JIM BAGGOTT

ORIGINS

THE SCIENTIFIC STORY OF CREATION

ORIGINS

ORIGINS

THE SCIENTIFIC STORY OF CREATION

JIM BAGGOTT

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries

First published in 2015
First published in paperback in 2017
Revised impression 2018

Impression: 1

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America

British Library Cataloguing in Publication Data
Data available

Library of Congress Control Number: 2017964288

ISBN 978–0–19–882600–2

Printed in Great Britain by
Bell & Bain Ltd., Glasgow

For Grace (aged 7),
who wants to be a planetary geologist

CONTENTS

ABOUT THE AUTHOR

Jim Baggott is an award-winning science writer. A former academic scientist, he now works as an independent business consultant but maintains a broad interest in science, philosophy, and history and continues to write on these subjects in his spare time. His previous books have been widely acclaimed and include:

Mass: The Quest to Understand Matter from Greek Atoms to Quantum Fields (Oxford University Press, 2017);

Farewell to Reality: How Fairy-tale Physics Betrays the Search for Scientific Truth (Constable, 2013);

Higgs: The Invention and Discovery of the 'God Particle' (Oxford University Press, 2012);

The Quantum Story: A History in 40 Moments (Oxford University Press, 2011);

Atomic: The First War of Physics and the Secret History of the Atom Bomb 1939–49 (Icon Books, 2009), short-listed for the Duke of Westminster Medal for Military Literature, 2010;

A Beginner's Guide to Reality (Penguin, 2005);

Beyond Measure: Modern Physics, Philosophy, and the Meaning of Quantum Theory (Oxford University Press, 2004);

Perfect Symmetry: The Accidental Discovery of Buckminsterfullerene (Oxford University Press, 1994);

The Meaning of Quantum Theory: A Guide for Students of Chemistry and Physics (Oxford University Press, 1992).

ACKNOWLEDGEMENTS

This has been an extraordinarily ambitious project, and I could not have contemplated it without the support of Latha Menon, my editor at Oxford University Press. I remain forever in her debt. Thanks must also go to Emma Ma, Jenny Nugee, and Kate Gilks at OUP, and Frances Topp and Paul Beversey, who have been instrumental in helping to turn my words and pictures into the book you now hold in your hands. I doubt that the book would be what it is were it not for John Gribbin, whose book *Genesis*, published in 1981, was an early source of inspiration (well, I remember cursing when I read it, because I realized that this was a book I would very much have liked to have written).

I'm never happier than when I'm learning something new. Writing *Origins* has taken me into some familiar territory, about which I've written before, but its wide-ranging scope has also drawn me into some less familiar spaces. I will therefore be eternally grateful to all the academic experts who have given of their valuable time to read through sections of the manuscript and offer helpful advice and criticism. These include Jennifer Clack at Cambridge University; George Ellis at the University of Cape Town; Andrew Knoll at Harvard University; Jeremiah Ostriker at Princeton University; Michael Russell at the Jet Propulsion Laboratory in Pasadena, California; Chris Stringer at London's Natural History Museum; Ian Tattersall at the American Museum of Natural History in New York City; Steven Weinberg at the University of Texas at Austin; and Simon White at the Max Planck Institute for Astrophysics in Garching, Germany. Any errors or misconceptions that remain are, of course, down to me.

Now, you know well enough that happiness doesn't come only from learning things and writing about them. It also derives from a great sense of place and the company of good people. Thanks then to Madeleine and Jur and Guido and Mikael, good people who have created two such fabulous places where some of these words were written, and to my wife Jini for just about everything else.

Jim Baggott
June 2015

LIST OF FIGURES

PREFACE

What is the nature of the material world? How does it work? What is the universe and how was it formed? What is life? Where do we come from and how did we evolve? How and why do we think? What does it mean to be human? How do we know?

These questions, many others just like them, and many subsidiary questions that logically follow are, collectively, the 'big questions' of human existence. They are questions that we have been asking ourselves for as long as we have been capable of rational thought.

We weave what answers we can discover or contrive into a creation story, and such stories have formed an essential foundation for every human culture throughout history. We have a seemingly innate and rather insatiable desire to want to comprehend our own place in the universe, to understand how we and everything around us came to be. This is a desire driven in part by simple curiosity. But I suspect it is also driven by a deeper emotional need to connect ourselves meaningfully with the world which we call home.

