileo also took rigorous observations of astronomical objects. In 1610 he made the first observation of Jupiter’s satellites, and the first observation of the phases of Venus, which is impossible according to the Ptolemaic geocentric model. His observations were important in putting to rest the Aristotelian theory of perfect and unchanging heavens.

By the time Newton came along, telescopes had been invented. Galileo had observed moons orbiting the Solar System planets, and hundreds of stars invisible to the naked eye. His 1610 treatise, aptly called Siderus Nuncius (“Starry Message”, or “Astronomical Report” in modern language) [18], was the first scientific work based on observations through a telescope. Mechanical clocks had been invented. The sphere of observed data had expanded [19]. Calculus provided the tool to make sense of this new flood of data. Thus, physicists of Newton’s generation found a very different scientific environment than the one in which Galileo had started off.
In 1687, Isaac Newton published in his “Mathematical Principles of Natural Philosophy”, known by its abbreviated Latin title as *Principia* [20]. This was a significant milestone in physics. Newton’s model of gravitation was, in his own words, a “universal” law. It applied to all bodies in the universe, whether it was cannonballs on Earth, or planets orbiting the Sun. For more than two centuries, Newton’s theory, was the standard physical description of gravity. There was no other attempt to find a different theory for the gravitational force, although the intervening years between Newtonian gravity and Relativity produced some important physical concepts such as de Maupertuis’s “Principle of Least Action” [21], further developed by Euler [22], Lagrange [23] and Hamilton [24,25]. The path of each particle is assigned a number called an action, which is the integral of the Lagrangian. In classical mechanics, the action principle is equivalent to Newton’s Laws. Lagrangian field theory is an important cornerstone of modern physics. The Lagrangian of any physical interaction, when subjected to an action principle, give us field equations and conservation laws for the theory. It is an expression of the symmetries in physical laws.

Newtonian gravity was the great success story of nineteenth century physics, the golden age of mathematical astronomy. It allowed astronomers to calculate the position of planets and asteroids with ever greater precision, and to confirm their calculations by observation. Thus the size of the known universe grew. Evidence started to accumulate suggesting that there might be other galaxies in the universe besides our own. In 1845, the planet Neptune was discovered, after Urbain le Verrier suggested pointing telescopes in a region of the Solar System which he predicted by Newtonian calculations [26,27]. The search was motivated in the first place by an anomaly in the orbit of Uranus which could not be otherwise explained using Newtonian theory [28]. The discovery of Neptune showed that Newtonian theory was valid even in the very farthest limits of the Solar System.

There was another anomaly which could not be explained—the excessive perihelion precession of Mercury by 43 arcseconds per century, confirmed by le Verrier himself. Urbain le Verrier thus holds the distinction of being one of the few experimentalists to have confirmed Newton’s theory and then disproved it. Astronomers attempted to explain this perihelion anomaly using Newtonian mechanics, which led them to speculate on the existence of Vulcan, a hypothetical planet whose orbit was even closer to the Sun [29].

The first doubts on Newtonian theory began to take shape just at the time when theorists were examining the full implications of the theory for complex, multi-body dynamical systems such as the Solar System. In 1890, Henri Poincaré published his magnum opus on the three-body problem [30], a masterpiece of celestial mechanics. At the time, Poincaré was working on another open question in physics: the aether. This led him to formulate a theory which was very close to Special Relativity [31], but which did not quite fit with Maxwell’s electromagnetism [32], and was ultimately flawed.

By the end of the 19th century, the necessary mathematical tools were in place which would enable the development of Special and then General Relativity. There is an intimate connection between physics and the language of mathematics which is often overlooked. The former, especially in modern times, depends on the latter. Could Aristotle have developed General Relativity? No. Because he had not the mathematical language. Equations and mathematical formulations are relatively recent in the history of physics. Even Newton, for all his fame as a mathematical genius, never wrote the equation \( F = -\frac{GMm}{r^2} \). He wrote a series of statements implying this law in (Latin) words: “Gravitatem, quæ Planetam unumquemque respicit, ese reciproce ut quadratum distantiae locorum ab ipsius centro”, and so on. It is hard to imagine how human beings could manipulate tensors and solve the field equations of Relativity in anything but numbers and symbols. Theories and physics do not happen in a cultural and scientific vacuum. They are human creations, and they depend intimately on tools for the transmission and communication of human knowledge.

The physical theory of gravity—the laws that govern gravitational interactions—remained unchanged until Einstein’s time. In 1905, Einstein published his Theory of Special Relativity (SR) [33]. Soon after, he turned to the problem of including gravitation within four-dimensional spacetime [34–37].
Newton’s formulation of the gravitational laws is expressed by the equations:

\[ \frac{d^2x^i}{dt^2} = -\frac{\partial \Phi}{\partial x^i}, \quad (1) \]
\[ \Delta \Phi = 4\pi G\rho, \quad (2) \]

where \( \Phi \) is the gravitational potential, \( G \) is the universal gravitational constant, \( \rho \) is the mass density, and \( \Delta = \nabla^2 \) is the Laplace operator. These equations cannot be incorporated into Special Relativity as they stand. The equation of motion (1) for a particle is in three-dimensional form, so it must be modified into a four-dimensional vector equation for \( \frac{d^2x^\mu}{d\tau^2} \). Similarly, the field Equation (2) is not Lorentz-invariant, since the three-dimensional Laplacian operator instead of the four-dimensional d’Alembertian \( \square = \partial^\mu \partial_\mu \) means that the gravitational potential \( \Phi \) responds instantaneously to changes in the density \( \rho \) at arbitrarily large distances. The conclusion is that Newtonian gravitational fields propagate with infinite velocity. In other words, instantaneous action in Newtonian theory implies action at a distance when reconsidered in the light of Special Relativity. This violates one of the postulates of SR. How do we reconcile gravity and Special Relativity?

3. The Development of General Relativity

3.1. From Special to General Relativity

The simplest relativistic generalisation of Newtonian gravity is obtained by representing the gravitational field by a scalar \( \Phi \). Since matter is described in Relativity by the stress-energy tensor \( T_{\mu\nu} \), the only scalar with dimensions of mass density (which corresponds to \( \rho \)) is \( T_{\mu\mu} \). A consistent scalar relativistic theory of gravity would thus have the field equation

\[ \square \Phi = 4\pi G T_{\mu\mu}. \quad (3) \]

However, when the equation of motion from this theory are applied to a static, spherically symmetric field \( \Phi \), such as that of the sun, acting on an orbiting planet, they would result in a negative precession, or retardation of the perihelion. Experimental evidence since the time of Le Verrier and his observation of the orbit of Mercury [38] clearly shows that planets experience a prograde precession of the perihelion. Moreover, in the limit of a zero rest-mass particle, such as a photon, the equations of motion show that the particle experiences no geodesic deviation. The existence of an energy-momentum tensor due to an electromagnetic field would also be impossible, since \( (T_{\text{electromagnetic}})_{\mu\mu} = 0 \). The theory therefore allows neither gravitational redshift, nor deviation of light by matter, both of which are clearly observable phenomena [39]. Another route to generalisation could be to represent the gravitational field by a vector field \( \Phi_{\mu} \), analogous to electromagnetism. Following through with this strategy, the “Coulomb” law in this theory gives a repulsion between two massive particles, which clearly contradicts observations. The theory also predicts that gravitational waves should carry negative energy, and, like the scalar theory, predicts no deviation of light. Like the scalar theory, then, the vector theory must be discarded.

What about a flat-space tensor theory? The gravitational field in this theory is described by a symmetric tensor \( h_{\mu\nu} = h_{\nu\mu} \). The choice of the Lagrangian in this theory is dictated by the requirement that \( h_{\mu\nu} \) be a Lorentz-covariant, massless, spin-two field.

In the 1930s, Wolfgang Pauli and Markus Fierz [40] were the first to write down this Lagrangian and investigate the resulting theory. The predictions of the theory for deviation of light agree with those of General Relativity, and are consistent with observations. Since the field equations and gauge properties are identical to those of Einsein’s linearised theory, the predictions for the properties of gravitational waves, including positive energy, agree with those obtained using the linearised theory in General Relativity. However, the theory differs from General Relativity in its predicted value for the
perihelion precession, which is $\frac{5}{3}$ of that given by GR. This disagrees with the value obtained from observations of Mercury’s orbit.

The theory has an even worse deficiency. If two gravitating bodies (that is, not test particles) are considered, and the field equations are applied to them, then the theory predicts that gravitating bodies cannot be affected by gravity, since they all move along straight lines in a global Lorentz reference frame. This holds for bodies made of arbitrary stress-energy, and since all bodies gravitate, then one must conclude that no body can be accelerated by gravity, which is an obvious self-inconsistency in the theory.

The only way in which a consistent theory of gravity can be constructed within Special Relativity is to consider the geometry of spacetime as the gravitational field itself. In other words, all matter moves in an effective Riemann space of metric $g^{\mu\nu} \equiv \eta^{\mu\nu} + h^{\mu\nu}$, where $\eta^{\mu\nu}$ is the Minkowski metric. The requirement of consistency leads us to universal coupling, which implies the Equivalence Principle.

The existence of curved spacetime can be deduced from purely physical arguments. In 1911, before he had fully developed General Relativity, Einstein [34] showed that a photon must be affected by a gravitational field, using conservation of energy applied to Newtonian gravitation theory. Schild [41–43] showed by a simple thought experiment, formulated within Special Relativity, that a consistent theory of gravity cannot be constructed within this framework. His argument is based upon a gravitational redshift experiment carried out in the field of the Earth, using a global Lorentz frame tied to the Earth’s centre. Successive pulses of light rising to the same height should experience a redshift, and therefore the pulse rate at the top should be slower than that at the bottom. But light rays are drawn at 45 degrees in Minkowski spacetime diagrams, so that top and bottom time intervals are equal, which is impossible if redshift occurs. Hence the spacetime must be curved. One therefore concludes that in the presence of gravity, Special Relativity cannot be valid over any sufficiently extended region.

General Relativity may be understood as a generalisation of Special Relativity over extended regions. Since Special Relativity can comfortably be described using tensor calculus, it was only natural to extend the flat Minkowski spacetime of Special Relativity to the curved spacetime of General Relativity. This was a physical application of Riemannian geometry [44,45], which had been developed in the second half of the 19th century. The idea of tensor calculus on curved manifolds was already mathematically well-established. Einstein’s innovation lay in identifying the Einstein tensor, itself related to the Riemann curvature tensor, as the “gravitational field” in the theory.

Einstein had been working on the problem for some years, starting in 1907. He arrived at the final, correct form in 1915 [46,47]. He was well-aware of the significance of his publication, and he gave it the succinct title of “The Field Equations of Gravitation” (Feldgleichungen der Gravitation). The correct field equations for the theory contained in this publication served as the starting point or subsequent derivations.

3.2. The Formalism of General Relativity

General Relativity is based on two independent but mutually supporting postulates.

The first postulate is sometimes referred to collectively as the Einstein Equivalence Principle:

- **The Strong Equivalence Principle**: The laws of physics take the same form in a freely-falling reference frame as in Special Relativity
- **The Weak Equivalence Principle**: An observer in freefall should experience no gravitational field. That is to say, an observer cannot determine from a local experiment whether the his laboratory is being accelerated by a rocket of static at the surface of a gravitating body. Gravity is erased up to tidal forces, which are determined by the size of the laboratory and its distance to the centre of the gravitational attraction.

The Equivalence Principle allows us to construct the metric and the equation of motion by transforming from a freely-falling to an accelerating frame. It can be mathematically expressed by
the assuming that all matter fields are minimally coupled to a single metric tensor $g_{\mu\nu}$. The distance between two points in 4-dimensional spacetime, called events, is:

$$d^2 = c^2 d\tau^2 = g_{\mu\nu} dx^\mu dx^\nu.$$  \hfill (4)

Throughout the text, we follow the Einstein summation convention for repeated indices, so that $c_i x^i = \sum_{i=1}^{n} c_i x^i$ for $i = 1, \ldots, n$. Greek indices are used for space and time components, while Latin indices are spatial ones only. We use the following metric signature: $(-+++)$.

The metric defines lengths and times measured by laboratory rods and clocks. This metric implies that the action for any matter field $\psi$ is of the form

$$S_{\text{matter}}[\psi, g_{\mu\nu}],$$ \hfill (5)

which gives us three important results. First, it implies the universality of freefall. Second, it implies that all non-gravitational constants are spacetime independent. Third, it implies that the laws of physics are isotropic. This equation defines how matter behaves in a given curved geometry, how light rays propagate, how stars, planets and galaxies move, and gives us verifiable observational consequences.

The second postulate is related to the dynamics of the gravitational interaction. This is assumed to be governed by the Einstein-Hilbert action:

$$S_{\text{gravity}} = \frac{c^3}{16\pi G} \int d^4x \sqrt{-g} R,$$ \hfill (6)

where $g_{\mu\nu}$ is a massless spin-2 field called the Einstein metric. General Relativity identifies the Einstein metric with the physical metric, that is: $g_{\mu\nu} = g^*_{\mu\nu}$. This implements the Strong Equivalence Principle.

The Einstein-Hilbert action defines the dynamics of gravity itself. Relativity is thus a geometrical approach to fundamental interactions. These are realised though continuous classical fields which are inseparably connected to the geometrical structures of spacetime, such as the metric, affine connection, and curvature.

The General Relativistic equation of motion is simply parallel transport on curved spacetime. It is given by

$$\frac{d^2 x^\mu}{d\tau^2} + \Gamma^\mu_{\alpha\beta} \frac{dx^\alpha}{d\tau} \frac{dx^\beta}{d\tau} = 0,$$ \hfill (7)

where $x^\mu$ is some set of coordinates for a point in spacetime. $\Gamma^\mu_{\alpha\beta}$ are the components of the affine connection (or metric connection). The fundamental theorem of Riemannian geometry states that the affine connection can be expressed entirely in terms of the metric:

$$\Gamma^\alpha_{\lambda\nu} = \frac{1}{2} g^{\alpha\nu} (g_{\mu\nu,\lambda} + g_{\lambda\nu,\mu} - g_{\mu\lambda,\nu}),$$ \hfill (8)

where the comma denotes a derivative, i.e., $g_{\mu\nu,\lambda} = \frac{\partial g_{\mu\nu}}{\partial x^\lambda}$.

We need to construct invariant quantities in GR (quantities that are the same for all observers). To achieve this, we need to contract the covariant $A_\mu$ and contravariant $A^\mu$ components of a vector or tensor $A$ by using the metric to raise or lower indices: $A_\mu = g_{\mu\nu} A^\nu$. Thus the equation of motion (7) can be made covariant by recasting it as the covariant derivative of the 4-velocity $U^\mu = \gamma(c, v)$:

$$\frac{D_\mu U^\mu}{d\tau} = 0,$$ \hfill (9)

where the covariant derivative is defined as

$$D_\mu A^\mu = dA^\mu + \Gamma^\mu_{\alpha\beta} A^\alpha dx^\beta.$$ \hfill (10)
The quantity $\gamma$ is the Lorentz factor:

$$\gamma = \frac{1}{\sqrt{1 - v^2/c^2}}. \quad (11)$$

The transformation from SR to GR is then carried out by mapping the Minkowski metric to a general metric: $\eta \rightarrow g$ and by mapping $\partial \rightarrow D$.

In GR, freely-falling bodies travel along a geodesic. Geometrically, this is the shortest distance between two points in spacetime. The path length along a geodesic is given by

$$S = \left( \int g_{\mu\nu} \, dx^\mu \, dx^\nu \right)^{1/2}. \quad (12)$$

In cosmology, it is essential for us to be able to describe spacetime which is not "empty". In the presence of a perfect fluid (an inviscid fluid with density $\rho$ and isotropic pressure $p$), the energy and momentum of spacetime is described by the energy-momentum tensor (or stress-energy tensor)

$$T_{\mu\nu} = \left( \rho + \frac{P}{c^2} \right) U^\mu U^\nu - pg_{\mu\nu}. \quad (13)$$

Classical energy and momentum conservation are generalized in GR as the four conservation laws

$$D_\mu T^{\mu\nu} = 0. \quad (14)$$

In other words, the stress-energy tensor has a vanishing covariant divergence. In the absence of a component possessing pressure or density, or both, the energy-momentum tensor is zero.

The central notion in General Relativity is that gravitation can be described by a metric. The Einstein equations give us the relation between the metric and the matter and energy in the universe:

$$G_{\mu\nu} = -\frac{8\pi G}{c^4} T_{\mu\nu}. \quad (15)$$

The left-hand side of this equation is a function of the metric: $G_{\mu\nu}$ is the Einstein tensor, defined as:

$$G_{\mu\nu} = R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R, \quad (16)$$

where $R_{\mu\nu}$ is the Ricci tensor, which depends on the metric and its derivatives, and the Ricci scalar $R$ is the contraction of the Ricci tensor ($R = g_{\mu\nu} R^{\mu\nu}$). The right-hand side of Equation (15) is a function of the energy: $G$ is Newton’s constant, and $T_{\mu\nu}$ is the energy-momentum tensor.

Einstein’s Relativity has three main distinguishing characteristics:

- it agrees with experiment
- it describes gravity entirely in terms of geometry
- it is free of any “prior geometry”

These characteristics are lacking in most of the other theories [48,49]. Apart from the issue of agreement with experiment, Einstein’s theory is unique in its physical simplicity.

Every other theory introduces auxiliary gravitational fields, or involves prior geometry. Prior geometry is any aspect of the geometry of spacetime which is fixed immutably, that is, it cannot be changed by changing the distribution of gravitating sources.

A rigorous mathematical definition of the unique simplicity of General Relativity is given by Lovelock’s theorem [50–52]. This is a generalisation of an earlier theorem by Élie Cartan [53], and may be formulated as follows:

In 4 spacetime dimensions, the only divergence-free symmetric rank-2 tensor constructed solely from the metric $g$ and its derivatives up to second differential order, and preserving diffeomorphism invariance, is the Einstein tensor plus a cosmological term.
In simple terms, the theorem states that GR emerges as the unique theory of gravity if the conditions of the theorem are followed. In fact, Lovelock’s theorem provides a useful scheme for classifying alternatives to General Relativity.

Einstein described both the demand for “no prior geometry” and for a “geometric, coordinate-independent formulation of physics” by the single phrase “general covariance”, but the two concepts are not quite the same.

While many physical theories can be formulated in a generally covariant way, General Relativity is actually based on the “no prior geometry” demand. This distinction was not always made, especially in the first decades after Einstein’s publications [54,55]. Erich Kretschmann’s famous objection in 1917 [56] concerned this point, since he regarded general covariance merely as formal feature that any theory could have, not as a special feature belonging to GR.

3.3. Newtonian Nostalgia: The First Wave of Alternative Theories

Newton’s theories had predicted observations of Solar System objects, comets and asteroids, with astounding precision. Why should they be tampered with? The first wave of alternative theories were driven more by theoretical considerations than by observations. Equations (1) and (2) can be generalized so that they are consistent with the postulates of Special and General Relativity. Several generalisations of this kind were attempted in the first few decades following the development of GR, motivated by lingering resistance to any deviation from Newtonian gravity.

One early theory, involving prior geometry, was formulated by Nordstrøm in 1913 [57]. In this theory, the physical metric of spacetime $g$ is generated by a background flat spacetime metric $\eta$, and by a scalar gravitational field $\phi$. Stress-energy generates $\phi$:

$$\eta^{\alpha\beta} \phi_{,\alpha\beta} = -4\pi \eta^{\alpha\beta} T_{\alpha\beta}$$

and $g$ is constructed from $\phi$ and $\eta$:

$$g_{\alpha\beta} = \phi^2 \eta_{\alpha\beta}.$$ (18)

Prior geometry cannot be removed by rewriting Nordstrøm’s equations in a form devoid of $\eta$ and $\phi$ [58]. Mass only influences one degree of freedom in the spacetime geometry, while the other degrees of freedom are fixed a priori. This prior geometry, if it existed, could be detected by physical experiments.

In the 1920s, Alfred North Whitehead [59] formulated a two-tensor theory of gravity in which the prior geometry is quite different from later theories such as Nǐ’s [48]. Whitehead’s theory is remarkable in that it agrees with Einstein’s in its predictions for the four standard tests (bending of light, gravitational redshift, perihelion shift, and time delay). It was accepted as a viable alternative for Einstein’s theory until Clifford Martin Will [60] showed that it predicts velocity-independent anisotropies in the Cavendish constant (the gravitational constant $G$ in Newtonian theory). This would produce time-dependent Earth tides which are clearly contradicted by everyday observations. Any valid theory of gravity must not only agree with relativistic experiments, but also with past experiments in the Newtonian regime.

One theory which disagrees violently with non-relativistic experiments is due to George David Birkhoff [61]. It was developed in the 1940s, and it predicts the same redshift, perihelion shift, deflection and time-delay as General Relativity, but it requires that the pressure inside gravitating bodies should be equal to the total density of mass-energy ($p = \rho$). This means that sound waves travel with the speed of light. This clearly contradicts everyday experiments.

Most of the early alternative theories were abandoned either because they were contradicted by observations, or because of internal inconsistencies in the theories themselves. One notable exception is Dicke-Brans-Jordan theory, sometimes called Brans-Dicke, or Jordan-Fierz-Brans-Dicke theory [62,63], developed in the 1960s by Robert H. Dicke and Carl H. Brans following earlier work by Pascual Jordan.
and Markus Fierz. The different names arise from the fact that the theory is a special case of Jordan’s, with \( \eta = -1 \). An alternative mathematical representation of the theory is given by [64].

This theory introduced auxiliary gravitational fields. Brans and Dicke took the equivalence principle as the starting point of their theory, and thus they describe gravity in terms of spacetime curvature, but their gravitational field, unlike Einstein’s, is a scalar-tensor combination. In this way it overcomes the difficulties associated with tensor or scalar-only theories mentioned earlier. The trace of the energy-momentum tensor \((T_M)_\mu^\mu\) (representing matter) and a coupling constant \(\lambda\) generate the long-range scalar field \(\phi\) via the equation

\[
\Box^2 \phi = 4\pi \lambda (T_M)_\mu^\mu. \tag{19}
\]

The scalar field \(\phi\) fixes the value of \(G\), which is therefore not a constant, but simply the coupling strength of matter to gravity. The gravitational field equations relate the curvature to the energy-momentum tensors of the scalar field and matter:

\[
R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R = -\frac{8\pi}{c^4} \left[ (T_M)_{\mu\nu} + (T_\phi)_{\mu\nu} \right], \tag{20}
\]

where \((T_M)_{\mu\nu}\) is the energy-momentum tensor of matter and \((T_\phi)_{\mu\nu}\) is the energy-momentum tensor of the scalar field \(\phi\). For historical reasons, it is usual to write the coupling constant as

\[
\lambda = \frac{2}{3 + 2\omega}, \tag{21}
\]

where \(\omega\) is the dimensionless ‘Dicke coupling constant’. In the limit \(\omega \to \infty\), we have \(\lambda \to 0\), so \(\phi\) is not affected by the matter distribution, and can be set to a constant \(\phi = 1/G\). Hence Dicke-Brans-Jordan theory reduces to Einstein’s theory in the limit \(\omega \to \infty\).

The equivalence principle is satisfied in this theory since the special-relativistic laws are valid in the local Lorentz frames of the metric \(g\) of spacetime. The scalar field does not exert any direct influence on matter. It only enters the field equations that determine the geometry of spacetime. On a conceptual level, Brans-Dicke theory can be seen as more fully Machian than Einstein’s theory. Einstein himself attempted to incorporate Mach’s Principle into his theory, but in Einstein’s General Relativity, the inertial mass of an object will always be independent of the mass distribution in the universe. In Brans-Dicke theory, the long-range scalar field is an indirectly coupling field, so it does not directly influence matter, but the Einstein tensor is determined partly by the energy-momentum tensor, and partly by the long-range scalar field.

Dicke-Brans-Jordan theory is self-consistent and complete, but experimental evidence based on Solar System tests, shows that \(\omega \geq 600\) [65], as a conservative estimate. Some calculations raise this limit even higher, with \(\omega \geq 10^4\) [66]. The Cassini mission set a comparable limit of \(\omega > 40,000\) [67]. Recent cosmological data from the Planck probe show that \(\omega \geq 890\) [68,69]. This is consistent with the Solar System bounds. Future cosmological experiments and data from large-scale structure could provide even better constraints [70].

Brans-Dicke theory is a special case of general scalar-tensor theories with \(\omega(\phi) = \) constant, where \(\phi\) is a value depending on the cosmological epoch. In these theories, the function \(\omega(\phi)\) could be such that the theory is very different from GR in the early universe or in future epochs, but very close to GR in the present. In fact, it has been shown that GR is a natural attractor for such scalar-tensor theories, since cosmological evolution naturally drives the fields towards large values of \(\omega\) [71,72].

3.4. Self-Consistency, Completeness, and Agreement with Experiment

Any viable theory must satisfy three fundamental criteria: self-consistency, completeness, and agreement with past experiment.
To be self-consistent, a theory must not contain any internal contradictions. The spin-two field theory of gravity [40] is equivalent to linearised General Relativity but it is internally inconsistent since it predicts that gravitating bodies should have their motion unaffected by gravity. When one tries to remedy this inconsistency, the resulting theory is nothing but General Relativity. Another self-inconsistent theory is due to Paul Kustaaanheimo [73,74]. It predicts zero gravitational redshift when the wave version of light (Maxwell theory) is used, and nonzero redshift when the particle version (photon) is used.

To be complete, a theory must be able to analyse the outcome of any experiment. This means that it must be compatible with other physical theories which describe any other forces that are present in experiments. This can only be achieved if the theory is derived from first principles, since the theoretical postulates must be as general as possible if the theory is to cover the widest range of phenomena.

A viable theory must agree with past experiment, which includes experiments in the Newtonian regime, and the standard tests of General Relativity. Its results must agree with those obtained from Newtonian theory in the weak field limit, and with GR in relativistic situations. It also means that the theory must agree with cosmological observations.

The experimental criterion also works the other way. Any alternative to General Relativity that claims to have a smaller set of limiting cases must be experimentally distinguishable, perhaps by future experiment. At some point, the divergence between GR and other theories must manifest itself physically, in the form of predictions which can be verified by experiment. This is perhaps the greatest challenge of current alternatives to GR.

3.5. Metric Theories and Quantum Gravity

Most theories of gravity incorporate two principles: spacetime possesses a metric; and that metric satisfies the equivalence principle. Such theories are called metric theories. There are some exceptions.

Soon after the publication of the theory of General Relativity, it became apparent that its formulation is incompatible with a Quantum Mechanical description of the gravitational field. It was Einstein himself who pointed out that quantum effects must lead to a modification of General Relativity [75]. Back then, the first successful applications of Quantum Mechanics to electromagnetism were starting to give useful results. These developments led to the question of whether General Relativity can be quantized.

This is a difficult question. First, Einstein’s field equations are much more complicated than Maxwell’s equations, and in fact are nonlinear. The physical reason for this is that the gravitational field is coupled to itself—the stress-energy tensor acts as the source for spacetime curvature, which in general contributes to the stress-energy tensor. This means that the equations seem to violate the superposition principle, which requires the existence of a linear vector space (see, e.g., [76,77]). This is the mathematical expression of wave-particle duality—a central tenet of Quantum Theory.

Second, to quantize the gravitational field we would have to quantize spacetime itself. The physical meaning of this is not completely clear.

Finally, there are experimental problems. Maxwell’s equations predict electromagnetic radiation, which was first observed by Hertz [78]. Quantization of the field results in being able to observe individual photons, and these were first seen in the photoelectric effect predicted by Einstein [79]. Similarly, Einstein’s equations for the gravitational field predict gravitational radiation [75], so there should be, in principle, the possibility of observing individual gravitons, which are the quanta of the field. The direct observation of gravitational waves was finally achieved in September 2015 by the LIGO instrument [80]. The detection of individual gravitons is more difficult and is beyond the capability of current experiments.

To develop a quantum theory of General Relativity, the fundamental interactions in GR would have to follow quantum rules. In Quantum Theory, particle interactions are described by gauge theories, so GR would have to follow the gauge principle. Although the gauge principle was first
recognized in electromagnetism, modern gauge theory, formulated initially by Chen Ning Yang and Robert Mills [81,82], emerged entirely within the framework of the quantum field programme. As more particles were discovered after the 1940s, various possible couplings between those elementary particles were being proposed. It was therefore necessary to have some principle to choose a unique form out of the many possibilities suggested. The principle suggested by Yang and Mills in 1954 is based on the concept of gauge invariance, and is hence called the gauge principle.

3.6. The Gauge Approach and Non-Metric Theories

The idea of gauge invariance, and the term itself, originated earlier, from the following consideration due to Hermann Weyl in 1918 [83,84]. In addition to the requirement of General Relativity that coordinate systems have to be defined only locally, so likewise the standard of length, or scale, should only be defined locally. It is therefore necessary to set up a separate unit of length at every spacetime point. Weyl called such a system of unit-standards a gauge system (analogous to the standard width, or “gauge”, of a railway track).

The gauge principle therefore may be formulated as follows: If a physical system is invariant with respect to some global (spacetime independent) group of continuous transformations $G$, then it remains invariant when that group is considered locally (spacetime dependent), that is $G \rightarrow G(x)$. Partial derivatives are replaced by covariant ones, which depend on some new vector field.

In Weyl’s view, a gauge system is as necessary for describing physical events as a coordinate system. Since physical events are independent of our choice of descriptive framework, Weyl maintained that gauge invariance, just like general covariance, must be satisfied by any physical theory.

In Euclidean geometry, we know that translation of a vector preserves its length and direction. In Riemannian geometry, the Christoffel connection [85] (or affine connection) guarantees length preservation, but a vector’s orientation is path dependent. However, the angle between two vectors, following the same path, is preserved under translation. Weyl wondered why the remnant of planar geometry, length preservation, persisted in Riemannian geometry. After all, our measuring standards (rigid rods and clocks), are known only at one point in spacetime. To measure lengths at another point, we must carry our measuring tools along with us. Weyl maintained that only the relative lengths of any two vectors at the same point, and the angle between them, are preserved under parallel transport. The length of any single vector is arbitrary. To encode this mathematically, Weyl made the following substitution:

$$g_{\mu\nu}(x) \rightarrow \lambda(x)g_{\mu\nu}(x), \quad (22)$$

where the conformal factor $\lambda(x)$ is an arbitrary, positive, smooth function of position. Weyl required that in addition to GR’s coordinate invariance, formulae must remain invariant under the substitution (Equation (22)). He called this a gauge transformation. The scale therefore becomes a local property of the metric.

Weyl’s theory enabled him to unify gravity and electromagnetism, the only two forces known at the time. However, Weyl’s original scale invariance was abandoned soon after it was proposed, since its physical implications seemed to contradict experiments. In particular, if two identical clocks $C_1$ and $C_2$ are transported on two different paths, which both end at the same point $Q$, the time-like vectors $l_1$ and $l_2$ given by $C_1$ and $C_2$ at $Q$ would be different in the presence of an electromagnetic field. Therefore the two clock rates would differ. As Einstein (probably the only expert who could keep an eye on Weyl’s theory at the time) pointed out, this concept meant that spectral lines with definite frequencies could not exist, since the frequency of the spectral lines of atomic clocks would depend on the atom’s location, both past and present. However, we know the atomic spectral lines to be definite, and independent of spacetime position [86–89].

Despite its initial failures, Weyl’s idea of a local gauge symmetry survived, and acquired new meaning with the development of Quantum Mechanics. According to Quantum Mechanics, interactions are realized through quantum (that is, non-continuous) fields which underlie the local
coupling and propagation of field quanta, but which have nothing to do with the geometry of spacetime. The question is whether General Relativity can be formulated as a gauge theory. This question has been discussed by ever since it was first posed [90–96].

If features of General Relativity could be recovered from a gauging argument, then that would show that the two formulations are not inconsistent. The first to succeed in this was Kibble [91], who elaborated on an earlier, unsuccessful attempt by Utiyama [90]. Kibble arrived at a set of gravitational field equations, although not the Einstein equations, constructing a slightly more general theory, known as “spin-torsion” theory. The inclusion of torsion in Einstein’s General Relativity had long been theorized. In fact the necessary modifications to General Relativity were first suggested by Élie Cartan in the 1920s [97–100], who identified torsion as a possible physical field.

The connection between torsion and quantum spin was only made later [91,101,102], once it became clear that the stress-energy tensor for a massive fermion field must be asymmetric [103,104]. The Einstein-Cartan (1920s) and the Kibble-Sciama (late 1950s) developments occurred independently. For historical reasons, spin-torsion theories are sometimes referred to as Einstein-Cartan-Kibble-Sciama (ECKS) theories, but Einstein-Cartan Theory (ECT) is the term more commonly employed.

The Einstein-Cartan Theory of gravity is a modification of GR allowing spacetime to have torsion in addition to curvature, and, more importantly, relating torsion to the density of intrinsic angular momentum. This modification was put forward by Cartan before the discovery of quantum spin, so the physical motivation was anything but quantum theoretic. Cartan was influenced by the works of the Cosserat brothers [105] who considered a rotation stress tensor in a generalized continuous medium besides a force tensor.

Cartan assumed the linear connection to be metric and derived, from a variational principle, a set of gravitational field equations. However, Cartan required, without justification, that the covariant divergence of the energy-momentum tensor be zero, which led to algebraic constraint equations, thus severely restricting the geometry. This probably discouraged Cartan from pursuing his theory. It is now known that the conservation laws in relativistic theories of gravitation follow from the Bianchi identities and in the presence of torsion, the divergence of the energy-momentum tensor need not vanish.

In simple mathematical terms, a non-zero torsion tensor means that

$$\Gamma^\mu_{\nu\sigma} = \Gamma^\mu_{\nu\sigma} - \Gamma^\mu_{\sigma\nu} \neq 0.$$  \hspace{1cm} (23)

Geometrically, it means that an infinitesimal geodesic parallelogram forms a non-closed loop. Torsion is therefore a local property of the metric. The Lagrangian action of Einstein-Cartan theory takes the usual Einstein-Hilbert form:

$$S = \int d^4x \sqrt{-g} \left( -\frac{R_{\mu\nu}(\Gamma)}{16\pi G} + \mathcal{L}_m \right),$$  \hspace{1cm} (24)

where $\Gamma$ is a general affine connection and $\mathcal{L}_m$ is the matter Lagrangian. The theory differs from GR in the structure of $\Gamma$, leading to a field theory with additional interactions.

Torsion vanishes in the absence of spin and the Einstein-Cartan field equation is then the classical Einstein field equation. In particular, there is no difference between the Einstein and Einstein-Cartan theories in empty space. Since practically all tests of relativistic theory are based on free space experiments, the two theories are, to all effects, indistinguishable via the standard tests of GR. The inclusion of torsion only results in a slight change in the energy-momentum tensor. Cartan’s theory holds the distinction of being complete, self-consistent and in agreement with experiment, but of being a non-metric theory of gravitation. The link between torsion and quantum spin means that it could be possible to study the divergence between the GR and ECKS theories at the quantum level. Such experiments have recently been proposed [106].
Kibble’s theory contains some features which were criticized [107]. It is now accepted that torsion is an inevitable feature of a gauge theory based upon the Poincaré group. Classical GR must be modified by the introduction of a spin-torsion interaction if it is to be viewed as a gauge theory. The gauge principle alone fails to provide a conceptual framework for GR as a theory of gravity.

In the 1990s, Anthony Lasenby, Chris Doran and Stephen Gull proposed an alternative formulation of General Relativity which is derived from gauge principles alone [108–113]. Their treatment is very different from earlier ones where only infinitesimal translations are considered [91,107]. There are a few other theories similar in their approach to that of Lasenby, Doran and Gull (e.g., [114,115]).

4. Why Consider Alternative Theories?

The motivation for considering alternatives to GR comes mainly from theoretical arguments, like scale invariance of the gravitational theory, additional scalar fields that emerge from string theories, Dark Matter, dark energy or inflation, or additional degrees of freedom that arise in the framework of brane-world theories.

In Table 1, we draw up a list of some of the more well-known alternatives to General Relativity. This list is far from exhaustive, but it serves to highlight the major elements which differentiate these theories. There are several works containing a more detailed listing and discussion of the various alternative theories (e.g., [39,116,117]).

Table 1. A “comparative morphology” of some of the major alternatives to General Relativity, in approximate chronological order. We have only listed the theories of particular historical significance. The current landscape, in which cosmologists seek to explain Dark Matter, dark energy, and inflation, offers far more theories. It is generally easier to incorporate the non-gravitational laws of physics within metric theories, since other theories would result in greater complexity, rendering calculations difficult. The only way in which metric theories significantly differ from each other is in their laws for the generation of the metric. Abbreviations: Tensor (T), V (Vector), S (Scalar), P (Potential), Dy (Dynamic), Einstein Equivalence Principle (EEP), i.e., uniqueness of freefall, Local Lorentz Invariance (LLI), Local Position Invariance (LPI), param (Parameter), ft (Function).

<table>
<thead>
<tr>
<th>Theory</th>
<th>Metric</th>
<th>Other Fields</th>
<th>Free Elements</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Newton 1687 [20]</td>
<td>Nonmetric</td>
<td>P</td>
<td>None</td>
<td>Nonrelativistic, implicit action at a distance</td>
</tr>
<tr>
<td>Poincaré 1890s–1900s [31,118]</td>
<td>Fails; does not mesh with electromagnetism</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nordstrøm 1913 [57]</td>
<td>Minkowski</td>
<td>S</td>
<td>None</td>
<td>Fails to predict observed light detection</td>
</tr>
<tr>
<td>General Relativity 1915 [46]</td>
<td>Dy</td>
<td>None</td>
<td>None</td>
<td>Viable</td>
</tr>
<tr>
<td>Whitehead 1922 [59]</td>
<td></td>
<td></td>
<td></td>
<td>Violates LLI; contradiction by everyday observation of tides</td>
</tr>
<tr>
<td>Cartan 1922–1925 [98]</td>
<td>ST</td>
<td></td>
<td></td>
<td>Still viable; introduces matter spin</td>
</tr>
<tr>
<td>Birkhoff 1943 [61]</td>
<td>T</td>
<td></td>
<td></td>
<td>Fails Newtonian test; demands speed of sound equal to speed of light</td>
</tr>
<tr>
<td>Milne 1948 [121]</td>
<td>Machian</td>
<td></td>
<td></td>
<td>Incomplete; no gravitational redshift prediction; contradicts cosmological observations.</td>
</tr>
<tr>
<td>Thiry 1948 [122]</td>
<td>ST</td>
<td></td>
<td></td>
<td>Unlikely; extremely constrained by results on $\gamma_{\text{PN}}$</td>
</tr>
<tr>
<td>Belinfante-Swihart 1957 [123]</td>
<td>Nonmetric</td>
<td>T</td>
<td>K param</td>
<td>Violates EEP; contradicted by Dicke–Braginsky experiments</td>
</tr>
<tr>
<td>Brans-Dicke 1961 [63]</td>
<td>Generic S</td>
<td>Dy</td>
<td>S</td>
<td>Viable for $\omega &gt; 500$</td>
</tr>
<tr>
<td>Ni 1972 [48]</td>
<td>Minkowski</td>
<td>T, V, S</td>
<td>1 param, 3 ft (</td>
<td>Violates LPI; predicts preferred-frame effects</td>
</tr>
<tr>
<td>Will-Nordtvedt 1972 [124]</td>
<td>Dy T</td>
<td>V</td>
<td></td>
<td>Viable but can only be significant at high energy regimes</td>
</tr>
<tr>
<td>Barker 1978 [125]</td>
<td>ST</td>
<td></td>
<td></td>
<td>Unlikely; severely constrained</td>
</tr>
<tr>
<td>Rosen 1973 [126,127]</td>
<td>Fixed</td>
<td>T</td>
<td>None</td>
<td>Contradicted by binary pulsar data</td>
</tr>
<tr>
<td>Rastall 1976 [128]</td>
<td>Minkowski</td>
<td>S, V</td>
<td>None</td>
<td>Contradicted by gravitational wave data</td>
</tr>
<tr>
<td>$f(R)$ models 1970s [129,130]</td>
<td>M, 1ST</td>
<td>S</td>
<td>Free ft</td>
<td>Consistent with Solar System tests; viable but severely constrained</td>
</tr>
<tr>
<td>MOND 1983 [131–133]</td>
<td>Nonmetric</td>
<td>P</td>
<td>Free ft</td>
<td>Nonrelativistic theory</td>
</tr>
<tr>
<td>DGP 2000 [134]</td>
<td>ST/Quantum</td>
<td></td>
<td></td>
<td>Appears to be contradicted by BAOs, CMB and Supernovae Ia unless DE added</td>
</tr>
<tr>
<td>TeVeS 2004 [135]</td>
<td>T, V, S</td>
<td>Dy S</td>
<td>Free ft</td>
<td>Highly unstable [138], ruled out by SDSS data [137]</td>
</tr>
</tbody>
</table>
5. From General Relativity to Standard Cosmology

When Einstein published his seminal GR papers it became almost immediately apparent that the theory could be applied to the whole universe, under certain assumptions, to obtain a relativistic cosmological description. If the content of the universe is known, then the energy-momentum tensor can be constructed, and the metric derived using Einstein’s equations. Einstein himself was the first to apply GR to cosmology in 1917 [138]. The first expanding-universe solutions to the relativistic field equations, describing a universe with positive, zero and negative curvature, were discovered by Alexander Friedmann [139,140]. This occurred before Edwin Hubble’s observations and the empirical confirmation, in 1929, that the redshift of a galaxy is proportional to its distance. Hubble formulated the law which bears his name: \( v = H_0 r \), where \( H_0 \) is the constant of proportionality [141]. The problem of an expanding universe was independently followed up during the 1930s by Georges Lemaître [142], and by Howard P. Robertson [143–145] and Arthur Geoffrey Walker [146].

These exact solutions define what came to be known as the Friedmann-Lemaître-Robertson-Walker (FLRW) metric, also referred to as the FRW, RW, or FL metric. This metric starts with the assumption of spatial homogeneity and isotropy, allowing for time-dependence of the spatial component of the metric. Indeed, it is the only metric which can exist on homogeneous and isotropic spacetime. The assumption of homogeneity and isotropy, known as the Cosmological Principle, follows from the Copernican Principle, which states that we are not privileged observers in the universe. This is no longer true below a certain observational scale of around 100 Mpc (sometimes called the “End of Greatness”), but it does simplify the description of the distribution of mass in the universe.

The FLRW metric describes a homogeneous, isotropic universe, with matter and energy uniformly distributed as a perfect fluid. Using the definition of the metric in Equation (4), it is written as:

\[
- \, ds^2 = c \, dt^2 - R^2(t) \left[ dr^2 + S_0^2(r) \left( d\theta^2 + \sin^2 \theta \, d\phi^2 \right) \right],
\]

where \( r \) is a time independent comoving distance, \( \theta \) and \( \phi \) are the transverse polar coordinates, and \( t \) is the cosmic or physical time. \( R(t) \) is the scale factor of the universe. The function \( S_0^2(r) \) is defined as:

\[
S_0^2(r) = \begin{cases} 
\sin(r) & (k = +1) \\
r & (k = 0) \\
\sinh(r) & (k = -1)
\end{cases}
\]

where \( k \) is the geometric curvature of spacetime, the values 0, +1, and −1 indicating flat, positively curved, and negatively curved spacetime, respectively.

Another common form of the metric defines the comoving distance as \( S_0^2(r) \to r \), so that

\[
- \, ds^2 = c \, dt^2 - R^2(t) \left[ \frac{dr^2}{1 - kr^2} + r^2 \left( d\theta^2 + \sin^2 \theta \, d\phi^2 \right) \right],
\]

where \( t \) is again the physical time, and \( r, \theta \) and \( \phi \) are the spatial comoving coordinates, which label the points of the 3-dimensional constant-time hypersurface.

The dimensionless scale factor \( a(t) \) is defined as

\[
a(t) \equiv \frac{R(t)}{R_0},
\]

where \( R_0 \) is the present scale factor (i.e., \( a = 1 \) at present). The scale factor is therefore a function of time, so it can be abbreviated to \( a \). The metric can then be written in a dimensionless form:

\[
- \, ds^2 = c^2 \, dt^2 = c^2 \, dt^2 - a^2 \left[ \frac{dr^2}{1 - kr^2} + r^2 \left( d\theta^2 + \sin^2 \theta \, d\phi^2 \right) \right],
\]
where $S_k(r)$ can be redefined as

$$S_k(r) = \begin{cases} 
R_0 \sin(r/R_0) & (k = +1) \\
r & (k = 0) \\
R_0 \sinh(r/R_0) & (k = -1). 
\end{cases} \tag{30}$$

Equivalently, using the definition in Equation (27),

$$- ds^2 = c^2 dt^2 - a^2 \left[ \frac{dr^2}{1 - k(r/R_0)^2} + r^2 (d\theta^2 + \sin^2 \theta d\phi^2) \right]. \tag{31}$$

The comoving distance is distance between two points measured along a path defined at the present cosmological time. It means that for objects moving with the Hubble flow, the comoving distance remains constant in time. The proper distance, on the other hand, is dynamic and changes in time. At the current age of the universe, therefore, the proper and comoving distances are numerically equal, but they differ in the past and in the future. The comoving distance from an observer to a distant object such as a galaxy can be computed by the following formula:

$$\chi = \int_{t_e}^t \frac{c}{a(t')} \, dt' \tag{32}$$

where $a(t')$ is the scale factor, $t_e$ is the time of emission of photons from the distant object, and $t$ is the present time.