There are many different versions of our creation story. This book tells the version according to modern science. It is the result of an enterprise which has involved (and continues to involve) thousands of scientists struggling to piece together parts of the puzzle, constantly speculating, hypothesizing, testing, debating, and revising. These scientists strive to ensure that the puzzle pieces are individually coherent and consistent. But if the story is to be unified and comprehensible, the pieces themselves must also fit together, from the large-scale grandeur of the universe to *Homo sapiens* to the smallest microorganism to the elementary particles from which all material substance is composed. This is a powerful constraint.

That said, there is no such thing as an 'authorized' or 'official' version of the scientific story of creation. But I'm modestly hopeful that if there were, then it *might* look something like the book you now hold in your hands. Without wishing to appear overly melodramatic, I believe all my efforts as a popular science author over more than twenty years have been building up to this. Deep down, this is a book I have always wanted to write.

Now, I'm sure that's all very well, but is this a book you will want to read? This is obviously something only you can judge, but here are a few things you might want to think about.

First, I think the time is right for a book like *Origins*. In the past few years there has been a glut of popular physics titles telling us about new theories of everything or arguing that we live in a universe which is but one of a multiplicity of universes. In truth, although this stuff is advertised as science, none of it is accepted outside of a relatively small community of theorists, and it actually explains nothing relevant to our story. I have written elsewhere about such unsubstantiated 'fairy-tale' physics and I believe (or, at least, hope) that readers are growing increasingly wary of it.*

But what do we do when we are assaulted by news headlines screaming of another dramatic breakthrough in our understanding of our origins, only for the conclusions to be retracted months later when it emerges that the analysis was faulty and the announcement premature? In these circumstances, it's all too easy to lose sight of what's regarded as accepted scientific fact. And then what do we do when scientists publish books arguing in favour of their pet theories—theories that perhaps very few of their colleagues in the scientific community buy into? It's difficult to know what to make of these. Should we believe them?

In *Origins* I've tried to distinguish unquestioned fact from majority explanation from debatable interpretation from pure speculation. This is a book for readers who want a reasonably clear, balanced, and (hopefully) unbiased perspective on what we think we know and can explain. Yes, there are gaps and

* See my book *Farewell to Reality*, published in 2013.

things we don't understand clearly, and there are places in the story where scientists have had no choice but to indulge their inner metaphysician. I've been sure to point these out.

Second, anyone interested in tracing the scientific account of creation from the big bang to the transition to behavioural modernity in humans will need to range over many different scientific frontiers. These include (deep breath) aspects of modern cosmology and particle physics; the primordial synthesis of hydrogen and helium; the formation of stars and galaxies; stellar evolution and nucleosynthesis; planetary formation and differentiation; the chemistry of life; the evolution of the genetic code and simple, single-celled organisms, complex cells, and multicellular organisms; the sequence of mass extinctions and radiations that profoundly shaped the evolution of marine and land animals; the rise (and fall) of the dinosaurs; the emergence of mammals, primates, early hominins, the genus *Homo*, the species *Homo sapiens*, and the evolution of human consciousness.

Now that's asking a lot, and there's only so much time in a day. There are many popular, accessible, and highly readable books on all these subjects but *Origins* is, I think, unique in attempting to pull the contemporary scientific story together in a single volume.

Telling this story will take us on a journey from the very 'beginning' of the universe to the origin and evolution of human consciousness, 13,820 million years later. Whatever your reason for joining me, you should be aware of the path we will take. I'm going to try to explain what makes the scientific story of creation so different, and why I believe it is more reliable and compelling and ultimately more satisfying than other versions you might have heard of.

I want to reassure you that this is a satisfaction born not of smug certainty, from some sense of scientific triumphalism or authoritarianism. Far from it. I recognize that, while answers are very interesting, it is often the unanswered *questions* that fascinate. The satisfaction comes from acknowledging that although we know (or think we know) much, we don't yet know everything. And the answers we do have are likely to be modified or replaced entirely as we discover more things about the world and about ourselves and our place in it.

Some may view this as a weakness, but over a lifetime I've learned to be very wary of people who lay claim to certain knowledge. Perhaps we can be reasonably certain of one thing. Just ten years from now the story will be different, in some subtle and some not-so-subtle ways. And I guarantee you that the new version will be *better* than the old.