The comoving distance defines the comoving horizon, or particle horizon. This is the maximum distance from which particles could have travelled to the observer since the beginning of the universe. It represents the boundary between the observable and the unobservable regions of the universe. If we take the time at the Big Bang as $t = 0$, we can define a quantity called the conformal time $\eta$ at a time $t$ as:

$$\eta = \int_0^t \frac{dt'}{a(t')} \tag{33}$$

This is useful, because the particle horizon for photons is then simply the conformal time multiplied by the speed of light $c$. The conformal time is not the same as the age of the universe. In fact it is much larger. It is rather the amount of time it would take a photon to travel from the furthest observable regions of the universe to us. Because the universe is expanding, the conformal time is continuously increasing.

The concept of particle horizons is important. It defines causal contact. The only objects not in causal contact are those for which there is no event in the history of the universe that could have sent a beam of light to both. This is at the origin of some of the big questions about the universe associated with the Big Bang model, which gave rise to the Inflationary paradigm (see [147]). We shall discuss this later.

5.1. Cosmological Expansion and Evolution Histories

The FLRW metric relates the spacetime interval $ds$ to the cosmic time $t$ and the comoving coordinates through the scale factor $R(t)$. The scale factor is the key quantity of any cosmological model, since it describes the evolution of the universe. The notion of distance is fairly straightforward in Euclidean geometry. In General Relativity, however, where we work with generally curved spacetime, the meaning of distance is no longer unique. The separation between events in spacetime depends on the definition of the distance being used.
By combining the GR field equation (Equation (15)) and the definition of the metric (Equation (31)), we obtain two independent Einstein equations, known as the Friedmann equations:

\[
\left(\frac{\dot{a}}{a}\right)^2 + \frac{k c^2}{a^2} = \frac{8 \pi G}{3} \rho \tag{34}
\]

and

\[
2 \left(\frac{\dot{a}}{a}\right) + \left(\frac{\ddot{a}}{a}\right) + k c^2 = \frac{-8 \pi G}{c^2} p . \tag{35}
\]

The Friedmann equations relate the total density \( \rho \) of the universe, including all contributions, to its global geometry. There exists a critical density \( \rho_c \) for which \( k = 0 \). By rearranging the Friedmann equation and using the definition of the Hubble parameter we then obtain

\[
\rho_c (t) = \frac{3 H^2 (t)}{8 \pi G} . \tag{36}
\]

A universe whose density is above this value will have a positive curvature, that is, it will be spatially closed (\( k = +1 \)); one whose density is less than or equal to this value will be spatially open (\( k = 0 \) or \( k = -1 \)).

A dimensionless density parameter for any fluid component of the universe (i.e., a component for whose gravitational field is produced entirely by the mass, momentum, and stress density) can be defined by

\[
\Omega (t) = \frac{\rho(t)}{\rho_c (t)} = \frac{8 \pi G \rho(t)}{3 H^2 (t)} . \tag{37}
\]

The current value of the density parameter is denoted \( \Omega_0 \).

Subtracting Equation (34) from Equation (35) yields the acceleration equation:

\[
\frac{\ddot{a}}{a} = -4 \pi G \left( \frac{\rho}{3} + \frac{p}{c^2} \right) . \tag{38}
\]

The geodesic Equation (12) allows us to compute the evolution in time of the energy and momentum of the various components particles which make up the universe. From this evolution, we can construct the fluid equation, or continuity equation, which describes the relation between the density and pressure:

\[
\dot{\rho} + 3 \frac{\dot{a}}{a} (\rho + \frac{p}{c^2}) = 0 . \tag{39}
\]

This is valid for any fluid component of the universe, such as baryonic and nonbaryonic matter, or radiation.

The foundations of the Concordance Model of cosmology depend on General Relativity. Any modification to the theory that changes the Einstein equations will have solutions that differ from the Friedmann equations.

The FLRW universe contains different mass-energy components which are assumed to evolve independently. This is physically valid at late cosmological times, when the components are decoupled, so the density evolutions are distinct. In Table 2, we give the equation of state and the evolution of the density and scale factor for different components of the universe. The quantities in this table are explained in detail in the following sections.
Table 2. The evolution of the various cosmological components. The quantities are the equation of state $w \equiv p/\rho c^2$, the density $\rho$, the pressure $p$, and the scale factor $a(t)$.

<table>
<thead>
<tr>
<th>Component</th>
<th>$w$</th>
<th>$\rho \propto a^{3(1+w)}$</th>
<th>$a(t) \propto t^{2/3(1+w)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Radiation (photons and relativistic neutrinos)</td>
<td>1/3</td>
<td>$\sim a^{-4}$</td>
<td>$\sim t^{1/2}$</td>
</tr>
<tr>
<td>Dust (includes CDM, baryons and non-relativistic neutrinos)</td>
<td>0</td>
<td>$\sim a^{-3}$</td>
<td>$\sim t^{2/3}$</td>
</tr>
<tr>
<td>Curvature</td>
<td>$-1/3$</td>
<td>$\sim a^{-2} \rightarrow a^{-4}$</td>
<td>$t$</td>
</tr>
<tr>
<td>Cosmic strings</td>
<td>$-1/3$</td>
<td>$\sim a^{-2} \rightarrow a^{-4}$</td>
<td>$t$</td>
</tr>
<tr>
<td>Domain walls</td>
<td>$-2/3$</td>
<td>$\sim a^{-1}$</td>
<td>$\sim t^2$</td>
</tr>
<tr>
<td>Inflation</td>
<td>$\rightarrow -1$</td>
<td>$\frac{1}{2} \dot{\phi}^2 + V(\phi)$</td>
<td>$\sim e^{H t}$</td>
</tr>
<tr>
<td>Vacuum energy</td>
<td>$-1$</td>
<td>constant</td>
<td>$\sim e^{H t}$</td>
</tr>
</tbody>
</table>

5.2. Matter (Dust)

Matter which is pressureless is referred to as “dust”. This is a useful approximation for cosmological structures which do not interact, such as individual galaxies. Substituting $p_m = 0$ in the equation of state for dust shows that the density of this component scales as:

$$\rho_m(a) = \frac{\rho_{m,0}}{a^3},$$

where $\rho_{m,0}$ is the current density. Assuming spatial flatness, the time evolution of the scale factor is then

$$a(t) = \left(\frac{t}{t_0}\right)^{2/3},$$

which gives us

$$H(t) = \frac{2}{3t}.$$

This is known as the Einstein-de Sitter (EdS) solution, and it describes the evolution of $H$ in a constant-curvature homogeneous universe with a pressureless fluid as the only component. It was first described by Einstein and Willem de Sitter in 1932 [148].

5.3. Radiation

In the early universe, the energy content was dominated by photons and relativistic particles (especially neutrinos). The expansion of the universe dilutes the radiation fluid, and the wavelength is increased by the expansion so that the energy decreases. From thermodynamics,

$$E_{\text{rad}} = \rho_{\text{rad}} c^2 = \alpha T^4,$$

where $T$ is the radiation temperature and $\alpha$ is the Stefan-Boltzmann constant. The equation of state for radiation can then be derived from the fluid Equation (39):

$$\rho_{\text{rad}}(a) = \frac{\rho_{\text{rad,0}}}{a^4}; \quad p_{\text{rad}} = \frac{\rho_{\text{rad}} c^2}{3}. $$

Combining this with the Friedmann equations, and assuming flatness ($k = 0$), we obtain the time dependence of the scale factor and the Hubble parameter:

$$a(t) = \left(\frac{t}{t_0}\right)^{1/2}; \quad H(t) = \frac{1}{2t}. $$
6. The Components and Geometry of the Universe and Cosmic Expansion

How do we relate the expansion of the universe to its contents? The total density of the universe in terms of its constituent components can be written as the sum of the densities of these components at any given time or scale factor:

\[ \rho = \rho_m + \rho_{\text{rad}} + \rho_{\text{DE}}, \]  

where the subscript “DE” denotes another component of the universe, called Dark Energy.

**Figure 2.** The density evolution of the main components of the universe. The early universe was radiation-dominated, until the temperature dropped enough for matter density to begin to dominate. The energy density of dark energy is constant if its equation of state parameter \( w = 1 \). Because the matter energy density drops as the scale factor increased, dark energy began to dominate in the recent past. At the present time \( (a(t) = 1) \), we live in a universe dominated by dark energy. For dark energy, the green band represents an equation of state parameter \( w = -1 \pm 0.2 \), showing how a small change in the value of this parameter can give very different evolution histories for dark energy. If the Concordance Model is correct, the universe will be completely dominated by dark energy in future epochs (shown by the dashed lines). The matter density will keep decreasing as the universe expands. Our Milky Way will merge with the Andromeda Galaxy, and eventually, the entire Local Group will coalesce into one galaxy. The luminosities of galaxies will begin to decrease as the stars run out of fuel and the supply of gas for star formation is exhausted. In the very far future, this galaxy will be in the only one in our Hubble patch, as all the other galaxies will pass behind the cosmological horizon. The night sky, save for the stars in the Local Group, will be very dark indeed. Stellar remnants will either escape galaxies or fall into the central supermassive black hole. Eventually, baryonic matter may disappear altogether as all nucleons including protons decay, or all matter may decay into iron. In either scenario, the universe will end up being dominated by black holes, which will evaporate by Hawking radiation. The end result is a Dark Era with an almost empty universe, and the entire universe in an extremely low energy state, with a possible heat death as entropy production ceases (see, e.g., [149,150]) What happens after that is speculative.
The total dimensionless density can then be written:

\[ \Omega = \Omega_m + \Omega_{\text{rad}} + \Omega_{\text{DE}}, \]  

(47)

where we have dropped the subscript for clarity, i.e., \( \Omega_{m,0} = \Omega_m \), etc. The Friedmann Equation (34) can now be rewritten using the equations of state for the different components:

\[ H^2(a) = \frac{8\pi G}{3} \left[ \rho_m a^{-3} + \rho_{\text{rad}} a^{-4} + \rho_{\text{DE}} e^{-3 \int_1^a [1+w(a')] d \ln a'} \right] - \frac{k c^2}{a^2}. \]  

(48)

This can be rearranged to give:

\[ H^2(a) = H_0^2 \left[ \Omega_m a^{-3} + \Omega_{\text{rad}} a^{-4} + \Omega_{\text{DE}} e^{-3 \int_1^a [1+w(a')] d \ln a'} + (1 - \Omega) a^{-2} \right], \]  

(49)

or, in terms of redshift:

\[ H^2(z) = H_0^2 \left[ \Omega_m (1+z)^3 + \Omega_{\text{rad}} (1+z)^4 + \Omega_{\text{DE}} e^{-3 \int_0^z [1+w(z')] d \ln z'} + (1 - \Omega)(1+z)^2 \right]. \]  

(50)

The term \( 1 - \Omega \) is sometimes replaced by \( \Omega_k \), the density due to the intrinsic geometry of spacetime. Equation (50) is of central importance since it relates the redshift of an object to the global density components and geometry of the universe.

The density evolution of the various components of the universe is shown in Figure 2.

7. The Hot Big Bang

In the Standard Model, it is generally accepted that the universe arose from an initial singularity, often termed the “Big Bang”, which occurred some 13.8 billion years ago (as measured by Planck [151]). This is not discussed here, but it should be noted that there are several proposals for the mechanism of this singularity. During this epoch, we are dealing with Planck scale physics, so most of these mechanisms involve quantum gravity. Other proposals (such as some superstring and braneworld theories) do away with the need for an initial singularity altogether.

7.1. The Cosmic Microwave Background

The radiation density \( \rho_{\text{rad}} \propto a^{-4} \), so the temperature evolution of the universe from an initial \( T_0 \) is:

\[ T = \frac{T_0}{a}. \]  

(51)

In other words, the universe cools down as it expands. Conversely, this means at early times, when the scale factor was close to zero, the temperature was very high (hence the term “Hot Big Bang”). The radiation left from the early hot universe, cooled by expansion, is known as the Cosmic Microwave Background, or CMB.

The properties of atomic and nuclear processes in an expanding universe provided the first clue for the existence of a hot Big Bang. This was a remarkable achievement of the Big Bang model, because it provided an explanation for the observed abundances of chemical elements in terms of nucleosynthesis. The processes that created nuclei and atoms could only have been possible in an early universe in thermal equilibrium, with black-body spectrum which cooled down as the universe expanded. This allowed Ralph Alpher, Robert Herman, Hans Bethe and George Gamow to predict the existence and temperature of the CMB in 1948 [152–155]. The universe therefore has a thermal as well as an expansion history. Hence the ‘Hot Big Bang’.

The first direct evidence for the Hot Big Bang came two decades later, with the observation of the CMB by Arno Penzias and Robert Woodrow Wilson in 1964 [156].
The confirmation of the thermal history of the universe, together with the discovery of charge parity violation in 1964 [157], provided clues about baryogenesis and the observed matter-antimatter imbalance in the universe. This inspired the first proposals for a mechanism for baryogenesis by Andrei Sakharov in 1967 [158], followed by electroweak symmetry breaking by Vadim Kuzmin in 1970 [159].

This is a remarkable demonstration of the success of the Concordance Model. The cosmological model fits very well with the predictions of particle physics, which in turn can be tested by cosmological observations. The Concordance Model of the structure and evolution of the universe requires a mechanism for baryogenesis as well as an explanation for Dark Matter and dark energy. The challenge for physical theories beyond (or within) the Standard Model is to explain the preference of matter over antimatter, and to explain the magnitude of this asymmetry. Cosmological observations can be used to address these challenges [160].

The CMB is an extremely isotropic source of microwave radiation, with a spectrum corresponding to a perfect blackbody at a temperature $T_0 = 2.7260 \pm 0.0013$ K [161]. Using the current temperature and $E_{\text{rad}} = \rho_{\text{rad}} c^2 = \alpha T^4$, the radiation density today is given by:

$$\Omega_{\text{rad}} = 2.47 \times 10^{-5} h^{-2}. \quad (52)$$

At some time in the early universe, the ambient radiation temperature corresponded to the ionisation potential of hydrogen, which is 13.6 eV. During this epoch, the universe was filled with a sea of highly energetic particles and photons—a hot ionised plasma. The particles were mainly electrons and protons. Other fundamental particles (quarks) existed earlier when the ambient energy corresponded to their rest mass. At some point, as the universe expanded and cooled, the energy of the photons was no longer sufficient to ionise the hydrogen, and within a relatively short time, all of the electrons and protons combined to form neutral hydrogen. The photons were then free to move through the universe. This process is known as decoupling and it occurred at a temperature of ~2500 K, when the universe was approximately 380,000 years old [162]. It is these decoupled photons which make up the CMB. The surface on the sky from which these photons originate is known as the surface of last scattering.

### 7.2. Matter-Radiation Equality

At the present epoch, neglecting dark energy, the universe is dominated by matter. This component is characterised by the fact that the matter particles can be treated in a non-relativistic regime, whereas photons and relativistic neutrinos both behave like radiation. The total contribution to the energy density from non-relativistic components (matter) and relativistic components (radiation and relativistic neutrinos) can be written as $\Omega_{\text{NR}}$ and $\Omega_{\text{R}} = \Omega_{\text{rad}} + \Omega_{\nu}$, respectively. Using the fact that $\Omega_m = \Omega_{m,0} a^{-3}$, the ratio of the contributions of the components is a function of the scale factor $a$:

$$\frac{\Omega_R}{\Omega_{\text{NR}}} = \frac{\Omega_{\text{rad}} + \Omega_{\nu}}{\Omega_m} = \frac{4.15 \times 10^{-5}}{\Omega_{m,0} h^2 a^{-3}}, \quad (53)$$

where we explicitly use the subscript 0 for the present-day values.

Then there must exist a scale factor for which the ratio is unity. This is given by:

$$a_{\text{eq}} = \frac{4.15 \times 10^{-5}}{\Omega_{m,0} h^2}, \quad (54)$$

or, in terms of redshift,

$$1 + z_{\text{eq}} = 2.4 \times 10^4 \Omega_{m,0} h^2. \quad (55)$$

The epoch at which the matter energy density equals the radiation energy density is called matter-radiation equality, and it has a special role in large-scale structure formation.
7.3. Neutrinos

Neutrinos have particular properties which give rise to a distinct evolution history. They are known to exist from the Standard Model of particle physics, and the Hot Big Bang model predicts the amount of neutrinos in the universe. Neutrinos can be thought of as “dark” matter because of their very small reaction cross-section, which implies negligible self-interaction. However, they are not cold Dark Matter. They are simply extremely light particle that can stream out of high-density regions. They therefore cause the suppression of perturbations on scales smaller than the free-streaming scale. Unlike photons and baryons, cosmic neutrinos have not been observed. However, particle physics allows us to chart the history of this particle during nucleosynthesis, and to relate the neutrino temperature to the photon temperature today [163–165].

The scale on which perturbations are damped by neutrinos is determined by the comoving distance that a neutrino can travel in one Hubble time at equality. For a neutrino mass \( \sim 1 \text{ eV} \), the average velocity, \( T_\nu/m_\nu \), is of order unity at equality. This leads to a suppression of power on all scales smaller than \( k_{eq} \). Note that this phenomenon depends on the individual neutrino mass, rather than the total neutrino mass. A lighter neutrino can free-stream out of larger scales, so the suppression begins at lower \( k \) for the lighter neutrino species. Heavier neutrinos constitute more of the total neutrino density, and so suppress small-scale power more than lighter neutrino species, which means that we need at least two parameters to model massive neutrino phenomenology to sufficient accuracy: the neutrino mass fraction \( \Omega_\nu \), or some expression of this quantity in terms of the total neutrino mass \( \sum m_\nu \), and the number of massive neutrino species \( N_\nu \).

Neutrinos introduce a redshift and scale dependence in the transfer function. We know that the perturbation modes of a certain wavelength \( \lambda \) can grow if they are greater than the Jeans wavelength. Above the Jeans scale, perturbations grow at the same rate independently of the scale. For the baryonic and cold Dark Matter components, the time and scale dependence of the power spectrum can therefore be separated at low redshifts. This is not the case with massive neutrinos, which introduce a new length scale given by the size of the comoving Jeans length when the neutrinos become non-relativistic. In terms of the comoving wavenumber \( k_{nr} \), this scale is given by:

\[
k_{nr} = 0.026 \left( \frac{m_\nu}{1 \text{ eV}} \right)^{1/2} \Omega_\nu^{1/2} h \text{ Mpc}^{-1}
\]

for three neutrinos of equal mass, each with mass \( m_\nu \). The growth of Fourier modes with \( k > k_{nr} \) is suppressed because of neutrino free-streaming. From the equation above, it is evident that the free-streaming scale varies with the cosmological epoch (since there is a dependence on \( \Omega_m \)), and therefore the scale and time dependence of the power spectrum cannot be separated.

Neutrinos are fermions, with a Fermi-Dirac distribution with assumed zero chemical potential. When they decoupled from the plasma, their distribution remained Fermi-Dirac, with their temperature falling as \( a^{-1} \). This decoupling occurred slightly before the annihilation of electrons and positrons, which occurred when the cosmic temperature was of the order of the electron mass (\( T \approx m_e \)). Neutrinos decoupled when the cosmic plasma had a temperature of around 1 MeV. The energy associated with this annihilation was therefore not inherited by the neutrinos, and the entropy was completely transferred to the entropy of the photon background. Thus:

\[
(S_e + S_\gamma)_{\text{before}} = (S_\gamma)_{\text{after}},
\]

where \( S_e \) and \( S_\gamma \) are respectively the entropy of the electron-positron pairs and the photon background, and ‘before’ and ‘after’ refer to the annihilation time.

The entropy per particle species, ignoring constant factors, is \( S \propto gT^3 \), where \( g \) is the statistical weight of the species. For bosons, \( g = 1 \) and for fermions, \( g = 7/8 \) per spin state. According to the Standard Model, the neutrino has one spin degree of freedom, each neutrino has an antiparticle, and there are three generations of neutrinos, also called “families” or “species” (\( \mu \), \( \tau \) and electron...
neutrinos). This means that the degeneracy factor of neutrinos is equal to 6. Before annihilation, the fermions are electrons (2 spin states), positrons (2), neutrinos and antineutrinos (6 spin states). The bosons are photons (2 spin states). We therefore have \( g_{\text{before}} = \frac{4}{7} + 2 = 11/2 \), while after annihilation \( g = 2 \) because only photons remain. Applying entropy conservation and counting relativistic degrees of freedom, the ratio of neutrino and photon temperatures below \( m_e \) is therefore:

\[
\frac{T_\nu}{T_\gamma} = \left(\frac{4}{11}\right)^{1/3},
\]

so that the present neutrino temperature is

\[
T_{\nu,0} = \left(\frac{4}{11}\right)^{1/3} T_{\text{CMB}} = 1.945 \text{ K}.
\]

The number density of neutrinos is then

\[
n_\nu = \frac{6 \zeta(3)}{11 \pi^2} T_{\text{CMB}}^3,
\]

where \( \zeta(3) \approx 1.202 \), which gives \( n_\nu \approx 112 \text{ cm}^{-3} \) at the present epoch [166]. In the early universe, neutrinos are relativistic and behave like radiation. So they contribute to the total radiation energy density \( \rho_{\text{rad}} \), which includes the photon energy density \( \rho_\gamma \):

\[
\rho_{\text{rad}} = \left[ 1 + \left(\frac{7}{8}\right) \left(\frac{4}{11}\right)^{4/3} N_{\text{eff}} \right] \rho_\gamma,
\]

where \( N_{\text{eff}} \) is the effective number of neutrino species. At late times, when massive neutrinos become non-relativistic, their contribution to the mass density is \( m_\nu n_\nu \), giving

\[
\Omega_\nu = \frac{\rho_\nu}{\rho_c} \approx \frac{N_i}{93.14 \text{ eV} \ h^2},
\]

where \( m_{\nu,i} \) is the mass of individual neutrino species and \( N_i \) is the number of massive neutrino species. This expression relates the total neutrino mass \( \sum m_\nu \) to the neutrino fraction \( \Omega_\nu \).

It can be seen from the above that this equation can be modified through a change in the effective number of neutrino species by many factors: a non-zero initial chemical potential, or a sizeable neutrino-antineutrino asymmetry, or even a fourth, ‘sterile’ neutrino [166–168]. The Standard Model predicts a value of \( N_{\text{eff}} = 3.046 \) for the effective number of neutrino species. This accounts for the three neutrino families together with relativistic degrees of freedom, since neutrinos are not completely decoupled at electron-positron annihilation and are subsequently slightly heated [169]. Any significant deviation from this value could be a signature of hidden physical effects, possibly requiring a modification of General Relativity [170].

Neutrino oscillation experiments do not, at present, determine absolute neutrino mass scales, since they only measure the difference in the squares of the masses between neutrino mass eigenstates. Cosmological observations, on the other hand, can constrain the neutrino mass fraction, and can distinguish between different mass hierarchies [166].

Observations of neutrino flavour oscillations in atmospheric and solar neutrinos, provide evidence of a difference between the masses of the different species or flavours, as well as for a non-zero mass. For three neutrino mass eigenstates \( m_1, m_2 \) and \( m_3 \), the squared mass differences are [171]:

\[
\Delta m_{12}^2 = m_1^2 - m_2^2, \quad \Delta m_{23}^2 = m_2^2 - m_3^2, \quad \Delta m_{13}^2 = m_1^2 - m_3^2.
\]
\[ \Delta m_{21}^2 = m_2^2 - m_1^2 \approx 7.5 \times 10^{-5} \text{eV}^2 \]
\[ \Delta m_{31}^2 = m_3^2 - m_1^2 \approx 2.5 \times 10^{-3} \text{eV}^2 \]
\[ \frac{\Delta m_{21}^2}{\Delta m_{31}^2} \approx 0.03. \]  

The ambiguity in the sign of the mass differences \( \Delta m \) allows for two possible mass hierarchies: the normal hierarchy given by the scheme \( m_3 \gg m_2 > m_1 \), or the inverted hierarchy \( m_2 > m_1 \gg m_3 \). Given Equation (63), constraining the total neutrino mass to a small enough maximum value could exclude an inverted hierarchy. Conversely, a total neutrino mass \( m_\nu \sim 2 \text{eV} \) is only possible with a degenerate neutrino mass scheme. Hence the interest in finding cosmological neutrino mass bounds.

The fact that cosmological constraints could be stronger than constraints from particle accelerators was noticed quite early (see [172]). The ‘closure limit’ gives us \( m_\nu < 90 \text{eV} \). This was first derived in the late 1960s and 1970s [173–176]. Since then, cosmological neutrino bounds have improved significantly, with different methods being used e.g., luminous red galaxies [177], CMB anisotropies [151,178], or weak lensing [179–181].

Joint Planck CMB and BAO observations give us \( m_\nu < 0.23 \text{eV} \), but various data combinations can change this figure, and strong priors on the value of the Hubble constant can provide tighter constraints [151].

8. Inflation: The Second Wave of Alternative Theories

In the late 1970s, General Relativity had been largely accepted by the scientific community. But a series of cosmological considerations led to renewed interest in alternative theories. These were not so much attempt to solve problems in the theory itself, but to find explanations for observations that were not explained by the theory.

General Relativity applied to the universe gave us the Hot Big Bang model: a universe expanding out of an initial highly energetic, dense state. The Hot Big Bang model was successful in explaining many interlinked phenomena which were subsequently confirmed by observation: the Hubble Law and the expansion of the universe, the thermal history of the universe, primordial nucleosynthesis, the existence of the cosmic microwave background, the relation between the temperature and scale factor, and finally the blackbody nature of the CMB. The remarkable fact is that these phenomena occur on extremely different scales, and are observed via different physical processes, and yet they all fit neatly within one model.

However, there are some observations which the Hot Big Bang model fails to explain. These cosmological problems are linked to the primordial universe, the most obvious being the following (for details see [182–186], and references therein):

- The Horizon Problem
- The Flatness Problem
- The Monopole Problem

The horizon problem arises from the structure of spacetime. In the standard cosmological model described by the FLRW equations, different regions of the universe observed today could have not been in causal contact with because of the great distances between them which are greater than the distance that could have been traversed by light since the Big Bang. The transfer of information (i.e., any physical interaction) or energy can occur, at most, at the speed of light, but these regions have the same temperature and other physical characteristics. In particular, we observe causally-disconnected regions of the CMB to be in thermal equilibrium. How could this have happened?

The horizon problem was first identified in the late 1960s. This led to to early attempts to model chaotic solutions to Einstein’s field equations near the initial singularity [187,188]. In the late 1970s, Alexei Starobinsky noted that quantum corrections to General Relativity should be important in the
very early universe. These corrections would lead to a modification of gravity, which induces an inflationary phase [189]. Starobinsky’s was the first model of inflation.

The flatness problem is one the so-called “coincidence problems” of modern cosmology. By the 1960s, observations had determined that the density of matter in the universe is comparable to the critical density necessary for a flat universe. So the contribution of curvature had to be of the same order of magnitude as the contribution of matter throughout the history of the universe. This represents a fine-tuning problem. Observations of the CMB have confirmed that the universe is spatially flat to within a few percent. Why is the global geometry of the universe so flat?

The magnetic monopole problem arises from the Hot Big Bang model. Grand Unified Theories predict the production of a large number of magnetic monopoles [190,191] in the early, extremely hot universe. Why have none ever been observed? If they exist at all, they are much more rare than the Big Bang theory predicts. This was noted by Zel’dovich and others in the late 1970s [192,193].

**Hot Big Bang Plus Inflation**

This gave rise to the idea of a model in which the early universe undergoes a period of exponential accelerated expansion. This theory, called “inflation”, was first formulated by Alan Guth in the 1980s [194] while he was trying to investigate why no magnetic monopoles are observed. It was realised that inflation solves the horizon and flatness problems, as well as explaining the absence of relic monopoles. Better still, it explains the origin of structure in the universe.

In the Standard ΛCDM Model, the initial perturbations from which structure evolved are assumed to have been seeded by the inflationary potential. Reconstructing the primordial power spectrum is no easy task, and poses two main problems. Observationally, we want to extract the amplitude and scale variation from the data. Theoretically, we seek to explain the origin of the perturbations. At present, the leading theoretical paradigm for the primordial fluctuations is inflation, which provides initial conditions for both large-scale structure and the cosmic microwave background radiation. The theory of inflation offers a plethora of models, each of which predicts a certain power spectrum of primordial fluctuations $P(k)$. Since the inflationary paradigm is linked to the theoretical description of the primordial power spectrum, it is necessary to briefly explain some of the main concepts here (for the full details, see [183,185]).

The precise definition of inflation is any period during which the scale factor of the universe is accelerating, that is, $\ddot{a} > 0$. This expression is equivalent to other definitions of inflation:

$$\frac{d}{dt} \left( \frac{H^{-1}}{a} \right) < 0 \quad \iff \quad \dot{e} \equiv - \frac{\dot{H}}{H^2} < 1 \quad \iff \quad \frac{d^2 a}{dt^2} > 0 \quad \iff \quad \rho + 3p < 0. \quad (64)$$

The first expression above has a remarkable physical interpretation. It means that the observable universe becomes smaller during inflation.

The basic theory of inflation states that from the initial Big Bang singularity to approximately $10^{-37}$ s, there existed a set of highly energetic scalar fields. By definition, $\Omega$ is driven towards 1 during inflation. Inflationary theories assume that gravity is described by GR, which means that the component driving inflation must satisfy $\rho + 3p < 0$. If for example, the universe was dominated during the inflationary phase by a scalar field (or set of fields) $\phi$ with a self-interaction potential $V(\phi)$. It is the form of this potential which differentiates the various inflationary theories. Most theories assume a “Mexican hat” potential, with a single field, while chaotic inflation assumes a simple power law potential with a slowly varying field [195]. The action for this potential is then [196]

$$S = - \int d^4x \sqrt{-g} \left[ \frac{m_{Pl}^2 R}{16\pi} - \frac{1}{2} (\nabla \phi)^2 + V(\phi) \right], \quad (65)$$
where $m_{Pl}$ is the Planck mass. As the universe cooled, the scalar field became trapped in a false vacuum, so its energy density became constant. The potential energy, however, is nonzero, so the pressure is negative. The scale factor during inflation has the de Sitter form:

$$a(t) = e^{(\Lambda_I/3)^{1/2}t},$$ \hspace{1cm} (66)

where $\Lambda_I$ represents the energy density of the inflationary field (sometimes called the inflaton).

Since the energy density of the inflaton field was very high, the associated magnitude of the negative pressure would have been very large as well. The scale factor is thought to have increased during inflation by $\sim e^{65}$, and any point in the universe which found itself in a false vacuum state would have undergone inflation. The accelerated expansion lasted until the field rolled down to a minimum, when it decays into the familiar particles of the Standard Model, and the universe can then be described by an FLRW model.

The inflationary paradigm provides an explanation for the origin of structure and for the observed geometry of the universe, in addition to solving the aforementioned cosmological problems.

First, inflation solves the flatness problem. Using Equation (66), the evolution of $\Omega$ during inflation can be written as:

$$|\Omega(t) - 1| \propto e^{-(4\Lambda_I/3)^{1/2}t},$$ \hspace{1cm} (67)

so that $|\Omega - 1|$ is driven very close to 0 as $t$ increases. This explains why the universe is flat.

It also means that this value has not deviated significantly from its initial value right after expansion. We can therefore safely assume spatial flatness throughout the history of the universe. Given the observational difficulties, this provides a theoretical motivation for taking the idea of a large $\Omega_{DE}$ seriously.

Second, inflation solves the horizon problem (Figure 3). Regions of the universe which are causally disconnected today evolved out of the same causally-connected region in the early universe. The observed uniformity of the CMB is no longer a problem.

Third, inflation explains why we have never observed magnetic monopoles. Due to rapid expansion of the universe during inflation, they become so rare in any given volume of space that we would be very unlikely to ever encounter one. Nor would they have sufficient density to alter the gravity and thereby the normal expansion of the universe following inflation. The problem of magnetic monopoles motivated the Guth’s development of his theory in 1981 [194]. The solution of the monopole problem, and problems related to other relics, was an early success of the inflationary paradigm, and inspired similar theories [197,198].

Fourth, the inflationary scenario provides a natural explanation for the origin of structure, providing a link between quantum mechanics and relativistic cosmological paradigm. This was realised soon after the development of the theory of inflation, and the details were worked out in the early 1980s [199–204].

An initially smooth background needs seed fluctuations around which gravitational collapse can occur. The inflationary scenario attributes their origin to quantum fluctuations in the inflaton field potential, so that the universe is not perfectly symmetric. Different points in the universe inflate from slightly different points on the potential, separated by $\delta \phi$. Inflation for these two points ends at different times, separated by $\delta t = \delta \phi / \dot{\phi}$. This induces a density fluctuation $\delta = H \delta t$ (see [184]). Since all the points undergoing inflation are part of the same potential field, the initial fluctuations are nearly scale invariant. This means that the density amplitude on the horizon scale will also be constant:

$$\delta_H = H \delta t = \frac{H^2}{2\pi \dot{\phi}} = \text{constant}. \hspace{1cm} (68)$$

In summary, inflation solves the three cosmological problems listed above:

- The Horizon Problem. Solution: the entire universe evolved out of the same causally-connected region.
• The Flatness Problem. Solution: any initial curvature is diluted by the inflationary epoch and driven to zero.
• The Monopole Problem. Solution: the rapid expansion of the universe drastically reduces the predicted density of magnetic monopoles, if they exist.

![Figure 3. How inflation solves the horizon problem. The light cones on the causal diagram of an inflationary FLRW model are at ±45°. The worldlines of comoving matter are vertical on this kind of diagram. The particle horizons are horizontal lines. Here we have shown the particle horizon for the CMB. Without inflation, conformal time would only go back to \( \tau_0 \), and different regions of the CMB which we observe today along our past light cone would never have been in causal contact. Because of inflation, conformal time is extended to the Big Bang singularity, so these regions would have been in causal contact at some point in our past light cone.]

How long did inflation last? The answer is given by looking closely at Equation (67) above. A convenient measure of expansion is the so-called \( e \)-fold number, defined as:

\[
N \equiv \ln \left( \frac{a_f}{a_i} \right) = \int_{t_i}^{t_f} H \, dt.
\]  
(69)

Here, \( a_i \) and \( a_f \) are the values of the scale factor at the beginning and end of inflation, while \( t_i \) and \( t_f \) are the corresponding proper times. The scale factor \( a \) is only physically meaningful up to a normalisation constant, so the \( e \)-fold number is defined with respect to some chosen origin. The reason
is that in cosmology, what is fixed is not the initial condition, but the current expansion—we cannot measure any $H$, but we measure $H_0$ then extrapolate backwards.

We can search for the minimum duration of inflation required to solve the horizon problem. At the very least, we require that the observable universe today fits in the comoving Hubble radius at the beginning of inflation:

$$ (a_0H_0)^{-1} < (a_iH_i)^{-1}. \quad (70) $$

The condition is the same for the horizon and flatness problems.

If we assume that the universe was radiation-dominated since the end of inflation (giving us $H \propto a^{-2}$), and ignore the relatively recent matter- and dark energy-dominated epochs, we obtain

$$ \frac{a_f}{a_i} = \frac{a_0}{a_f}. \quad (71) $$

In the general case, the condition becomes:

$$ \frac{a_f}{a_i} \geq \frac{a_0}{a_f}, \quad (72) $$

or in terms of the number of $e$-folds,

$$ N_f - N_i \geq N_0 - N_f. \quad (73) $$

In other words, there should be as much expansion during inflation as after inflation.

The solution to the horizon problem is the same as the solution to the flatness problem. Taking into account the present energy density of the universe, we need a minimum of about 50 to 60 $e$-folds. This already gives us a useful criterion for realistic inflation models. The most recent Planck results show a preference for a higher number of $e$-folds: $78 < N < 157 \ [205]$.

Most models of inflation are slow-roll models, in which the Hubble rate varies slowly $[185,206,207]$. This model was first developed by Andrei Linde in 1982 $[208]$. It solved a major problem in Guth’s early theory. Instead of tunnelling out of a false vacuum state, inflation occurs by a scalar field rolling down a potential energy gradient. When the field rolls very slowly compared to the expansion of the universe, inflation occurs. Hence the name “slow-roll inflation”. However, when the gradient becomes steeper, inflation ends and reheating can occur. It is beyond the scope of this review to go into the detail of the theory, but it is necessary for us to briefly refer to the link between this theory and the spectral index of primordial fluctuations, which is an important observational parameter in the Concordance Model of cosmology.

To quantify slow roll, cosmologists typically use two parameters $\epsilon$ and $\eta$ which vanish in the limit that $\phi$ becomes constant. The first parameter is defined as:

$$ \epsilon \equiv \frac{d}{dt} \left( \frac{1}{H} \right) = -\frac{\dot{H}}{aH^2}, \quad (74) $$

which is always positive, since $H$ is always decreasing. The second complementary variable which defines how slowly the field is rolling is:

$$ \eta \equiv \frac{1}{aH\phi^{(0)}} \left[ 3aH\dot{\phi}^{(0)} + a^2V' \right], \quad (75) $$

where $\phi^{(0)}$ is the zero-order field, and $V$ is the potential.

The scalar spectral index can be defined in terms of some function, usually a polynomial, involving the two slow-roll parameters $\epsilon$ and $\eta$. As an example we shall give two such parameterisations:

$$ n = 1 - 4\epsilon - 2\eta \ [209] $$

and

$$ n = 1 - 6\epsilon + 2\eta \ [206]. $$

The rate of change of $n$ can also be expressed in terms of inflationary parameters: $dn/d\ln k = 16\epsilon\eta + 24\epsilon^2 + 2\xi^2 \ [210]$, where
\[ \xi^2 \equiv \frac{m_{Pl}^4}{V} \left( \frac{d^3}{d\phi^3} \right) \frac{V'}{\sqrt{2}}, \] (76)

\( m_{Pl} \) being the reduced Planck mass (4.342 \times 10^{-6} \text{g}).

Therefore, by extracting the values of \( \epsilon \) and \( \eta \) from the data, using methods such as weak lensing, we can directly probe the potential of the inflaton field. Likewise for the tilt or spectral index of the primordial power spectrum. Slow-roll inflation predicts that the spectral index of primordial fluctuations should be slightly less than 1. The reason for this is simple. For inflation to end, the Hubble parameter \( H \) has to change in time. This time-dependence changes the conditions at the time when each fluctuation mode exits the Hubble horizon and therefore gets translated into a scale-dependence.

Inflation accounts for the observed spatial flatness of the universe, and the absence of magnetic monopoles. These predictions have been confirmed by various probes, most notably by precision measurements of CMB anisotropies, starting with the Cosmic Background Explorer (COBE) in 1992 [211–213], then with the Wilkinson Microwave Anisotropy Probe (WMAP) which ran for nine years from 2001 to 2010. WMAP data placed tight constraints on the predicted burst of growth in the very early universe, providing compelling evidence that the large-scale fluctuations are slightly more intense than the small-scale ones, which is a subtle prediction of many inflation models [162,214–219]. Significantly, WMAP found evidence that the scalar spectral index is less than 1 (a 2\( \sigma \) deviation), implying a deviation from scale invariance for the primordial power spectrum. As explained above, this is a major prediction of inflation, and this observation reinforced the evidence in favour of the theory. Conclusive proof of a scale-dependent primordial power spectrum (a 5\( \sigma \) deviation from \( n_s = 1 \)) was provided by the Planck CMB anisotropy probe in 2013 [220,221] and confirmed in 2015 [151,205].

One current experimental challenge is to observe the B-modes of polarisation of the CMB caused by primordial gravitational waves produced by inflation. Their detection by the BICEP2 experiment was announced in early 2014. However, more accurate modelling of the signal over the next few months, which allowed the observation to be explained by polarised dust emission in our Galaxy, decreased the statistical confidence of the initial result [222]. This was confirmed by Planck data in 2016 [223]. Upcoming large-scale structure surveys, such as the Euclid satellite mission, or 21-cm radiation surveys such as the Square Kilometre Array, may measure the power spectrum with greater precision than current CMB probes, and could provide further evidence in favour of the inflationary paradigm [224–226].

9. The First Unknown Component: Dark Matter

The first evidence for Dark Matter came from astronomy rather than cosmology. Newtonian physics and General Relativity both provide very precise rules for the dynamics of galaxies: the mass determines the rotation velocity. Starting in the 1920s, astronomers noticed that amount of visible matter in galaxies did not match the observed rotation curves. These curves relate the tangential velocity of the constituent stars (or gas) about the centre of the galaxy to their radial distance. Observations of the velocities of globular clusters about galaxies showed that at large radii the velocities are approximately constant, implying that the amount of mass in the galaxies is much higher than the visible mass.

The first suggestion of the existence of hidden matter, motivated by stellar velocities, was made by Jacobus Kapteyn in 1922 [227]. Radio astronomy pioneer Jan Oort also hypothesized the existence of Dark Matter in 1932 [228]. Oort was studying stellar motions in the local galactic neighbourhood and found that the mass in the galactic plane must be greater than what was observed. This measurement was later determined to be erroneous.

In 1933, Fritz Zwicky, who studied galactic clusters while working at the California Institute of Technology, made a similar inference [229]. Zwicky applied the virial theorem to the Coma galaxy cluster and obtained evidence of unseen mass that he called dunkle Materie in German, or “Dark Matter”. Zwicky estimated its mass based on the motions of galaxies near its edge and compared that
to an estimate based on its brightness and number of galaxies. He estimated that the cluster had about 400 times more mass than was visually observable. The gravity effect of the visible galaxies was far too small for such fast orbits, thus mass must be hidden from view. Based on these conclusions, Zwicky inferred that some unseen matter provided the mass and associated gravitation attraction to hold the cluster together.

In 1937, Zwicky made the bold assertion that galaxies would be unbound without some form of invisible matter [230]. Zwicky’s estimates were off by more than an order of magnitude, mainly due to an obsolete value of the Hubble constant. The same calculation today shows a smaller fraction, using greater values for luminous mass. However, Zwicky did correctly infer that the bulk of the matter was dark.

More evidence started to accumulate for the existence of some non-emitting component which was now being called Dark Matter [231]. In 1959, Kahn and Woltjer [232] pointed out that the motion of Andromeda towards us implied that there must be Dark Matter in our Local Group of galaxies. Dynamical evidence for massive Dark Matter halos around individual galaxies came later, starting in the 1970s, when rotation curve data from multiple galaxies confirmed the Dark Matter halo hypothesis [233–236]. Like baryonic matter, Dark Matter is a fluid with vanishingly small pressure. Unlike baryonic matter, it has no interaction with photons, making it both dark and transparent. It also has a vanishingly small self-interaction beside gravity. One result of this is that the Dark Matter halos surrounding galaxies are rounder than the galaxies themselves [237].

In the last few decades, cosmology has contributed one important piece of information: the amount of Dark Matter. The observed value of the matter density in the universe is $\Omega_m = 0.3089 \pm 0.0062$. However, the density of baryonic matter is $\Omega_b = 0.0486 \pm 0.0010$ [151]. The missing mass is made up of Dark Matter.

The name “Dark Matter” is an indication of its nonbaryonic nature: it cannot be observed by emission of photons, so observers need to find a way around this problem. Current evidence for the existence of Dark Matter comes from a variety of sources besides galactic dynamics [231]. The two most important ones are CMB anisotropies and gravitational lensing. In addition, Big Bang nucleosynthesis provides evidence that some of the Dark Matter may be baryonic. The inventory of observed baryons in the local universe falls short of the total anticipated abundance from Big Bang nucleosynthesis, implying that most of the baryons in the universe are unseen [238].

Anisotropies in the CMB are related to anisotropies in the baryonic density field by the Sachs-Wolfe effect [239]. This means that the baryon density field variation at the time of decoupling can be linked to CMB anisotropies. If all matter were made of baryons, the amplitude of the density fluctuations should have reached $\delta \sim 10^{-2}$ at the present epoch. However, we observe structures with $\delta \gg 1$ at the present epoch (e.g., galaxies and galaxy clusters). The discrepancy can only be explained by the presence of additional matter, which created potential wells for the baryons to fall into after decoupling. These potential wells would have had to be formed by a weakly interacting fluid that decoupled well before baryons and began to cluster much earlier. Such a fluid would only interact via the gravitational and possibly the weak nuclear force. As the baryons accumulated in the potential wells, their pressure would have built up, leading to oscillations in the baryon fluid, termed “baryon acoustic oscillations” (BAO) [240,241]. These oscillations leave an imprint on the CMB power spectrum, which has been confirmed observationally, and which constrains the mass density, leading to a further confirmation of the existence of this missing mass.

The phenomenon of gravitational lensing includes cosmic shear, weak lensing, cosmic magnification. Although the theory of cosmic shear had been worked out from the 1960s to the early 1990s [242], the first detection had to await the development of instruments sensitive enough to make the required observations, and image analysis software to accurately correct for unwanted effects when measuring the shapes of galaxies. In March 2000, four groups independently announced the first discovery of cosmic shear [243–246]. Since then, cosmic shear has established itself as an important technique in observational cosmology.
Gravitational lensing shows that the amount of lensing of galaxies around galaxy clusters is too high to be caused by the visible matter. Apart from the stars themselves, a galaxy cluster also has a gas component, but X-ray observations show that this is still not enough to account for the extra mass. The cluster must therefore have a non-emitting halo of Dark Matter around it.