THIS THING CALLED SCIENCE

The physical chemist and novelist C. P. Snow famously wrote about the 'two cultures'—essentially science and the humanities—in an article published in the *New Statesman* magazine in October 1956, just a few months before I was born. Alas, this schism is in some ways wider now than ever before. There remains a persistent perception that science is not about 'us', that it somehow lacks human empathy. Science, so the argument goes, involves the application of a methodology derived from a rather cold, inhuman logic. It is obsessed with materialist mechanisms which demean our human spirituality. Science, the argument continues, can tell us some things about the physical, chemical, and biological mechanics of 'how?' but it can't address our deeper, equally compelling questions concerning 'why?'

I personally believe that *Origins* is very firmly about 'us'. This is *our* story. It is about how the world on which we live came to be, how life began and evolved to produce us: conscious, intelligent beings capable of a scientific investigation of their beginnings.

Science works because when we apply it we impose on ourselves a fairly rigorous discipline. We demand things from ourselves that are often counter-intuitive and counter-cultural and which require considerable intellectual effort. Obviously, we impose exacting standards in the elucidation and reporting of scientific facts. We demand scientific theories that broadly fit these facts, which hopefully make predictions that are potentially accessible to observation and experiment and which provide genuine insight and understanding (although there are lots of grey areas).

But science also demands that we adopt an attitude or perspective which is generally referred to as the 'Copernican Principle'. Although scientists are fundamentally interested in 'us', they presume that we are not particularly *special*. When we step outside the sphere of our distinctly human preferences and prejudices, as *Origins* will recount, we do indeed discover that we are not uniquely privileged observers of the universe we inhabit. To all appearances, the universe is not designed with us in mind. As we will see, science tells us that we are very much a naturally evolved *part* of this reality, but we are not the *reason* for it.

In other words, the universe has no *purpose,* at least if we seek to interpret purpose in a specifically human context. If teleology concerns the search for purpose or evidence of design in nature, then by definition science is firmly anti-teleology. This, in my view, doesn't make science inhuman. What it means is that it is not possible for humans to apply the scientific method without first adopting more than a little *humility*.

And herein lies the rub. Many of our 'why?' questions are typically driven by our singularly human search for purpose and meaning in life. It should therefore come as no surprise to find that science will struggle to answer them if, indeed, they have answers at all. Science is simply not set up this way and, if needed, we must reach for other belief systems to provide such answers as we can find.

Does this make science any less valid? No, it does not. Even though it cannot address many of our 'why?' questions, the simple truth is that we do not know its limits. One of the most remarkable aspects of modern science is its relatively new-found capacity to provide answers to questions that not so long ago might have been considered the preserve of high priests, priests who historically have had a habit of conflating 'why?' with 'how?'.

There is much we can still learn.

THE STORY IN OUTLINE (SPOILER ALERT!)

Origins tells the scientific story of creation in a chronological sequence, beginning with the origin of the universe—of space, time, and energy—in a 'big bang'. The first three chapters trace the origin and expansion of the universe through to something called the moment of recombination, about 380 000 years later, which releases the flood of hot electromagnetic radiation we now identify with the cold cosmic background. By this stage in the history of the universe, the basic building blocks are all available—space, time, energy, matter (dark matter and hydrogen and helium atoms), and light.

Much of our understanding of this early stage in the history of creation is derived from the fusion of so-called inflationary cosmological models and something called the standard model of particle physics. For the earliest stages in this history, we are obliged to reach for educated guesses and extrapolate theories to energy regimes well beyond their domain of applicability or validity. There's a lot we really don't understand about the very earliest stages in the evolution of the universe.

Nevertheless, the discovery of something that looks very much like the standard model's Higgs boson, at CERN in July 2012, suggests that our ignorance of the real state of affairs in the history of the universe might be limited to the first trillionth of a second of its existence. Now, I would respectfully suggest, that's not so bad.

In choosing to tell the story chronologically, I've had to accept some challenges. It means that the opening chapters involve us in some fairly heavy and demanding physics and, since much of the observational evidence that shapes our understanding of the early universe is derived from events that occur later in its history, there are places in the story where I have to ask you temporarily to suspend your demand for the evidence.