Various Dark Matter candidates have been proposed [247–251]. However, all these candidates have one common characteristic: a very small reaction cross-section, making them extremely difficult to detect directly [252–255]. Experiments have, however, placed limits on the mass of Weakly Interacting Massive Particles (WIMPs), which are the current best candidate for Dark Matter (together with axions). WIMPs are an entire new class of fundamental particle outside of the Standard Model that result from supersymmetry [256,257]. These results show that even the lightest Dark Matter particle should have a mass which is not below $\sim 10$ MeV. We also know that $\Omega_{\text{CDM}} = 0.2589 \pm 0.0057$ [151]. The conclusion is that $\Omega_{\nu} \ll \Omega_{m}$, implying that hot Dark Matter (i.e., neutrinos) cannot account for the Dark Matter density $\Omega_{\text{CDM}}$.

An alternative to Dark Matter is to explain the missing mass by means of a modification of gravity at large distances or more specifically at small accelerations. In 1983, Morderhai Milgrom proposed a phenomenological modification of Newton’s law which fits galaxy rotation curves [131–133]. The theory, known as Modified Newtonian Dynamics (MOND) automatically recovers the Tully-Fisher law. The theory modifies the acceleration of a particle below a small acceleration $a_0 \sim 10^{-10}$ ms$^{-2}$, which therefore enters the theory as a universal constant. The gravitational acceleration at large distances then reads $a = \sqrt{GMa_0/r}$ at large distances, instead of the Newtonian law $a = \sqrt{GM/r^2}$.

There are two main difficulties with MOND. First, it does not explain how galaxy clusters can be bound without the presence of some hidden mass [258,259]. Second, attempts to derive MOND from a consistent relativistic field theory have failed. One such attempt is the Tensor-Vector-Scalar Theory (TeVeS) [135] is more successful and actually relativistic but not apparently necessary since it still requires dark energy and Dark Matter. Many models are unstable [260], or require actions which depend on the mass $M$ of the galaxy, thereby giving a different theory for each galaxy. Moreover, modified gravity theories have serious difficulties reproducing the CMB power spectrum and the evolution of large-scale structure [261–263].

The greatest challenge to modified gravity theories, and also the clearest direct evidence of Dark Matter, comes from observations of a pair of colliding galaxy clusters known as the Bullet Cluster [264] in which the stars and Dark Matter separate from the substantial mass of ionised gas. The Dark Matter follows the less substantial stars and not the more massive gas.

The modifications of gravity proposed as alternative to the Dark Matter paradigm illustrate the need for tests of GR at large distances and low accelerations. They also illustrate the problems faced by models which favour goodness of fit over parsimony. Modified gravity theories can give an excellent phenomenological fit through an adjustment of the values of the extra parameters, but there is no universal principle to determine these values. This requirement for simplicity and predictivity is met by General Relativity.

10. The Second Unknown: Dark Energy and the New Wave Alternative Theories

The current motivation for alternative theories seems to be the search for an explanation of the observed accelerating expansion of the universe. Let us consider the justification for the dark energy paradigm within the inflationary $\Lambda$CDM model, and the process which led to its acceptance by the scientific community (see [265]).

Round about the time that GR was developed, the universe was thought to be static. There was no compelling reason to think otherwise. Einstein realised that his equations implied a non-static universe, so in 1917 he revised his field equations of GR to read [138]:

\[ G_{\mu\nu} - \Lambda g_{\mu\nu} = G_{\mu\nu} - 8\pi G \rho_{\Lambda} g_{\mu\nu} = 8\pi G T_{\mu\nu} \]  \hspace{1cm} (77)
where $\rho_\Lambda = \Lambda / 8\pi G$ is proportional to the cosmological constant $\Lambda$.

It can be seen from this equation that Einstein did not consider the cosmological constant to be part of the stress-energy term. One could, of course, put $\rho_\Lambda g_{\mu\nu}$ on the right-hand side of the equation and count it as part of the source term of the stress-energy tensor and simply consider $\rho_\Lambda$ to be the vacuum energy. This is not just a semantic distinction. When $\rho_\Lambda$ takes part in the dynamics of the universe, then the field equation is properly written with $\rho_\Lambda$, or its generalisation, as part of the stress-energy tensor:

$$G_{\mu\nu} = 8\pi G (T_{\mu\nu} + \rho_\Lambda g_{\mu\nu}).$$ (78)

The equation describing gravity is then unchanged from its original form—there is no new physical theory. Instead, there is a new component in the content of the universe.

This component must satisfy Special Relativity (that is, an observer can choose coordinates so that the metric tensor has Minkowskian form). An observer moving in spacetime in such a way that the universe is observed to be homogeneous and isotropic would measure the stress-energy tensor to be

$$T_{\mu\nu} = \begin{pmatrix} \rho & 0 & 0 & 0 \\ 0 & p & 0 & 0 \\ 0 & 0 & p & 0 \\ 0 & 0 & 0 & p \end{pmatrix}. \quad (79)$$

This means that the new component in the stress-energy tensor looks like an ideal fluid with negative pressure:

$$p_\Lambda = -\rho_\Lambda. \quad (80)$$

In modern concordance cosmology, this component is usually termed “dark energy”. If the equation of state parameter of dark energy is constant, i.e., $w(z) = -1$, then its energy density will be constant regardless of the expansion of the universe.

Einstein inserted the cosmological constant because he felt that the non-static universe predicted by the formalism of GR was incorrect, given the data available in 1917 [138]. At the time, observations of the universe were limited primarily to stars in our own galaxy, with observed low velocities, so there was solid observational evidence justifying the assumption that the universe was static. Einstein’s goal was to obtain a universe that satisfied Mach’s principle of the relativity of inertia. However, observational evidence started to accumulate for another paradigm. In 1917, Vesto Slipher [266] published his measurements of the spectra of spiral nebulae, which showed that most were shifted towards the red. The breakthrough came when the linear redshift-distance relation was formulated by Hubble [141], who showed that the universe was expanding. Einstein then dropped his support for the cosmological constant.

In the FLRW cosmological model, the expansion history of the universe is determined by the mass density of the different components, whose sum is normalised to unity:

$$\Omega_{m,0} + \Omega_{\text{rad},0} + \Omega_X,0 + \Omega_{k,0} = 1, \quad (81)$$

where the 0 subscript indicates the present epoch. We use the term $\Omega_X$ to show that this equation does not assume anything about the nature of the additional energy component (dark energy). In fact we could have used $\Omega_\Lambda$ or $\Omega_{\text{DE}}$ in the current Concordance Model.

Big bang nucleosynthesis and observations of large scale structure provide a good determination of the mass content of the universe, allowing $\Omega_m$ and $\Omega_{\text{rad}}$ to be fixed. However, observations in the 1980s and 1990s started to show inconsistencies with the cosmological model at the time, which was a matter-dominated, expanding universe with a present-e epoch Hubble constant of $H_0 \simeq 0.7 \text{ km s}^{-1} \text{ Mpc}$ and $\Omega_\Lambda = 0$ [267,268]. This was the so-called “age problem”, where the predicted age of the universe seemed to be younger than the age of the oldest stars. Angular-diameter distances to the last scattering surface at $z = 1100$ measured from the CMB are in fact 1.7 times smaller than those predicted by
an isotropic and homogeneous universe containing only pressureless matter (see [269]). Since the inflationary scenario, which by then was well established, predicts a flat $\Omega_{\text{total}} = 1$ universe, there was a problem with the cosmological model.

It was realised that one of the three assumptions of the cosmological model had to be wrong. Either the universe contains exotic matter with a negative pressure, or standard General Relativity is wrong, or the universe is not homogeneous and isotropic. The solution could also lie in some combination of the three. Most of the research since the late 1990s has followed the first approach, and the term “Concordance Model” refers to an FLRW universe, following General Relativistic cosmology, containing dark energy.

Within the FLRW framework, two main proposals were put forward: one was $\Lambda$CDM, in which there is a contribution to the energy density from a term similar to the cosmological constant (or the cosmological constant itself), and the other was $\nu +$CDM, where the missing mass came from massive neutrinos ($m_\nu \simeq 7 \text{eV}$) (e.g., [270,271]).

The first strong evidence of dark energy came in 1998 and 1999, when observations of the luminosities of type Ia supernovae indicated that the expansion of the universe is accelerating [272,273]. Concurrently, other observations constrained the neutrino mass to $m_\nu \ll 7 \text{eV}$, thus discounting the $\nu +$CDM model and confirming $\Lambda$CDM as the Concordance Model (e.g., [265,274]). It is not clear when the term “dark energy” was first used, but it seems to have been around 1998. The term is analogous to “Dark Matter”, which had been in use for some time [275].

Since then, numerous observations have confirmed cosmic acceleration, including supernovae, the cosmic microwave background, large-scale structure and baryon acoustic oscillations (see, e.g., [276–278]). The values of the present epoch matter and radiation components are well established:

$$\Omega_{m,0} = \frac{8\pi G \rho_{m,0}}{3H_0^2} \sim 0.3, \quad \Omega_{\text{rad},0} = \frac{8\pi G \rho_{\text{rad},0}}{3H_0^2} \sim 1 \times 10^{-4},$$

(82)

where $H_0$ is the present value of the Hubble parameter $H(a = 1)$.

The data also indicate that the universe is currently nearly spatially flat:

$$|\Omega_K| \ll 1.$$ 

(83)

This is normally taken to imply that the spatial curvature $K = 0$, since

$$\Omega_{K,0} = 0 \equiv \frac{-K}{a_0^2 H_0^2} \sim 0.$$ 

(84)

Thus it also justifies the inflationary paradigm. However, inflation only tells us that $\Omega_K \to 0$, so that the curvature may have had a nonzero value in the past. In the present universe, however, the distinction is negligible. In any case, Equation (81) implies that there has to be a nonzero $\Lambda$ (a constant term added to the Einstein equation) such that

$$\Omega_{\Lambda,0} \equiv \frac{\Lambda}{3H_0^2} \sim 0.7.$$ 

(85)

Inserting these values into the Friedmann equation leads to the dramatic conclusion that the expansion of the universe is accelerating:

$$\ddot{a}_0 = H_0^2 \left( \Omega_{\Lambda} - \frac{1}{2} \Omega_m - \Omega_{\text{rad}} \right) > 0,$$

(86)

where $a_0$ is the present value of the scale factor $a(t)$. 
Note that this conclusion only holds if the universe is homogeneous and isotropic (i.e., a Friedmann-Lemaître model). In such a universe, the distance to a given redshift $z$ and the time elapsed since that redshift are tightly related via the only free function, $a(t)$. If the universe is isotropic around us, but not homogeneous, that is, a non-Copernican Tolman-Bondi-Lemaître model [279], then this relation would be lost and present data might not imply acceleration. A Copernican model where this relation again breaks down is the inhomogeneous universe, where the acceleration can be produced via nonlinear averaging—the backreaction of inhomogeneities.

Dark energy is a fluid component whose equation of state is:

$$p_{\text{DE}} = w \rho_{\text{DE}}. \tag{87}$$

This is the equation of state in its most general form, since $w$ can be any function of redshift, scale factor or cosmic time, with the constraint that $w \leq 0$ (i.e., the fluid has a negative pressure). Assuming that $w = w(a)$, we have the following density-scale relation:

$$\rho_{\text{DE}}(a) = \rho_{\text{DE},0} e^{-\frac{3}{2} \int_1^a \left[ 1 + w(a') \right] d \ln a'}. \tag{88}$$

It can be seen that in the special case of a constant $w = -1$, the fluid equation implies that the density is constant.

If dark energy is a cosmological constant, it still leaves the question of its physical nature. Its observed value of $\Lambda \approx 3 \times 10^{-122} c^3 / \hbar G$ is so small that it is hard to interpret as the vacuum energy.

One possibility is that the final value of the cosmological constant is zero, and that cosmic acceleration is due to the potential energy of a scalar field, with some sort of mechanism to dynamically relax it to a small value. This notion leads to models of dark energy which invoke a slowly-rolling cosmological scalar field to source accelerated expansion, similar to cosmological inflation.

How can a scalar field drive cosmic acceleration? The action of a scalar field minimally coupled to Einstein gravity is

$$S = \int d^4x \sqrt{-g} \left( \frac{m_{\text{Pl}}^2 R}{2} - \frac{1}{2} (\partial \phi)^2 - V(\phi) \right). \tag{89}$$

The stress-energy tensor for the scalar field is given by

$$T^\phi_{\mu\nu} = \partial_\mu \phi \partial_\nu \phi - g_{\mu\nu} \left( \frac{1}{2} (\partial \phi)^2 + V(\phi) \right). \tag{90}$$

For a homogeneous field such that $\phi = \phi(t)$, a cosmological scalar acts like a perfect fluid with equation of state $w = P/\rho$ given by

$$w_\phi = \frac{\dot{\phi}^2 - V(\phi)}{\dot{\phi}^2 + V(\phi)}. \tag{91}$$

The observed expansion leads us to a value of $w_\phi \simeq -1$, which requires a very slowly-rolling field: $\dot{\phi}^2 \ll V(\phi)$. There has also been a lot of interest in constructing quintessence models which can produce an equation of state of the “phantom” type ($w_\phi < -1$).

The equation of state of dark energy that has the potential to distinguish between dark energy candidates. The most important distinction that can be made between different dark energy models is whether the energy density of this component is constant, filling space homogeneously, or whether it is some form of quintessence field whose energy density can vary in time and space. There is a multitude of alternative models, such as $f(R)$ [280–282], or Chameleon Models [283,284]. It is therefore useful to consider the redshift evolution of $w$, so that $w$ is an arbitrary function of redshift $z$. There are a number of different parameterisations of $w(z)$ ([285], and references therein).
The most common parameterisation is sometimes termed the Chevallier-Linder-Polarski (or CPL) parameterisation [285,286]:

$$w(z) = w_0 - \frac{dw}{da} (1 - a), \quad (92)$$

where the scale factor $a = 1/(1 + z)$. If we define

$$w_a = - \frac{dw}{a da \ln a}, \quad (93)$$

then the equation becomes

$$w(a) = w_0 + w_a (1 - a), \quad (94)$$

which is the most commonly used form.

It has been shown that this parameterisation is stable and robust over large redshift ranges [287]. A wide range of functional forms of $w(a)$ can be parameterised by the $w_0 - w_a$ combination. However, there are some dark energy models which it cannot reproduce (see [288–291]).

The problem with the dark energy paradigm, stated simply, is that the parameters are not constrained well enough to rule out certain models. We have fairly good bounds on the dark energy density, but the dark energy equation of state is still poorly constrained. Even for a constant $w$ model, corresponding to $\Lambda$CDM, the bounds are such that a time-varying $w(a)$ could mimic a constant $w$, thereby disguising underlying physics (see [292–294]).

11. The Evolution of Large-Scale Structure

After the epoch of matter-radiation equality, and before the onset of dark energy domination, the mass-energy content of the universe became dominated by matter. From an initially smooth background (as evidenced by CMB observations), structures have evolved to a scale of more than 100 Mpc, with the term “large scale structure” being used to refer to objects modelled on this scale. At this scale, the mass-energy inhomogeneities can be modelled as perturbations on a homogeneous and isotropic unperturbed background spacetime. Below this scale we observe galaxy clusters, individual galaxies, and stars. The model of structure formation must be accurate enough to provide an good description of the universe on a wide range of scales.

The standard model for the formation of structure assumes that at some early time there existed small fluctuations, which grew by gravitational instability. The origins of these fluctuations are unclear, but they are thought to arise from quantum fluctuations of the primordial universe, uncorrelated and with Gaussian amplitudes, which were then amplified during a later inflationary phase [185]. The assumption that the amplitudes of the relative density contrasts is much smaller than unity means that we can think of the primordial fluctuations as small perturbations on a homogeneous and isotropic background density. This ensures that we can describe them using linear theory.

Heuristically, the mechanism of structure formation can be understood in terms of gravitational self-collapse. Matter collapses gravitationally around initial mass overdensities. This increases the relative density of that region, causing further collapse of more matter, and amplifying the effect. The linear theory of structure formation needs to be relativistic, because the perturbations on any length scale are comparable or larger than the horizon size at sufficiently early times. The horizon size is defined as the distance $ct$ which light can travel in time $t$ since the Big Bang. Dissipative effects and pressure also affect structure formation, as explained below (for details of the theory, see [295,296]).

The relative density is the density $\rho$ at a particular point in space $x$ relative to the mean $\bar{\rho}$ at some time parameterised by the scale factor $a$, and can be expressed as a dimensionless density contrast:

$$\delta(x,a) = \frac{\rho(x,a) - \bar{\rho}(a)}{\bar{\rho}(a)}. \quad (95)$$
This quantity can be understood as the dimensionless density perturbation of some background matter distribution.

There are two types of density perturbations that can occur within a matter-radiation fluid. If the fluid could be compressed adiabatically in space, the perturbations have a constant matter-to-radiation ratio everywhere. Since the energy density of radiation is proportional to \( T^4 \), and the number density is proportional to \( T^3 \), the energy densities of radiation and matter are related by:

\[
\delta_{\text{rad}} = \frac{4}{3} \delta_m. \tag{96}
\]

Isocurvature perturbations occur when the entropy density is perturbed, but not the energy density. Since the total energy density remains constant, there is no change in the spatial curvature and

\[
\rho_{\text{rad}} \delta_{\text{rad}} = \rho_m \delta_m. \tag{97}
\]

Perturbations can occur at different scales, or ‘modes’. The latter term is used when the amount of perturbation on a particular scale is expressed using Fourier analysis. The Fourier transform pair of \( \delta(x) \) is:

\[
\hat{\delta}(k) = \int d^3 x \delta(x) e^{i k x};
\]

\[
\delta(x) = \frac{1}{(2\pi)^3} \hat{\delta}(k) e^{-i k x}, \tag{98}
\]

with each mode assumed to evolve independently. In the Einstein-de Sitter regime, linear adiabatic perturbations scale with time as follows:

\[
\delta \propto \begin{cases} 
    a(t)^2 & \text{(radiation domination)} \\
    a(t) & \text{(matter domination)} 
\end{cases} \tag{99}
\]

while isocurvature perturbations are initially constant and then decline:

\[
\delta \propto \begin{cases} 
    \text{constant} & \text{(radiation domination)} \\
    a(t)^{-1} & \text{(matter domination)} 
\end{cases}. \tag{100}
\]

In both cases, the overall shape of the spectrum of the perturbations over all modes is preserved, while the amplitude changes with time. The evolution described above is affected on small scales by a number of processes.

11.1. Evolution on Small Scales

During the radiation-dominated epoch the growth of certain modes is suppressed. This behaviour can be modelled in terms of the horizon scale \( \lambda_H(a) \), which is the distance \( c t \) that light could have travelled since the initial singularity (a comoving horizon size). A mode \( k \) is said to enter the horizon when \( \lambda = \lambda_H(a) \), where \( \lambda = (2\pi)/k \). If \( \lambda < \lambda_H(a_{\text{eq}}) \) then a mode enters the horizon during the radiation-dominated epoch. The time scale for collapse of matter during this epoch is larger than the typical expansion time scale (\( t \sim 1/H(a) \)) due to the relatively rapid expansion \( \rho_{\text{rad}} \propto a^{-4} \). The growth of these modes is therefore suppressed. After the epoch of matter-radiation equality (\( a = a_{\text{eq}} \)), these perturbations can then start to collapse gravitationally. We can define the suppression factor for a particular mode as the factor by which the amplitude is reduced had it not entered the horizon:

\[
f_{\text{sup}} = \left( \frac{a_{\text{enter}}}{a_{\text{eq}}} \right)^2 = \left( \frac{k_0}{k} \right)^2. \tag{101}
\]
where the mode evolves as $\propto a^2$ until it enters the horizon at $a_{\text{enter}}$ and is suppressed until $a_{\text{eq}}$, when its evolution resumes as $\propto a$. The second equality in the above equation comes from applying an Einstein-de Sitter approximation where $k_0 = 1/\lambda_H(a_{\text{eq}})$ (see [297]).

Pressure opposes gravitational collapse for modes with a wavelength less than the Jeans length [298], sometimes called the free-streaming scale, defined as

$$\lambda_J = c_s \sqrt{\frac{\pi}{G\rho}}. \quad (102)$$

During the radiation-dominated epoch, the sound speed $c_s = c/\sqrt{3}$ and the Jeans length is always close to the horizon size. The Jeans length then reaches a maximum at $a = a_{\text{eq}}$ and then begins to decrease as the sound speed declines. This means that on scales larger than the comoving horizon size, perturbations are only affected by gravity, and the spectrum starts to turn over at this point (where the effects of pressure begin to dominate). The comoving horizon size at $z_{\text{eq}}$ is given by:

$$R_0 r_H(z_{\text{eq}}) \approx 16.0 \Omega_m h^2 \text{Mpc}. \quad (103)$$

Another important scale occurs where photon diffusion erases perturbations in the matter-radiation fluid. This process is termed Silk damping [299]. The scale at which it occurs is characterised by the distance travelled by the photon in a random walk by the time of last scattering:

$$\lambda_S \approx 16.3 (1 + z)^{-5/4} (\Omega_b^2 \Omega_m h^6)^{-1/4} \text{Gpc}. \quad (104)$$

All of the effects mentioned above are particularly important where the behaviour of massive neutrinos is concerned. Heuristically we can understand the complexity of their behaviour by considering them as a component whose equation of state changes as the universe evolves. From a component which behaves like photons (since the particles have a very small mass and relativistic speeds), massive neutrinos lose energy and start behaving like baryonic matter [300,301].

### 11.2. Growth of Perturbations in the Presence of Dark Energy

All of the above effects were described in an Einstein-de Sitter universe. In a universe with a smooth non-clustering dark energy component below the horizon scale, the matter perturbation fields evolves according to:

$$\ddot{\delta} + 2H\dot{\delta} - (3/2)H^2 \Omega_m \delta = 0$$

$$\delta'' + (2 - q)a^{-1}\dot{\delta}' - (3/2)\Omega_m a^2 \delta = 0, \quad (105)$$

where a dot denotes a time derivative and a dash denotes a derivative with respect to $a$. The term $q$ is the deceleration parameter. This can be interpreted in the following way: the perturbations grow according to a source term which involves the amount of matter ($\Omega_m$) but the growth is suppressed by the friction term due to the expansion of the universe. The latter is also known as the Hubble drag.

If we define the growth as the ratio of the amplitude of a perturbation at a time $a$ to some initial amplitude, i.e.,

$$D(a) = \frac{\delta(a)}{\delta(a_{\text{initial}})} \quad (106)$$

the equation becomes, for a general dark energy scenario where $w = w(a)$ (see [302])

$$D'' + \frac{3}{2} \left( 1 - \frac{w(a)}{1 + X(a)} \right) \frac{D'(a)}{a} - \frac{3}{2} \left( \frac{X(a)}{1 + X(a)} \right) \frac{D}{a^2} = 0, \quad (107)$$
where

\[ X(a) = \frac{\Omega_m}{\Omega_{DE}} e^{-3} f'_a \ln d'w(a') \]  

(108)

is the ratio of the matter density to the dark energy density. For large \( X \) (i.e., \( \Omega_m \sim 1 \) where \( \Omega_{DE} \sim 1 - \Omega_m \)) we recover the matter-dominated behaviour (\( D \sim a \)). To parameterise deviations from this behaviour we define the “normalised growth” as \( G = D/a \). The evolution equation is then:

\[
G'' + \left[ \frac{7}{2} - 3 \left( \frac{w(a)}{1 + X(a)} \right) \right] \frac{G''}{a} + 3 \left( \frac{1 - w(a)}{1 + X(a)} \right) \frac{G}{a^2} = 0. 
\]

(109)

This equation allows us to physically interpret the effects of dark energy. In the presence of dark energy, the Hubble drag term is increased, so that growth is suppressed in a universe with an accelerating expansion. This is similar to the suppression due to radiation dominance.