Chapter 4 continues the story with the formation of the first stars and galaxies, some time between 300 and 550 million years after the big bang. It describes the fundamentally important roles played by tiny inhomogeneities in the distribution of matter, thought to be imprinted on the large-scale structure of the universe by inflation, and the mysterious dark matter believed to account for almost 27% of the total mass-energy of the universe today. Chapters 4 and 5 detail what we currently understand about the evolution of stars and of stellar nucleosynthesis, which produces a range of chemical elements from hydrogen to iron. Cataclysmic supernova explosions are required to explain the existence of elements heavier than iron.

The sprinkling of a broad variety of chemical elements in the dust and vapour ejected from exploding stars leads to the production of interstellar molecules, many of which are now recognized to be important in the chemistry of life. These molecules seed the interstellar clouds which slowly gather together and eventually collapse to form new third-generation stars with associated planetary systems. Chapter 6 describes the formation of the solar system about 4.6 billion years ago, from a spinning giant molecular cloud which contracts and condenses. Dust in the outer parts of the cloud condenses to form first rock and metals. These accrete to form planetesimals, which combine eventually to form the inner planets.

Our attention switches in Chapter 7 to the early history of the Earth, its differentiation into core, mantle, crust, ocean, and atmosphere and the factors influencing surface composition and temperature. The Earth acquires a Moon in a collision with another planet-size body, which we call Theia. Earth's fluid mantle convects, and the continents start to move. Realignment of the outer planets precipitates the Late Heavy Bombardment, as billions of billions of tonnes of rock and ice crash to the planet's surface. We trace the subsequent evolution of the Earth to the point where conditions are right for the emergence of life.

Chapter 8 opens with a relatively stable warm wet Earth sprinkled with a variety of organic chemicals known to be essential to life and with natural deep-ocean geological systems that may act as factories for the conversion of simple inorganic chemicals into complex biochemical systems. The fossil record shows that primitive, single-celled organisms existed on Earth about 3.5 billion years ago, just a billion years after the Earth is first formed. The spontaneous generation of living from non-living matter is called *abiogenesis*. How this happens remains essentially mysterious. We have some compelling theories but there is as yet no 'standard model' for the origin of life.

Irrespective of how it comes about, we do know that the basic biological structures that emerge about 3.5 billion years ago establish a template that will be replicated in all subsequent life forms. Chapter 9 describes how early single-celled organisms use photosynthesis to terraform the planet by pumping their waste oxygen into the atmosphere. Oxygen opens up new opportunities for evolution to experiment with. Some single-celled organisms now merge with each other to form complex cells, and go on to form larger and larger multicellular creatures. After 2.8 billion years of evolution, the first primitive animals appear.

The story of the last 540 million years or so of Earth's history is a song of ice and fire. As Chapter 10 recounts, the planet swings violently between inhospitable ice ages and periods of deadly volcanic eruptions and at least one asteroid impact. Life clings on, both in the oceans and now on the land. Evolution drives frantic periods of diversification of different animal species, only for these to be cruelly pegged back in a series of planet-wide mass extinctions. In the mass extinction event 252 million years ago, called the 'Great Dying', almost 95% of all marine animals and a substantial proportion of all land animals are wiped out. But each extinction is followed by an evolutionary 'radiation' in the near-empty ecology that results. From the ashes of the Great Dying comes the age of the dinosaurs, until they too are destroyed by an asteroid, 66 million years ago.

Among the creatures that survive this last calamity are small mammals. Chapter 11 describes the evolutionary radiation that now occurs, paying particular attention to the primates. As primate species diversify, they follow different evolutionary lines, much like the branches of a tree. The first to branch away are tarsier monkeys, followed by New World monkeys, Old World monkeys, gibbons, orangutans, and gorillas. Finally, about 5–7 million years ago, chimpanzees part evolutionary company with the line of hominins.

Tracing the lineage of modern humans from this point is fraught with difficulty and much is contested. But the fossil evidence supports a line that brings us through early hominins dated 4–7 million years old, to species of the genus *Australopithecus*, 2–4 million years old, to early *Homo* species 2 million years ago. Of these, *Homo sapiens* appears in Africa just 200 000 years ago.

There can be little doubt that human consciousness is what sets humans apart. The ability to have abstract thoughts, imagine concepts, and develop intelligence has allowed humans to break free from the prison of the present, and is the basis for the proliferation of a humanity of 7.4 billion across the Earth. Chapter 12 attempts to describe the origin and evolution of consciousness, its foundation in genetics, and its close fundamental ties to the development of language and society.