11.3. The Power Spectrum of Matter

In an FLRW universe, the homogeneity and isotropy assumption means that any statistical properties must also be homogeneous and isotropic. The implication for the matter perturbation field is that its Fourier modes must be uncorrelated (due to homogeneity). Usually, we assume that the mode amplitudes are Gaussian. This assumption is well motivated since the theory for the seed fluctuations assumes that they have a quantum origin. Due to the central limit theorem, the sum of a sufficiently large number of mode amplitudes will tend towards a Gaussian distribution (see, e.g., [303,304]).

Such a field, with uncorrelated modes, and a Gaussian distribution of mode amplitudes is called a Gaussian random field, and can be entirely described by its two-point correlation function:

\[
\langle \delta(x) \delta^*(y) \rangle = C_{\delta\delta}(|x - y|).
\]

(110)

The angled brackets denote an ensemble average (an average over a multitude of realisations). The value of \( \delta \) at a given point in the universe will have a different value in each realisation, with a variance \( \langle \delta^2 \rangle \). Since we can only observe one realisation of our universe (in other words, at most only a finite region in this one universe), we apply the ergodic principle: The average over a sufficiently large volume is equal to the ensemble average.

In Fourier space, the correlation function can be written as:

\[
\langle \hat{\delta}(\mathbf{k}) \hat{\delta}^*(\mathbf{k}') \rangle = \int d^3x e^{i\mathbf{k}\cdot\mathbf{x}} \int d^3x' e^{-i\mathbf{k}'\cdot\mathbf{x}'} \langle \delta(x) \delta^*(x') \rangle.
\]

(111)

Replacing \( x' = x + y \), and substituting Equation (110), this can be written as:

\[
\langle \hat{\delta}(\mathbf{k}) \hat{\delta}^*(\mathbf{k}') \rangle = \int d^3x e^{i\mathbf{k}\cdot\mathbf{x}} \int d^3y e^{-i\mathbf{k}'\cdot(x+y)} C_{\delta\delta}(|y|)
\]

\[= (2\pi)^3 \delta_D(\mathbf{k} - \mathbf{k}') \int d^3y e^{-i\mathbf{k}'\cdot(y)} C_{\delta\delta}(|y|)\]

\[= (2\pi)^3 \delta_D(\mathbf{k} - \mathbf{k}') P_{\delta}(|\mathbf{k}|). \]

(112)  

(113)  

(114)

The power spectrum has been defined as the Fourier transform of the correlation function:

\[ P_{\delta}(|\mathbf{k}|) = \int d^3y e^{i\mathbf{k}\cdot\mathbf{y}} C_{\delta\delta}(|\mathbf{y}|). \]

(115)

The standard convention in cosmology is to abbreviate \( P_{\delta}(|\mathbf{k}|) \) to \( P(k) \), where \( k = |\mathbf{k}| \). The power spectrum can be expressed in dimensionless form as the variance per \( \ln k \), so that:

\[ \Delta^2(k) = \frac{k^3P(k)}{2\pi^2}. \]

(116)
11.3.1. Nonlinear Evolution

The power spectrum gives us the evolution of the initial matter density fluctuations. However, the linear evolution breaks down at small scales, when complex structures begin to form, and overdensities can no longer treated as perturbations on a smooth background. This is the nonlinear regime of gravitational evolution. The scale above which nonlinearities cannot be ignored is approximately set by \( \Delta(k_{\text{NL}}) \sim 1 \), which corresponds to \( k_{\text{NL}} \sim 0.2 \, h \, \text{Mpc}^{-1} \) in most cosmological models. The standard way to model nonlinear evolution is by using phenomenological fits based on \( N \)-body simulations.

One strategy is to build a model based on a stable clustering hypothesis [305–307], which assumes that the nonlinear collapsed objects form isolated, virialised systems that are decoupled from the expansion of the universe.

A different approach is used in the halo model [308]. Here, the density field is decomposed into individual clumps of matter with some density profile and varying mass. By using this model to calculate the number of clumps within a given volume, the galaxy halo profile can be calculated. This is the equivalent of the power spectrum for these matter halos. A functional relation between the linear power spectrum and this halo profile is then derived and calibrated using large \( N \)-body simulations. This relation is then used to calculate the nonlinear power spectrum, using a fitting formula, for instance (see, e.g., [309,310]).

Whichever approach is used, it must account for a range of small-scale physical processes, such as baryonic physics, stellar formation, galactic magnetic fields, and Dark Matter and neutrino properties, which have become more significant with the ability of future astrophysical experiments to probe the nonlinear regime with ever increasing precision (see, e.g., [311–317]).

The need for realistic models of the universe has come to the fore in recent years, due the massive improvement in the quality and volume of cosmological measurements. Most of the \( N \)-body simulations rely on a perturbative approach. They use an FLRW cosmological background (perfectly homogeneous and isotropic), and assume that any sub-horizon inhomogeneous structure of the universe will contribute to an average expansion on horizon-sized volumes driven by the horizon-averaged density. With the next generation of cosmological probes, these may not be accurate enough to model the universe realistically. There are several ongoing efforts to build fully relativistic, nonlinear, inhomogeneous and asymmetric models using numerical methods [318–320]. This is a significant contribution to the study of the backreaction effect and the question of the expansion rate of the universe.

11.3.2. The Primordial Perturbations

In the very early universe, the tiny initial perturbations are thought to have formed a Gaussian random field whose covariance function is diagonal and nearly scale-invariant. This form, known as the Harrison-Peebles-Zel’dovich spectrum, was assumed in most cases within the Standard Model, as it corresponds very closely to the observed power spectrum in the universe.

This type of spectrum was first proposed in the 1970s by Edward Robert Harrison [321], Yakov Zel’dovich [322], and Phillip James Edwin Peebles [323], who were working independently, as the spectrum for initial density fluctuations. This hypothesis was subsequently closely borne out by observations. The defining characteristic of a Harrison-Peebles-Zel’dovich spectrum is that it describes a fractal metric, where the degree of perturbation is the same on all scales (hence the term “scale-invariant”), so that \( P(k) \propto k \). If we assume scale invariance for the power spectrum on large scales, and combine this with Equation (101), this implies the following general shape for the matter power spectrum in the Einstein-de Sitter scenario:

\[
P(k) \propto \begin{cases} 
  k & \text{for } k \ll k_0 \\
  k^{-3} & \text{for } k \gg k_0.
\end{cases}
\]

(117)
The actual form of the spectrum depends in non-trivial ways on the parameters in the cosmological model, including the ‘slope’ of the initial power spectrum $n_s$, where $P(k) \propto k^{n_s}$. In a scale-invariant spectrum in the linear regime, the fiducial value of $n_s$ is taken to be 1.

Cosmological observations were consistent with scale invariance up until the early data releases from the Wilkinson Microwave Anisotropy Probe (WMAP) [214,324], but showed some tension. Subsequent observations cast further doubts on scale invariance [216]. In 2013, the Planck probe, mapping the anisotropies in the CMB, led to an important and conclusive result: it ruled out scale invariance at over 5σ. The primordial power spectrum was found to be scale-dependent, with $n_s = 0.9603 \pm 0.0073$ [221]. This was confirmed by the 2015 data release, which found that $n_s = 0.968 \pm 0.006$ [205]. This significant result is a powerful demonstration of the importance of multiple sources of data, joint observations from different probes, and the increasing reliance on complex statistical techniques for cosmological model selection [325–328].

Thus it can be seen that the current Concordance Model, consisting of $\Lambda$CDM with an inflationary epoch in the early universe, was built in stages over the last 100 years. It is the result of a process of accumulation of evidence and testing of competing models and theories. At each step, theory provides the basis for adjustments to the model, and observations from different probes provided the evidence (Figure 4). The current model should in no way be seen as “true”. It is merely the best model that fits all the data available so far. Future data may very well require an adjustment to the Concordance Model.

![Figure 4. How the Concordance Model of Cosmology was developed. Theories and observations motivated the development of cosmological models, which were adjusted as new observations challenged the older models.](image)

12. How Do We Test General Relativity?

The assumption of metric coupling given by Equation (5) has been tested accurately many times over the last 100 years, at scales from $10^{-4}$ m in laboratories [329,330], up to $10^{14}$ m in the Solar System [331]. These experiment test the implications of the metric coupling: the spacetime...
independence of non-gravitational constants, the isotropy of the coupling of all matter field to a unique metric tensor, the universality of free-fall, and gravitational redshift.

The standard way to test constraints on $S_{\text{gravity}}$ and therefore the dynamics of General Relativity is by using the parameterised post-Newtonian (PPN) formalism [65]. This assumes that gravity is described by a metric over all scales. The idea is to write the most general form that $g_{\mu
u}$ can take in the presence of matter, when considering correction of order $1/c^2$ with respect to the Newtonian limit. This method was first used in 1923 by Arthur Eddington [332]. In its simplest form, it provides us with two phenomenological parameters $\beta_{\text{PPN}}$ and $\gamma_{\text{PPN}}$ entering the Schwarzchild metric in isotropic coordinates:

$$g_{00} = -1 + \frac{2Gm}{r c^2} + 2\beta_{\text{PPN}} \left(\frac{2Gm}{r c^2}\right)^2$$
$$g_{ij} = \left(1 + 2\gamma_{\text{PPN}} \frac{2Gm}{r c^2}\right) \delta_{ij}. \quad (118)$$

According to General Relativity, $\beta_{\text{PPN}} = \gamma_{\text{PPN}} = 1$. The experimental constraints on these parameters are summarised in Table 3.

<table>
<thead>
<tr>
<th>Method</th>
<th>Constraint</th>
<th>Experiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shift of perihelion of Mercury</td>
<td>$</td>
<td>2\gamma_{\text{PPN}} - \beta_{\text{PPN}} - 1</td>
</tr>
<tr>
<td>Lunar laser ranging</td>
<td>$</td>
<td>4\beta_{\text{PPN}} - \gamma_{\text{PPN}} - 3</td>
</tr>
<tr>
<td>Very long baseline interferometry</td>
<td>$</td>
<td>\gamma_{\text{PPN}} - 1</td>
</tr>
<tr>
<td>Time-delay variation</td>
<td>$\gamma_{\text{PPN}} - 1 = (1.2 \pm 2.3) \times 10^{-5}$</td>
<td>Cassini spacecraft [67]</td>
</tr>
<tr>
<td>Planetary perihelion precessions</td>
<td>$\beta_{\text{PPN}} - 1 = (-2 \pm 3) \times 10^{-5}$</td>
<td>Solar System ephemerides to 2013 [337]</td>
</tr>
</tbody>
</table>

It should be noted that the PPN formalism assumes that there is no characteristic length scale for the gravitational interaction, and therefore it does not allow testing of finite-range effects. These too have been constrained to be very close their General Relativistic value of zero. The deviation of amplitude $\alpha$ from a Newton potential on a characteristic scale $\lambda$ is typically $\alpha < 10^{-2}$ on scales ranging from a few millimetres to Solar System size [329]. This implies no deviation from GR over more than 15 orders of magnitude in length scale.

The Solar System tests constrain Eddington’s two PPN parameters to a tiny region very close to 1. The formalism has been generalized to include eight additional phenomenological parameters [65] to describe any possible deviation from GR at the first post-Newtonian order. They have all been constrained to be very close to their General Relativistic values. The latest data use Solar System ephemerides, which include perihelion measurements for Mercury and the other planets, as well as lunar ephemerides. Independent teams of astronomers have estimated corrections to the standard first-order post-Newtonian General Relativistic formalism to be all statistically compatible with zero [337,338]. This leads us to conclude that General Relativity is the only theory consistent with Solar System experiments at the post-Newtonian order.

It should be noted that, so far, it is only the first-order post-Newtonian, static, Schwarzschild-like part of the spacetime metric that has been modelled and tested using Solar System dynamics. The first-order post-Newtonian gravitomagnetic or Lense-Thirring part of the spacetime metric has neither been modelled nor tested yet in the Solar System. However, it has recently been pointed out that this could be possible in the next few years by focusing on particular models (Sun and planets, planets and spacecraft, or planets and planets) [339]. This would open the field to the possibility of constraining the PPN parameters using a more complete model of General Relativistic effects.

What about larger or smaller scales? The size of the universe was around $10^{-35}$ m in the beginning, and the present size is around $10^{26}$ m. GR remains untested at these extreme scales.
But one fundamental assumption of GR is that it describes gravity at all scales (i.e., the theory assumes the analytical continuity of solutions).

The evolution of the universe itself therefore provides a useful test of General Relativity. The challenge lies in testing a wide range of potentials (weak and strong field regimes) over cosmological scales. The strong field regime is tested using compact objects. However, over cosmological scales, the kind of objects that would produce gravitational potentials approaching the strong field regime simply do not exist. One solution is to observe the evolution of the universe and check whether the evolution of large-scale structure corresponds to the predictions of General Relativity.

13. Cosmological Tests

General Relativity has been submitted to 100 years of Solar System tests, which it has passed with flying colours. The discoveries of the last two decades: cosmic acceleration, the scale dependence of the primordial power spectrum, and also the open question of Dark Matter, have made it clear that GR must also be tested at astrophysical and cosmological scales.

This presents a serious challenge. We derive our knowledge of the universe from measurements of distances and times, and statistical properties like the distribution of matter. The main difficulty in extending the Solar System tests to cosmological scales is that these measurements depends strongly on the construction of cosmological models.

These models depend on four hypotheses:

1. A theory of gravitational interactions.
2. A description of the matter in the universe and the non-gravitational interactions such as electromagnetic emissions.
3. A hypothesis on the symmetry.
4. A hypothesis on the topology, or the global structure of the universe.

Some of the hypotheses are hard to verify, and some have unverifiable implications. Assuming the symmetry of our solutions means that we also assume the laws of physics are the same throughout the universe, including its unobservable part outside the cosmological horizon (delimited by a radius of around 15.7 Gpc). This is a very strong assumption, but one that is unverifiable.

Any cosmological model needs all four hypotheses. The first two are the physical theories, but their equations cannot be solved without some kind of assumption on the symmetry of the solutions, given by the third hypothesis. The fourth hypothesis is then an assumption on the global properties of these solutions.

The simplest \( \Lambda \)CDM Concordance Model assumes that gravity is described by General Relativity (Hypothesis 1), and that the universe contains the particles and fields of the Standard Model of particle physics \([171]\), together with Dark Matter and a cosmological constant (Hypothesis 2). The Einstein equations require an effective stress-energy tensor averaged out on large scales, so the model requires an extra assumption on the averaging procedure. This averaging and the validity or otherwise of this assumption is the subject of research on the backreaction effect. The model assumes the Copernican Principle (Hypothesis 3), and the continuity of the solutions of Einstein equations across all spatial sections (Hypothesis 4).

The number of alternative cosmological models proposed, especially since the 1990s, is too numerous to list here. Many of them are impossible to rule out simply by fitting them to the data, because their predictions are so close to those of \( \Lambda \)CDM. It is easier to test the four hypotheses rather than trying to test the observables predicted by the models. As with any null test, a significant violation would indicate the need to modify one or more hypotheses.

As far as tests of GR (the theory of gravity in \( \Lambda \)CDM) are concerned, many experiments test more than one assumption. We list the main experiment types, and the assumptions they test, in Table 4.
Table 4. What we are actually testing. Experimental tests of General Relativity (GR) often probe more than one assumption, and isolating the effects is a challenge in itself.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Assumption Tested</th>
</tr>
</thead>
<tbody>
<tr>
<td>Solar System tests</td>
<td>Metric coupling</td>
</tr>
<tr>
<td>Quadrupolar shift of nuclear energy levels</td>
<td>Isotropy [340–342]</td>
</tr>
<tr>
<td>Lunar laser ranging and orbiting gyroscopes</td>
<td>Universality of freefall [343] and structure of metric</td>
</tr>
<tr>
<td>Space-borne clocks</td>
<td>Gravitational redshift [344]</td>
</tr>
<tr>
<td>Shift in perihelion of planets</td>
<td>Structure of metric [345]</td>
</tr>
<tr>
<td>Time invariance of physical constants</td>
<td>Metric coupling [346]</td>
</tr>
<tr>
<td>Detection of gravitational waves</td>
<td>Lorenz gauge condition and inhomogeneous wave equation</td>
</tr>
</tbody>
</table>

13.1. Testing the Description of Matter and Non-Gravitational Interactions

General Relativity can be tested by measuring the fundamental constants of the theory [346–349]. Any variation would require a modification of GR [350,351]. A local measurement of a fundamental constant, such as a determination of the fine-structure constant from the Oklo phenomenon [352–354], is actually a cosmological-scale measurement along the time dimension. Astrophysical probes such as 21-cm radiation [355] or the cosmic microwave background [356], can be used to test the constancy of the fine-structure constant $\alpha$, or to constrain simultaneous variations of $\alpha$ and Newton’s gravitational constant $G$ [357]. A non-constant $G$ would have serious implications both for Newtonian physics and for General Relativity. This motivates ongoing efforts to devise new methods to measure this quantity, and to push the limits of experimental accuracy in order to test the constancy of $G$ [358–362].

13.2. Testing the Assumption of Symmetry

The assumption of symmetry can be tested by checking for any deviations from isotropy. This requires statistical ensembles of data, so the best observables are the CMB, and, on a smaller scale, large-scale structure [363–365].

13.3. Testing the Gravitational Interactions

The gravitational interactions have been tested many times at laboratory scales and all the way up to Solar System scales. The challenge today is to test them at cosmological scales.

Tests of General Relativity may be placed in three broad categories: laboratory, astrophysical, and cosmological. Experiments cannot span all length scales, and so they cannot test all theories. We are forced to design experiments which can test alternative theories, or the effects of GR, at particular length scales (Table 5).

Laboratory tests probe effects from sub-millimetre scales to a few hundred metres. Galileo’s experiments on weights dropped from the Leaning Tower of Pisa [366] are one example of a laboratory-scale test. The torsion balance experiments of Cavendish and Eötvös [367], and their modern versions [368–371], are another.

Astrophysical tests probe gravity from Solar System scales all the way up to galactic cluster scales. They include experiments such as laser ranging off the Moon (lunar laser ranging) and several of the Earth’s artificial satellites [372–375], and proposals to extend laser ranging to other planets in the Solar System (planetary laser ranging) [376–380], radar astrometry on near-Earth objects [381], observations of the precession of the perihelion as well as higher-order effects for Mercury and other planets [338,382,383], observations of spacecraft in the Solar System [337], actual or proposed measurements of the acceleration of the Pioneer [384–386] and New Horizons probes [387–389], observations of the orbits of compact objects such as neutron stars [390–392], and measurements of galaxy rotation curves [393,394]. In the last decade, observations of extrasolar planetary systems have been proposed and used in order to test GR [395–404]. All of these systems are characterised by their spherical symmetry. Many can be approximated by a test object orbiting at a distance $r$ around a
The gravitational field of the more massive central object is then probed by observing the orbit of the test body.

Table 5. Bridging the length scales to test the cosmological model. Experimental tests of gravity and dark sector couplings, at their typical length scales. Massive gravity (MG) screening mechanisms would show up at short ranges, while smooth dark energy manifests itself at cosmological scales. We give the experimental accuracy from current and future experiments planned over the next decade. The comparison between growth and expansion history comes from combined BAO, supernova, weak lensing, redshift distortion, cosmic microwave background (CMB) lensing, and cluster data. Lensing effects and dynamical mass comparisons can be carried out over a range of scales: inside galaxies, using strong lensing and stellar velocities, and on cosmological scales, using cross-correlations. This is a route to testing screening effects in alternative theories. Laboratory and Solar System tests can also probe dark sector couplings besides short-range effects, but many of the constraints obtained depend on the cosmological model.

<table>
<thead>
<tr>
<th>Test</th>
<th>Length Scale</th>
<th>Theories Probed</th>
<th>Current Status (and Future)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Growth vs. expansion history</td>
<td>100 Mpc (\cdots) 1 Gpc</td>
<td>GR with smooth dark energy</td>
<td>10% accuracy (2%–4%)</td>
</tr>
<tr>
<td>Lensing vs. Dynamical mass</td>
<td>0.01 (\cdots) 100 Mpc</td>
<td>Test of GR</td>
<td>20% accuracy (5%)</td>
</tr>
<tr>
<td>Astrophysical tests</td>
<td>0.01 AU (\cdots) 1 Mpc</td>
<td>MG screening mechanisms</td>
<td>(~10%) (Up to 10 times improvement)</td>
</tr>
<tr>
<td>Laboratory and Solar System tests</td>
<td>1 mm (\cdots) 1 AU</td>
<td>PPN parameters in MG</td>
<td>Constraints are model dependent (Up to a tenfold improvement)</td>
</tr>
</tbody>
</table>

For such systems, the deviation of the metric from the Minkowski form is characterised by the magnitude of the Newtonian gravitational potential

\[
\epsilon = \frac{G M}{r c^2}.
\] (119)

The strongest gravitational fields accessible to an observer occur in the limit \(\epsilon \to O(1)\). In this limit, the central object is a black hole and the test object is close to the event horizon.

We know that the Riemann curvature tensor is an essential quantity in General Relativity. The approximate magnitude of this tensor is expressed by the Kretschmann scalar (the fully contracted Ricci scalar) for the Schwarzschild metric:

\[
\xi \propto \frac{G M}{r^2 c^2}.
\] (120)

Note that this expression is more complicated for rotating objects [405]. For the purposes of our description, however, we may use this simple expression, which we call the curvature.

The third broad category is the cosmological tests. First, the measurements are made on a statistical ensemble. The masses must be treated as power spectra. Second, the gravitational field assigned to each wavenumber \(k\) is very weak. The description of the system often depends on the cosmological model. At this scale, the description must take into account the background perturbations and the expansion of the universe.

The position of the various systems on a potential-curvature parameter space is shown in Figure 5. The potential accessible to observers is bounded by the line \(\epsilon \simeq 1\). There is no limit to the maximum curvature that can be observed except for the Planck limit, where \(r \simeq 1.6 \times 10^{-33}\) cm. This lies many orders of magnitude above the boundaries of the figure. General Relativity is a complete theory, and does not fail below the Planck scale. To test or even probe GR below the Planck scale is a different matter, and a goal that has not yet been achieved. Many alternative theories of gravity that attempt to explain cosmic acceleration do so through modifications to GR at the Planck scale, so it is important to explore ways in which GR can be tested at these scales.
Figure 5. The parameter space for quantifying the strength of a gravitational field. The horizontal axis measures the potential. The vertical axis measures the spacetime curvature of the gravitational field at a radius $r$ away from a central object of mass $M$. Regions of this parameter space with potential greater than 1 represent distances from a gravitating object that are smaller than the event horizon radius and are therefore inaccessible to observers. The red vertical line on the right-hand side of the plot marks the horizon limit. This is a schematic plot, and in no way do we show an exhaustive list of objects and systems that have been used or could be used to test GR. The region of Solar System-scale tests is broadly bounded by the Moon, Gravity probe B [406,407], Mercury, and the Pioneer and New Horizons spacecraft. We have included the Voyager spacecraft in the diagram, even though, unlike the Pioneer probe, it was never suitable for tests of GR. The famous Hulse-Taylor binary pulsar [408], although a neutron star binary, is also roughly in the region of Solar System tests. Black holes and neutron stars are in the strong field regime, and the former are at at the limit of the event horizon boundary. Adapted from [409].

What is the minimum curvature? The unperturbed FLRW metric is isotropic, and the unperturbed Kretschmann scalar is a function of time only. The curvature for the homogeneous universe drops as the universe expands. The present universe has a curvature which is just above the boundary of the region marked ‘Dark energy’, since dark energy is not yet completely dominant over pressureless matter. However, the curvature will approach this limit asymptotically. In this paradigm, this represents a fundamental minimum curvature scale. It is shown in the figure by the region labelled “Dark energy”.

Galaxies are astrophysical probes of GR. Their innermost regions can be modelled as test particles orbiting a central supermassive black hole. Galactic rotation curves, which can only be explained by introducing Dark Matter, describe galaxy velocities up to the outermost regions. Systems below a constant acceleration scale of approximately $1.2 \times 10^{-10}$ ms$^{-2}$ cannot be modelled without adding a contribution to the gravitational field in the form of Dark Matter. This constant acceleration is the diagonal boundary of the region labelled “Dark matter” in the figure.
Note that the regions of parameter space occupied by Dark Matter and dark energy overlap—there is a degeneracy between these two unknown components of the cosmological model. However, it is not impossible to distinguish between their effects, since their properties are different. In particular, Dark Matter forms clumps just like baryonic matter, while dark energy does not.

GR has not been tested in the region between curvatures of $\sim 10^{-40}$ and $\sim 10^{-50}$. This corresponds to the region between Solar System scales and cosmological scales probed by galaxy surveys and the CMB. The challenge lies in finding systems which span these scales. In theory they do exist, in the form of galaxies, clusters and superclusters. Their rotation curves transition from Schwarzchild orbits in their innermost regions, to outer regions dominated by Dark Matter. However, our observations are hampered by the fact that we are limited by the resolution of our telescopes, so we can only observe the outer regions. In addition, the untested region is situated between a region where GR is extremely well-constrained (by Solar System tests), and a region where Dark Matter and dark energy have to be invoked, and where we must take the cosmological model into account, because the effect of large-scale structure on background dynamics becomes non-negligible. It has been suggested that this backreaction effect may be at the origin of the observed cosmic acceleration.

Cosmological tests of General Relativity may be broadly classified as follows:

- Tests of the consistency between the expansion history and the growth of structure. A discrepancy in the equation of state parameter of dark energy $w$, inferred from the two approaches can indicate a breakdown of the GR-based smooth dark energy cosmological paradigm.
- Detailed measurements of the linear growth factor across different scales and redshifts.
- Comparison of the cosmological mass distribution inferred from different probes, especially redshift space distortions and lensing.

14. Possible Modifications of GR and Cosmological Implications

It is useful to identify the regimes in which modifications to GR may appear. This enables us to get a clearer picture of the capabilities and limitations of current and future experiments to tests these alternative theories.

14.1. Weak and Strong-Field Regimes

In order to test gravity in the strong-field regime, we need to observe compact objects with a very high density [410–412]. Black holes are good candidates. Their compactness and mass takes $GM/rc^2$ close to the maximum limit of unity. However, they have a serious drawback. Because of the “no hair theorem”, they are not characterised by any coupling to a scalar field and therefore cannot be used to test for this effect, and to discriminate between scalar-tensor theories and GR.

Neutron stars, on the other hand, are still very compact bodies, but they can be strongly coupled to a hypothetical scalar field. This property has been used to test relativistic parameters by observing the Hulse-Taylor binary pulsar PSR B1913+16 [413], and the neutron star–white dwarf binary PSR J1141-6545 [414]. Gravitational time delay has been tested using other binary pulsars [415–419], and pulsars have also been suggested as ideal probes of the Lense-Thirring effect [420,421]. General Relativity passes these tests with flying colours.

14.2. Small and Large Distances

Distance-dependent modifications can be induced by a massive degree of freedom, which will cause a Yukawa-like coupling. General Relativity is very well constrained on the size of the Solar System, and there are several tests constraining Yukawa interactions at Solar System scales [422–424], but there are no constraints on scales larger than $10^3 h\ Mpc^{-1}$, at least without assuming some cosmological model. Some theories put forward to explain cosmic acceleration, such as Chameleon mechanisms [425], are essentially modifications of GR at cosmological distance scales.
14.3. Low and High Accelerations

Galaxy rotation curves and galaxy dynamics motivated the Dark Matter paradigm. The Tully-Fisher law \cite{426} tells us that Dark Matter cannot be explained by a modification of General Relativity at a fixed distance. MOND instead explains it by modifying gravity at low accelerations below the typical acceleration $a_0 \sim 10^{-8} \text{ cm} \cdot \text{s}^{-2}$.

14.4. Low and High Curvature

Curvature $R$ is important in distinguishing possible extensions of the Einstein-Hilbert action. A curvature-dependent may become important even if the potential $\Phi$ remains small. In the Solar System, the curvature $R_\odot \sim 4 \times 10^{-28} \text{ cm}^{-2}$. The curvature of the homogeneous universe according to the Friedmann equation is

$$R_{\text{FLRW}}(z) = 3H_0^2 \left( \Omega_m (1 + z)^3 + 4\Omega_\Lambda \right),$$

from which we can see that the curvature of the universe evolved with time, from $\sim 10^{-33} \text{ cm}^{-2}$ at the time of nucleosynthesis, to $\sim 10^{-56} \text{ cm}^{-2}$ at $z = 1$. The curvature scale associated with the cosmological constant is $R_\Lambda = (1/6)\Lambda$, so the phenomenology of the cosmological constant occurs in low curvature regime

$$R < R_\Lambda \sim 1.2 \times 10^{-30}R_\odot.$$  

This is of particular interest to paradigms which seek to explain cosmic acceleration through the backreaction effect. For cosmological-scale perturbations, we are always in the weak field regime. However, the curvature perturbation associated with large-scale structure becomes of the order of the background curvature at redshift $z \sim 0$, even if we are still in the weak field limit. This means that the effect of large-scale structure on the background dynamics may be non-negligible \cite{427}.

In summary, in order to explain the dark energy or Dark Matter problem by modifying General Relativity, the modifications have to be either at large scales (typically Hubble scales), low accelerations (typically below $a_0$, or small curvatures (typically $R_\Lambda$). The regions corresponding to dark energy and Dark Matter curvature and potentials are shown in Figure 5.

14.5. Cosmological Probes

The idea of testing General Relativity using large-scale structure was first proposed in 2001 \cite{428}. This relies on the ingenious idea that if gravity is well-described by General Relativity, and the universe well-described by $\Lambda$CDM, then on sub-Hubble scales, and considering only scalar perturbations, the spacetime metric can be written as

$$ds^2 = -(1 + 2\Phi) \, dt^2 + (1 - 2\Psi) a^2(t) \gamma_{ij} \, dx^i \, dx^j$$

where $\Phi$ and $\Psi$ are the two potentials, and $\gamma_{ij}$ is the metric of the spatial section \cite{429–431}. The Einstein equations reduce to the Poisson equation

$$\triangle \Psi = 4\pi G \rho_{\text{matter}} a^2 \delta_{\text{matter}}$$

and

$$\Psi - \Phi = 0,$$

since the matter anisotropic stress is negligible. The spectrum of the two gravitational potentials has to be proportional to the matter power spectrum. The scale dependence of the gravitational potential $P_\Phi$ and of the matter distribution $P(k)$ are related by:

$$k^4 P_\Phi(k, a) = \frac{9}{4} \Omega_m H_0^2 a^{-2} P(k, a).$$
If the Poisson equation is modified by some modification of gravity, the matter power spectrum changes shape according to the cosmological model assumed, so the above relation is not model-independent. However, the fact that the two spectra differ is independent of the cosmological model. It therefore provides a test of the underlying gravitational theory. Such a test can be carried out by comparing weak lensing data to galaxy surveys.

Various similar approaches have been proposed [432]. This approach allows us to test various classes of alternative theories by means of large-scale structure. Some of these theories include Dvali-Gabadadze-Porrati models [134,433,434], quintessence [435–439], and scalar-tensor theories [440–442]. The difficulty lies in finding a parameterisation of the perturbation equations which is consistent with the one used for the background evolution, since both assume the same theory for gravitational interactions.

The cosmic microwave background is another potential testing ground for General Relativity. The amplitude and position of the peaks of the cosmic microwave background allows us to probe the potential wells present during the recombination era by extracting information on the primordial power spectrum created during inflation. Planck data, alone or in combination with weak lensing has been used to test GR [444,445] and modified gravity [446–448].

Despite the fact that GR is extremely well tested on laboratory and Solar System scales, cosmology provides plenty of scope for alternatives to General Relativity. Let us consider cosmic acceleration, which is now a confirmed observation. It requires an explanation. The cosmological constant paradigm rests on three assumptions: that the observations are correct, that GR is the correct theory of gravity, and that our FLRW model of the universe is correct.

More than two decades of precision measurements have removed any doubt that the observed acceleration may be due to incorrect modelling of experimental errors. So in order to do away with the cosmological constant, we must seek possible answers in the other two assumptions.

Backreaction is an “alternative” to the alternative theories, or to the Dark Matter paradigm. It is firmly within the General Relativity paradigm, and seeks to explain cosmic acceleration by modelling the universe and the structures within it, and therefore its expansion history, in more detail than is the case with the FLRW model. It keeps the first and last assumption, but does away with the assumption of an FLRW universe. However, is it enough to explain the observed acceleration?

If GR is assumed to be the correct theory together with the other two assumptions, cosmological measurements are usually interpreted as providing evidence for Dark Matter and a nonzero cosmological constant or dark energy. This poses conceptual problems. Why is the observed value of the cosmological constant so small in Planck units? It also poses a coincidence problem. Why is the energy density of the cosmological constant so close to the present matter density? No dynamical solution of the cosmological constant problem is possible within GR—the cosmological constant is not the attractor of some dynamical function.

This opens the field to possible modifications of GR. Should GR be modified at low and high energies? This is a serious challenge for theorists. Einstein’s theory is the unique interacting theory of a Lorentz-invariant massless spin-2 particle. New physics in the gravitational sector must introduce additional degrees of freedom. These additional degrees of freedom must modify the theory at low or high energies, or both, while being consistent with GR in the intermediate-energy regime, that is, at length scales $1\mu \lesssim \ell \lesssim 10^{11}$ m, where the theory is extremely well tested.

Figure 6 illustrates some of the difficulties in testing the completeness of General Relativity. There is an obvious “scale desert” between Solar System scales, and cosmological scales. At the other end of the scale, there is a gap which is often overlooked: between sub-millimetres scales, which is the current limit where GR has been tested, and the Planck scale. A modification of gravity at very small scales would be apparent in this regime. Even ignoring quantum effects, it is a serious challenge to test gravity in a regime in which vastly stronger forces come into play.
Figure 6. The parameter space for experiments. The horizontal axis is the typical length scale of the object in question. The vertical axis measures the gravitational potential. The red vertical line on the left-hand side marks the Planck scale. The vertical line on right-hand side of the plot marks radius of the observable universe, or Hubble radius. Experimental verification of GR is impossible beyond these limits. The radius of the surface of last scattering is only slightly smaller than our Hubble radius. Assuming GR implies assumptions far beyond the range that has been experimentally tests. For instance, if we define the Planck mass as $m_{\text{Planck}} = \sqrt{\frac{\hbar c}{G}}$, we are assuming that the gravitational constant remains constant down to the Planck length. This extrapolates the inverse square law over a scale of more than $10^{-30}$ from what has been tested. The green region is where Solar System tests have been carried out. Beyond $\sim 100$ Mpc, assumption of a homogeneous and isotropic metric becomes accurate enough to use in physical models. The blue region shows length scales at which the FLRW metric is valid. By way of comparison to this parameter space, the nonlinear regime for perturbation theory, which gives use the matter power spectrum, covers $\sim 10^{22}$ to $\sim 10^{23}$ m, the linear, quasi-static regime covers $\sim 10^{23}$ to $\sim 10^{25}$ m, and beyond that is the superhorizon regime. These length scales ranging from 1 Mpc to above 1 Gpc fit on a tiny part of the horizontal axis above, shown by the thick blue horizontal line.

15. The Nature of Dark Energy and the Implications for General Relativity

The Concordance Model of cosmology assumes that General Relativity is correct, an assumption which is justified by the tests which GR has undergone. Within this model, $\sim 95\%$ of the content of the universe is unaccounted for. Dark matter, which makes up around 25% of the mass-energy of the universe is a matter-like component which is cold (sub-relativistic) and weakly interacting. The discrepancy between the observed acceleration of the expansion of the universe and the predictions of GR leads to the conclusion that there must be a cosmological component with a negative equation of state parameter making up around 70% of the mass-energy content of the universe: dark energy.

But the dark energy paradigm does not fix the nature of this component. There exist many theories which attempt to explain its nature. In particular, one can ask whether the basic assumptions of the Concordance Model—homogeneity and isotropy—are correct.
If the universe is not homogeneous and isotropic, the FLRW equations are no longer valid. Over the last decade, one line of research has attempted to explain the accelerated expansion by exploring the implications of an inhomogeneous universe on a General Relativistic cosmology. This effect is the backreaction. This approach does away with dark energy.

There are various ongoing investigations on the effect of the backreaction due to an inhomogeneous universe [269,427,449–455], with different lines of research offering different interpretations of the Buchert equations, where the Friedmann equations are supplemented by an additional backreaction term [456]. Whether one can explain all of the observed expansion history of the universe as a consequence of the growth of inhomogeneities without invoking some additional fluid component is the subject of ongoing debate [452,457]. The backreaction, even if it turns out to be incapable of replacing the dark energy paradigm, is still a subject worth investigating, if anything as a correction to the homogeneity assumption, which obviously breaks down at small scales.

If we assume that the Copernican Principle holds, then the universe is well described by a Friedmann-Robertson-Walker spacetime. The dynamics of the background expansion are determined by the content of the universe: the list of fluids (perfect fluids, due to the assumption of the Copernican Principle), with their equations of state. Within the dark energy paradigm, we can distinguish two main strategies for formulating hypotheses:

1. there is some new kind of component in the universe, or
2. there is some new property of gravity.

Let us first recall that General Relativity rests on two assumptions: the gravitational interaction is described by a massless spin-2 field, and matter is minimally coupled to the metric, which implies the weak equivalence principle. The Einstein-Hilbert action described by Equation (6) implies that

$$S_{\text{gravity}} = \frac{c^3}{16\pi G} \int R \sqrt{-g} \, d^4x + S_{\text{matter}}[\text{matter}; g_{\mu \nu}],$$

(127)

where $R$ is the Ricci scalar of the metric tensor $g_{\mu \nu}$, and $S_{\text{matter}}[\text{matter}; g_{\mu \nu}]$ is the action of the matter fields. If we only consider field theories, this gives us a useful classification scheme for the different theories that seek to explain the nature of dark energy.

The first strategy listed above assumes that gravitation is described by General Relativity, and introduces new forms of gravitating components beyond the Standard Model of particle physics. This adds a new term $S_{DE}[\psi; g_{\mu \nu}]$ to the action in Equation (127) while keeping the Einstein-Hilbert action of all the standard and Dark Matter unchanged.

The second strategy modifies gravity, and therefore extends the action, either by modifying the Einstein-Hilbert action of the coupling of matter. These theories also involve new forms of matter.

A cosmological constant $\Lambda$ is the simplest modification which can be made to gravity, and it is equivalent to dark energy with a constant equation of state. To explain the observed acceleration, the new form of matter must have an equation of state $w < -1/3$. Dark energy can also be attributed to the energy of the vacuum, although the energy predicted by the Standard Model of particle physics is either 0 (using super-symmetry), or $10^{120}$ orders of magnitude larger than the observed cosmological value [458]. There are ongoing attempts to solve this ‘fine-tuning problem’ using string theory [459–461], causal sets [462,463], or by using anthropic arguments [464,465].

The other approach is to attribute dark energy to a scalar field whose potential has evolved in some way that it currently exerts a negative pressure. Such fields, in theories within the framework of GR, are termed Quintessence. Their distinguishing feature is that they allow the equation of state of dark energy to evolve. Alternatives to Quintessence within the same approach include K-essence, Phantom Fields, or the Chaplygin Gas [466–469].

Another strategy is to depart from General Relativity and modify the laws of gravity and posit dark energy as the manifestation of an effect arising from extra dimensions, or higher-order corrections. Within this category, the more successful theories have been of two types. Dvali-Gabadadze-Porrati
(DGP) dark energy considers the universe as a 4D brane within a 5D Minkowskian bulk [134]. The weakness of gravity relative to the other forces is explained by gravity “leaking” into the higher dimensions as it acts through the bulk [470], whereas the other forces act within the brane. The other class of theories is \( f(R) \) gravity, where the Ricci scalar \( R \) in the Lagrangian is replaced by some function \( f(R) \). Such theories correspond to scalar-tensor gravity with vanishing Brans-Dicke parameter [471].

We can therefore identify four main classes of theories, as shown in Figure 7. Classes 1 and 2 assume GR and introduce new forms of gravitating matter, the difference being that in class 2, the distance-duality relationship may be violated due to mechanisms such as photon decay. Classes 3 and 4 modify gravity in some way. In class 3, a new field introduces a long-range force so that gravity is no longer described by a massless spin-2 graviton. In this class, there may be a variation of the fundamental constants. Class 4 includes theories in which there may exist an infinite number of new degrees of freedom, such as brane models or multidimensional models. These models predict a violation of the Poisson equation on large scales. The constraints which may be placed on these theories, and the experiments to test them, correspond to the regimes in which gravity is modified, as summarised in Section 14.

An exhaustive list of dark energy or cosmic acceleration theories is beyond the scope of this review. We simply note that most of them give different predictions for the equation of state of dark energy, its evolution, or the expansion history of the universe. To distinguish between these proposals we need to track the evolution of these parameters throughout the history of the universe.

![Figure 7](image)

Figure 7. The four main classes of dark energy theories, within the two broad strategies, classified as modifications of the General Relativistic action. Classes 1 and 2 assume the gravitational metric coupling of GR, whereas classes 3 and 4 modify this metric coupling, and are therefore modifications of gravity. In the upper line of classes (1 and 3), new fields dominate the matter content of the recent universe. Adapted from [431].

16. The Current Status of General Relativity

General Relativity has been subjected to a multitude of tests in its 100 years of existence. As of 2016, the main predictions of GR have been tested and confirmed. Whereas it is sufficient for most purposes at ordinary accelerations and energy scales to use Newtonian calculations, General Relativity has found its way into daily life, in the Global Positioning System and geodesy. The postulates of General Relativity have been confirmed with ever-increasing accuracy. Deviations from the Einstein Equivalence Principle are now constrained to below \( \sim 10^{-14} \), deviations from Local Lorentz Invariance down to \( \sim 10^{-20} \), and deviations from Local Position Invariance down to \( \sim 10^{-6} \).

General Relativity has been probed down to scales of \( \sim 10^{-6} \) metres in laboratories, and up to 1000 AU in space-based experiments and observations. Astrophysical and cosmological observations...
have probed GR at scales from 1 Mpc up to gigaparsec scales. In the Solar System, the dynamics of GR have been tested with radar and laser-ranging. We can track the ephemerides of the planets in the Solar System, right up to the minor outlying planets (such as Sedna, with an aphelion just short of 1000 AU). The farthest objects whose trajectory has been followed from the moment they were “thrown” are now outside the Solar System. They are the Pioneer 10 and 11 spacecraft, and are now around 70 AU distant, and Voyager 1, which is at a distance of 135 AU, making it the only object to have reached interstellar space. Communication with the Pioneer spacecraft ceased in 2003. The Pioneer spacecraft exhibited an anomalous constant acceleration towards the Sun which could not be explained using GR. This prompted a reexamination of all the recorded data. It is now generally accepted that the anomaly is caused by thermal radiation, and that once this is accounted for, there is no remaining anomalous acceleration [472].

Gravitational waves, among the last untested predictions of GR, were first detected in late 2015, ushering a whole new era of observational astrophysics, in which the strong field regime can be probed and where the predictions of GR can be tested.

The evidence for General Relativity is extremely strong. The theory has passed all tests in the weak-field limit at Solar System scales, including tests of the assumptions (the Equivalence Principle) and the predictions specific to GR (frame-dragging, gravitational time dilation), and in the strong field with the observation of a binary black hole merger and the resulting gravitational waves. We give a summary of the experimental milestones in Table 6. However, there are still issues that allow room for speculation.

The first is the question of the completeness of GR. Is it valid at all scales? There is a scale gap in our tests of GR between laboratory, Solar System and galactic scales, and cosmological ones. This gives rise to a multitude of domains of validity for different alternative theories, whereas ideally we should seek a universal theory that can explain phenomena at all scales.

The second issue is the accuracy of our approximations. General Relativity may be conceptually simple in that it is based on a minimum number of postulates. But the resulting field equations, when applied to real physical systems, can be very hard to solve. We get around this by making approximations such as spherical symmetry, or homogeneous and isotropic perfect fluids, which allows us to obtain analytical solutions. However, the accuracy of these approximations may not always be good enough. This is evident in the case of cosmological perturbations and large-scale structure. At which scales is it valid to use a Friedmann-Lemaître-Robertson-Walker metric? Can the backreaction explain some or all of the observed cosmic acceleration? For smaller systems such as aspherical collapsing bodies, we still need accurate models in order to match theory and observation. Such questions have spurred the development of numerical methods in General Relativity. Physicists now have the necessary computing power to go far beyond simple first-order approximations.

The remaining open questions are of a cosmological nature: Dark Matter and dark energy, and inflation. Dark matter and dark energy account for around 25% and 70% of the mass-energy content of the universe, respectively. They are not a problem for the theory of gravity itself, but it does mean that we do not know the nature of 95% of the content of the universe. Inflation solves a number of cosmological problems, but whatever theory we choose to explain inflation, we still need to introduce new physics beyond GR.
### Table 6. The major predictions of GR, and their experimental status.

<table>
<thead>
<tr>
<th>Effect</th>
<th>Milestones</th>
<th>Current, Future and Proposed Experiments</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gravitational waves</td>
<td>LIGO [80]</td>
<td>Advanced LIGO [500] eLISA <a href="https://www.elisascience.org">https://www.elisascience.org</a></td>
<td>Recently confirmed in two events [80,501]</td>
</tr>
</tbody>
</table>
17. Future Developments

The current Concordance Model of cosmology was built in successive (and sometimes concurrent) steps. General Relativity applied to a spacetime under the Copernican Principle, filled with pressureless matter, produced the Einstein-de Sitter model. Motivated by the observed Hubble expansion, it resulted in the Big Bang model. Clues from element abundances, baryon asymmetry, and knowledge of nucleosynthesis from the Standard Model of particle physics meant that the universe had to have a thermal history, which resulted in the Hot Big Bang Model. When evidence for missing mass became incontrovertible, cold Dark Matter had to be added to the inventory of cosmic components. This model worked well, but not well enough. It could not explain the observed homogeneity of the universe across regions which were causally disconnected, nor its flatness. So Inflation was introduced. Observations of an accelerated cosmic expansion motivated a search for explanations within the then current paradigm, which resulted in various hypotheses: a curved geometry, supermassive neutrinos, or perhaps a particular cosmological topology. In the end, the paradigm had to be shifted yet again with the introduction of dark energy.

The Concordance Model can explain the observations with just six parameters: the physical baryon density parameter $\Omega_b h^2$, where $h$ is the Hubble parameter, the physical Dark Matter density parameter $\Omega_c h^2$, the age of the universe $t_0$, the scalar spectral index $n_s$, the curvature fluctuation amplitude $\Delta^2$, and the reionisation optical depth $\tau$. That such a degree of fit is offered by such a simple model is remarkable.

The success of the Concordance Model has been its ability to include physical effects at extremely different scales, from primordial nucleosynthesis to large-scale structure evolution, in one coherent theory. However, this does not allow us to state that the $\Lambda$CDM model is correct. It merely implies that deviations from $\Lambda$CDM are too small compared to the current observational uncertainties to be inferred from cosmological data alone. This leaves room for some very fundamental open questions, which we have described in this review.

The science of cosmology finds itself at a critical point where it has to make sense of the vast quantity of data that has become available. Different probes have allowed us to piece together interlocking information which, so far, confirms the Concordance Model. The cosmic microwave background has provided conclusive evidence of a flat geometry, super-horizon features, the correct harmonic peaks, adiabatic fluctuations, Gaussian random fields, and most recently, a departure from scale invariance. We have not yet observed primordial inflationary gravitational waves. Large-scale structure observations, which probe the recent universe, provide firm evidence in favour of the Concordance Model’s explanation of the evolution of density perturbations and the growth of structure, and provide a bridge between the effects of long-range gravitational interactions and shorter-range forces.

The recent detection of gravitational waves in two events (possibly three), one hundred years after they were predicted by Einstein, directly validates General Relativity in several ways. It shows that GR is correct in the strong-field regime, that black holes really exist, and black hole binaries too, and it proves that gravitational waves are a real physical phenomenon, and not just a mathematical artefact of GR.

We are fast approaching the point where cosmological observations will be limited only by cosmic variance, i.e., no more data will be available from our Hubble patch [508]. What does that imply for the development of new theories? How do we test the predictivity of these theories without new data?

We will have to look for new effects in old data, and for new correlations in future data. Large-scale structure is extremely useful in testing General Relativity and cosmological models, but the future may bring other observational windows. In particular, the next fifty to one hundred years may see the development of a gravitational wave astronomy. This is likely to be an even more significant development than even CMB astronomy. The observation of primordial gravitational waves would provide vital information on the inflationary epoch [509,510]. B-mode polarisation of the CMB offers an indirect pathway to the observation of this gravitational wave
background [511,512]. The second major development could be the observation of a cosmic neutrino background [513], which is the result of neutrino decoupling in the lepton era. This would push the observations along our past light cone even further back in redshift, providing information on the universe before recombination and the CMB.

17.1. Plausible Conclusions from Incomplete Information

The statistical questions facing cosmologists pose some particular problems. We observe a finite region of our universe, which is itself a single realisation of the cosmological theory. We can only observe whatever is on or inside our past light cone, as shown in Figure 1. Not only are we limited by cosmic variance, we also have just a single data point for the cosmological model.

There exist several alternatives to GR and to \( \Lambda \)CDM that have not been ruled out by experiment. Constructing viable physical models is not just a question of fitting the model to the data. It is a question of model selection, which requires robust statistical techniques that allow us to make sensible decisions using our incomplete information. Bayesian inference provides a quantitative framework for plausible conclusions [514,515]. We can identify three levels of Bayesian inference.

1. Parameter inference (estimation). We assume that a model \( M \) is true, and we select a prior for the parameters \( \theta \), or the \( \text{Prob}(\theta|M) \).
2. Model comparison. There are several possible models \( M_i \). We find the relative plausibility of each in the light of the data \( D \), that is we calculate the ratio \( \text{Prob}(D|M_i)/\text{Prob}(D|M_0) \).
3. Model averaging. There is no clear evidence for a best model. We find the inference on the parameters which accounts for the model uncertainty.

At the first level of Bayesian inference, we can estimate the allowed parameter values of the theory. If we assume General Relativity as our theory, we still need to fix the values of the various constants. This is the rationale behind ongoing efforts to measure quantities like Newton’s constant \( G \) more accurately, despite the fact that the parameter has been around for three centuries. What are the energy densities of the various components of a \( \Lambda \)CDM universe? If we include a dark energy equation of state parameter \( w \), what value does it take?

Next, we can ask which parameters we should include in the theory. Should we include a cosmological constant in General Relativity? Or should we include dynamical dark energy parameter? Although current data are consistent with the six-parameter \( \Lambda \)CDM model based on GR, there are more than twenty candidate parameters which might be required by future data (see [516]). We cannot simply include all possible parameters to fit the data, since each one will give rise to degeneracies that weaken constraints on other parameters, including the \( \Lambda \)CDM parameter set (e.g., [517–519]).

The landscape of alternative cosmological models is even larger if we relax our assumptions on the theory of gravity [520]. The goal in data analysis is usually to decide which parameters need to be included in order to explain the data. For physicists, those extra parameters must be physically motivated. That is, we need to know the physical effects to which our data are sensitive, so that we can relate these effects to physics. At the current state of knowledge, we have to acknowledge the possibility of more than one model. We therefore require a consistent method to discard or include parameters. This is the second level of Bayesian inference—model selection.

Bayesian model selection penalises models which introduce wasted parameter space. We can always construct a theory that fits the data perfectly, even better than GR, but we would need to introduce extra free parameters (e.g., extra fields or couplings between matter and the metric). This is the mathematical equivalent of Occam’s razor. We seek a balance between goodness of fit (the degree of complexity) and predictive power (consistency with prior knowledge). General Relativity fits all the data with the minimum number of parameters. In cosmology, \( \Lambda \)CDM is the best model because it only involves one new parameter and no new fields.

The problem of model selection in relation to GR is as old as the theory itself. But only recently have cosmologists have started to use Bayesian methods for cosmological model
selection (e.g., [521–524]), when the astrophysical data began to have the necessary statistical power to enable model testing. Bayesian techniques are starting to be applied to General Relativity itself [525]. With the next generation of astrophysical probes in the pipeline, model selection is likely to grow in importance [514].

Bayesian model selection cannot be completely free of assumptions. In cosmology, there is some model structure which depends on a number of unverifiable hypotheses about the nature of the universe. The Copernican Principle is one such hypothesis [9,526,527].

The third level of Bayesian inference is model averaging. In the current scenario, there is firm evidence for General Relativity as the best model for gravitational interactions. However, it is still useful to quantify our degree of certainty (or doubt), for one simple reason: we do not have the final list of alternative theories. In other words, when we choose GR against any number of alternative theories, we have no knowledge about other alternatives outside that list (such as theories yet to be developed, for example). At best, we know which alternatives are ruled out by the data. This level of Bayesian inference is the application of a principle that has been called ‘Cromwell’s Rule’ [528]: even if all the data show our theory to be correct, we should allow a non-zero probability, even if tiny, that the theory is false.

The utility of alternative theories becomes evident when we apply Bayesian model selection. For science to advance by falsification, it is not enough to claim that the present theory is false. We need to know which alternative theory is favoured instead. Newtonian gravity would have likely have survived the 1919 eclipse if General Relativity had not been formulated.

17.2. Experimental Progress

There has been rapid progress in constraining cosmological parameters and models over the last two decades, with a multitude of experiments observing the CMB, large-scale structure, galaxies and supernovae. We will just provide a summary of the most recent data sets.

The first are anisotropies in the CMB, where the main statistic is the angular power spectrum of fluctuations $C_\ell$, and polarisation of the CMB. The most recent and current are: WMAP (http://map.gsfc.nasa.gov), Planck (http://www.cosmos.esa.int/web/planck), Atacama Cosmology Telescope (ACT) (http://act.princeton.edu), South Pole Telescope (SPT) (https://pole.usc.edu), Atacama Cosmology Telescope polarisation-sensitive receiver (ACTPol) [529], SPTPol, Spider (http://spider.princeton.edu). Polarbear (http://bolo.berkeley.edu/polarbear), Background Imaging of Cosmic Extragalactic Polarization (BICEP2) (http://bicepkeck.org,https://www.cfa.harvard.edu/CMB/bicep2), Keck Array (http://bicepkeck.org,https://www.cfa.harvard.edu/CMB/keckarray).


The third source of data are weak lensing, which use the fact that images of distant galaxies are distorted and correlated by intervening gravitational potential wells to produce statistics such as the convergence power spectrum $C_\kappa$ [530,531]. Some current experiments are: Dark Energy Survey (DES) (https://www.darkenergysurvey.org), Red Cluster Sequence Lensing Survey (RCSLenS) (http://www.rcslens.org), Canada-France Hawaii Telescope Lensing Survey (CFHTLenS) (http://www.cfhtlens.org),

The final source of data are catalogues of peculiar velocities. By measuring redshifts and radial distances of galaxies and clusters it is possible to reconstruct a radially projected map of large-scale motions. Progress in this field will come from all three data sets above.

The main science goal of the next generation of cosmological probes is to test the Concordance Model of cosmology. Some major experiments will be operational in the next decade.

Planck, decommissioned in 2013, marked a major milestone in CMB experiments. Its proposed successor is the ground-based programme CMB-S4 [532], which should reach sensitivities below $10^{-3} \, \mu \text{K}$ whose main aim is to achieve higher resolutions, probe larger scales, and measure new observables such as polarisation.

There are various future experiments to map the mass-energy content of the universe (including baryons and Dark Matter), either in the planning phase, or close to completion. The Large Synoptic Space Telescope (LSST) (https://www.lsst.org), which should achieve first light in 2019, is a ground-based telescope which will map the entire sky.

The Euclid space telescope (http://www.euclid-ec.org), due for launch in 2020, will map galaxies and large-scale structure over the whole sky at visible and near-infrared wavelengths, providing a catalogue of 12 billion sources at 50 million. Euclid will probe the recent universe, when galaxies have formed and dark energy starts to dominate. Its main scientific objective is to understand the origin of the accelerated expansion of the universe by probing the nature of dark energy using weak-lensing observables (which include cosmic shear, higher-order distortions, and cosmic magnification), and galaxy-clustering observations [225].

The Square Kilometre Array (SKA) (https://www.skatelescope.org), which will begin operations in 2020, is a multi-wavelength radio telescope, built across multiple sites to achieve the largest collecting area ever [70,533]. It should provide the highest resolution images of the radio sky, thus providing maps of large-scale structure, and observing pulsars which should provide direct tests of General Relativity. The SKA will observe the epoch between the emission of the CMB and the formation of the first galaxies. Neutral hydrogen surveys (or 21 cm intensity mapping) [534], offer yet another promising probe, as do galaxy redshift surveys [535].

Joint observables will be key in the next generation of experiments. These include the Sachs-Wolfe effect, and the Sunyaev-Zel’dovich effect [536]. We will also need to extract more observables from the CMB. We need information on in order to constrain inflationary models.

As experiments probe larger scales at better resolutions (low $\ell$ and high $\ell$), the data analysis and the cosmological tests (dark energy, Dark Matter, the properties of cosmic neutrinos) will require accurate calculations on the growth of large-scale structure, which can only be achieved using $N$-body simulations [537–544].

At the other end of the length scale, there is particle physics. The Very Large Hadron Collider, the successor to the Large Hadron Collider (https://home.cern/topics/large-hadron-collider) (which can achieve energies around 14 TeV), is still in its conceptual phase. It could, if built, probe energies around 100 TeV, allowing it to test physics beyond the Standard Model, possibly including Supersymmetry and Grand Unified Theories, and thus provide clues on the nature of Dark Matter and dark energy (see, e.g., [545]).

Tremendous progress is also being made in Milky Way astrophysics. With the launch on 2013 of the Gaia mission (http://www.cosmos.esa.int/web/gaia), we will soon have improved data on Solar System ephemerides, and on the orbits and tidal streams in the Milky Way. This will allow precise tests of GR at Galactic scales. In particular, tidal streams provide an opportunity to probe GR at late cosmological times and to close the gap between astronomical and cosmological scales.

Now that gravitational waves have been detected, gravitational wave physics is set to become an established branch of astrophysics. The Laser Interferometer Gravitational-Wave Observatory (LIGO) (https://www.ligo.caltech.edu), Advanced LIGO (aLigo) (https://www.advancedligo.mit.edu)
and Virgo (http://www.virgo-gw.eu) will subject GR to a battery of test in the strong regime. Also targeting the regime of strong curvatures and potentials, the planned Event Horizon Telescope (http://www.eventhorizontelescope.org) is a network of millimetre and sub-millimetres telescopes being used for very-long baseline interferometry to directly image supermassive black holes in galactic centres.

17.3. Theoretical and Computational Progress

On the theoretical front, there are four main lines of development in theories of gravity. They are all motivated by current open questions in physics. The ongoing attempt to find a Grand Unified Theory continues to motivate the development of string theory [546] as a framework for gravity. Aside from the theoretical difficulties of a mathematically complex theory, the challenge for string theory is to produce physical predictions which can be experimentally tested. The second approach is brane theories or supergravity, in which spacetime has more than four dimensions. The resulting field theory combines Supersymmetry (from particle physics) and General Relativity [547]. The third approach is quantum gravity [548], in which spacetime, as a dynamical field, is a quantum object. This implies a violation of Lorentz Invariance near the Planck scale, which in turn means that some particle decays forbidden by Special Relativity are allowed, and possibly charge-parity-time violations too. This motivates the search for signatures of quantum gravity in particle physics experiments.

The final approach, which is closer to the theoretical framework of the Concordance Model of cosmology, is a phenomenological use of General Relativity. This includes the various alternative theories that seek to explain cosmic acceleration and the missing mass, and also the various theories which provide dark energy and Dark Matter candidates.

The early alternatives to General Relativity were motivated by theoretical considerations. The current alternatives are mainly motivated by the open questions in cosmology. Cosmology is now in the age of Big Data. In the last decade, the data finally caught up with the theory, and we are now in a position to test many of the current alternative theories using statistical techniques.

The principles that underpin statistical techniques are simple enough. Under minimal assumptions about signal and noise, it is simply a question of maximising the Gaussian likelihoods. However, in practice they are extremely complex. The foreground contamination, the signal, the possibly non-Gaussian noise and the systematics all have to be modelled. This requires data simulation, so the number of maps that is generated is far larger than the number that is actually observed by the instrument (see, e.g., [549–554]).

A low signal-to-noise ratio, and higher-resolution observations of fainter signals over a larger frequency range, result in massively big data sets. For statistical probes such as the CMB and large-scale structure, these data sets have to be analysed as a whole, in order to correlate data. In addition to the volume of data collected, this generates a huge amount of computational data, and requires the appropriate computing power to carry out multiple complex calculations. We are witnessing a Moore’s Law in cosmology, where the data volume of experiments increases by a factor of around 1000 every ten years, as shown in Figure 8.

In addition to the data volume challenge, we have an algorithmic challenge. The number of computations for cosmological data analysis depends on four main factors: the number of observations $N_i$, the number of pixels $N_p$, the number of multipoles (a function of the resolution and frequency range) $N_\ell$, and the number of iterations $N_i$. The first three quantities determine the data volume. The data simulation scales at least as $O(N_i)$. The map-making scales as $O(N_i N_\ell \log N_i)$. The maximum-likelihood power spectrum estimation scales as $O(N_i N_\ell N_p^3)$. 
Figure 8. The growing data volume of experiments. The data volume of each experiment is shown as an order-of-magnitude multiple of the data volume of the 1965 Bells Labs experiment which detected the CMB. The labels show the year of ‘first light’ for each experiment. CMB surveys are marked by blue dots, while red dots show large-scale structure surveys. A first light date for CMBPol (light blue triangle) has not yet been fixed. Note that the vertical axis is logarithmic: the date volume increases about a thousandfold every ten years (grey dashed line). Note too that we plot CMB, large-scale structure, space and ground-based probes on the same graph. Ground-based probes will always tend to have a larger data volume than space-borne probes, due to the bandwidth limit on data transmission from spacecraft. Longer-running experiments will also have a larger data volume.

Planck marked a milestone in data science. It was the first CMB experiment in which the whole data treatment process was parallelised, and where Monte Carlo methods were used in order to cut down on the number of data realisation iterations that were carried out.

There are two main considerations in data analysis: the amount of data, and the complexity of the theory. Future experiments will require sophisticated techniques, and considerable computing power to process the vast amounts of data. Gaia is already collecting data [555], while Euclid, the SKA, Enhanced LIGO, and CMB-S4 will soon be operational. Farther ahead, the Evolved Laser Interferometer Space Antenna (eLISA) (https://www.elisascience.org), due for launch in 2034, will also generate huge volumes of data, and will require the necessary computing power to test GR directly, by comparing the data to simulated black hole collapse, inspiralling binaries.

What direction will experimental cosmology take over the coming decade? Extracting fainter signals, such as CMB polarisation, or going to very high resolutions requires larger data volumes to provide a higher signal-to-noise ratio, and it requires more complex models to control fainter systematic effects. Even the Solar System tests of General Relativity will depend upon vastly greater data volumes and computational complexity to get the full relativistic ephemerides, given the ever-increasing number of objects being tracked in the Solar System and the significantly greater precision of the data for each object.
This enables us to make a sensible prediction on future developments. The science we are able to extract from present and future data sets will be determined by the limits on our computational capability, and our ability and willingness to exploit it.

17.4. Conclusions

GR may well survive for another 100 years. After all, Newtonian gravity was around for 200 years. GR has just reached its peak, when data and computing power have caught up with the theory. We are at a pivotal moment in the history of GR. We are on the point of confirming beyond reasonable doubt all its predictions throughout its entire domain of validity.

We have seen how modern cosmology is faced with big questions which touch the very foundations of physics. What is this form of matter which interacts only with gravity and apparently with nothing else? Why is the expansion of the universe accelerating? What caused the universe to undergo a period of rapid expansion soon after the Big Bang? These questions, motivated by cosmological observations, lead to questions about fundamental physics. Are there forces and interactions besides the four we know of, that is, gravity, electromagnetism, and the strong and weak nuclear forces? Are there particles beyond the Standard Model? What determines the value of the fundamental constants of nature? What is the real structure of spacetime? Are there extra dimensions?

Science needs data, so each of these questions must be addressed through careful experiment. The challenge of modern experimental physics is to probe nature at extreme distances and energies, well outside the capabilities of the instruments that were available to Einstein. It has certainly come a long way, as shown by the detection of gravitational waves in 2015, a feat which was thought to be impossible by many of Einstein’s contemporaries.

General Relativity is not the final theory of gravity, for there is no such thing. As General Relativity turns 100, we would do well to celebrate it with a healthy does of scientific scepticism. Long live General Relativity, and a big welcome to its eventual replacement, whether in our lifetime or not.

Acknowledgments: George F. Smoot acknowledges support through his Chaire d’Excellence Université Sorbonne Paris Cité and the financial support of the UnivEarthS Labex programme at Université Sorbonne Paris Cité (ANR-10-LABX-0023 and ANR-11-IDEX-0005-02). I.D. acknowledges that the research work disclosed in this publication is partially funded by the REACH HIGH Scholars Programme—Post-Doctoral Grants. The grant is part-financed by the European Union, Operational Programme II—Cohesion Policy 2014–2020.

Author Contributions: All the authors conceived the idea and contributed equally.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this work:

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>AU</td>
<td>Astronomical Unit</td>
</tr>
<tr>
<td>CDM</td>
<td>Cold Dark Matter</td>
</tr>
<tr>
<td>CMB</td>
<td>Cosmic Microwave Background</td>
</tr>
<tr>
<td>EEP</td>
<td>Einstein Equivalence Principle</td>
</tr>
<tr>
<td>ECKS</td>
<td>Einstein-Cartan-Kibble-Sciama</td>
</tr>
<tr>
<td>eV</td>
<td>electronvolt</td>
</tr>
<tr>
<td>FLRW</td>
<td>Friedmann-Lemaître-Robertson-Walker</td>
</tr>
<tr>
<td>GR</td>
<td>General Relativity</td>
</tr>
<tr>
<td>Gy</td>
<td>Gigayear ($10^9$ years)</td>
</tr>
<tr>
<td>ΛCDM</td>
<td>Cold Dark Matter</td>
</tr>
<tr>
<td>LLI</td>
<td>Local Lorentz invariance</td>
</tr>
<tr>
<td>LLR</td>
<td>Lunar laser ranging</td>
</tr>
<tr>
<td>LPI</td>
<td>Local position invariance</td>
</tr>
<tr>
<td>Mpc</td>
<td>Megaparsec</td>
</tr>
<tr>
<td>PPN</td>
<td>Parameterised post-Newtonian</td>
</tr>
<tr>
<td>SR</td>
<td>Special Relativity</td>
</tr>
<tr>
<td>TeV</td>
<td>Teraelectronvolts ($10^{12}$ electronvolts)</td>
</tr>
</tbody>
</table>
References

7. Popper, K.R. Logik der Forschung; Mohr Siebeck: Tübingen, Germany, 1934.
23. Lagrange, L. Méchanique Analytique; Chez La Veuve Desaint: Paris, France, 1788.
64. Dicke, R.H. Mach’s principle and invariance under transformation of units. Phys. Rev. 1962, 125, 2163–2167.
78. Hertz, H. Untersuchungen über die Ausbreitung der Elektrischen Kraft; J.A. Barth: Leipzig, Germany, 1894.
84. Weyl, H. Raum-Zeit-Materie; Springer: Berlin, Germany, 1918.


484. Iorio, L. Constraining the preferred-frame $a_1$, $a_2$ parameters from solar system planetary precessions. *Int. J. Mod. Phys. D* **2014**, *23*, 1450006.

Universe 2016, 2, 23