I've mapped out a 'timeline of creation' in Table 1. This lists the origins of its various singular components, from space, time, and energy all the way to human consciousness, measured in time from the present (2017) and from the big bang as 'time zero'. To put this into some kind of perspective that we mere mortals might grasp, I've also mapped the time measured from the big bang onto a single hypothetical 24-hour 'day of creation'.

On this reckoning, the universe 'begins' at midnight. Particles with mass appear the merest whisper of a fraction of a second afterwards, and the universe is bathed in light at the moment of recombination two seconds later, as primordial electrons latch themselves to primordial hydrogen and helium nuclei. Stars and galaxies first appear between 12:30 and 1:00 a.m., with complex molecules starting to make their appearance sometime between 3 a.m. and 6:30 a.m., in time for breakfast.

We're then obliged to sit on our hands for most of the day—9–10 hours—as we wait for the Sun and Earth to appear, at nearly 4 p.m. At some time during this wait, the expansion of the universe flips. The matter in the universe that has thus far been slowing down the rate of expansion becomes so dilute that 'dark energy'—the energy of 'empty' space—takes over and starts to accelerate the expansion once again.

Life emerges around 6 p.m., and complex cells and multicellular organisms around 8:30 p.m. A few hours later we see the beginnings of the diversification of animal species in the 'Cambrian explosion'. Modern humans make their first appearance at about 1 second before midnight. Human consciousness

TABLE 1: The Timeline of Creation

Chapter	The Origin of . . .	Measured from the Present (2017)*	Measured from the Big Bang	Mapped to a Single 'Day of Creation'
1	Space, Time and Energy	13.8 Ga	'0'	Midnight
2	Mass	13.8 Ga	10^{-12} s	A fraction after midnight
3	Light	13.8 Ga	380 000 Yrs	2 seconds after midnight
4	Stars and Galaxies	13.5-13.3 Ga	300–550 million Yrs	Between 12:30 to 1:00 am
5	Molecules	10-12 Ga	1.8–3.8 billion Yrs	Between 3:00–6:30 a.m.
6	Solar System	4.6 Ga	9.2 billion Yrs	About 4 p.m.
7	Earth			
8	Life	3.5 Ga	10.4 billion Yrs	Almost 6 p.m.
9	Complex Cells and Multicellular Organisms	~2 Ga	11.8 billion Yrs	About 8:30 p.m.
10	Species (Animal Species Diversity)	540 Ma	13.4 billion Yrs	A little after 11:00 p.m.
11	*Homo sapiens*	200 ka	13.8 billion Yrs	About 1 second to midnight
12	Human Consciousness	50 ka	13.8 billion Yrs	About 300 milliseconds to midnight

* Ga = billions of years ago, Ma = millions of years ago, ka = thousands of years ago, Yrs = years, s = seconds. Based on an estimate of the age of the universe of 13.82 billion years established by recent results (21 March 2013) from the European Space Agency's Planck satellite.

develops throughout this second, but has begun to realize its full potential with the transition to behavioural modernity in the 'Great Leap Forward' which happens with just 300 milliseconds (thousandths of a second) left on the clock.

As *Origins* will, I hope, make abundantly clear, these have been a very busy few hundred milliseconds.

1

IN THE 'BEGINNING'

The Origin of Space, Time, and Energy

Don't be fooled. No matter what you might have read in some recent popular science books, magazine articles, or news features, and no matter how convincing this might have seemed at the time, be reassured that *nobody* can tell you how the universe began. Or even if 'began' is a word that's remotely appropriate in this context.

There's a good reason for this. As we will see in what follows, we know the universe is expanding. By extrapolating backwards, we know therefore that there must have been a moment in its history when all the energy in the universe was compacted to an infinitesimally small point, from which it burst forth in what we call the 'big bang'.

How do we know? This chapter will provide some of the answers to this question, and I'll provide the scientific evidence for the big bang and the expanding universe as subsequent chapters relate the story of its evolution through its first 380 000 years. Suffice to say that something like the big bang *must* have happened, and our best estimate is that it happened about 13.8 billion years ago, give or take a few hundred million years.

Describing the very 'beginning' of the universe is problematic because, quite simply, none of our scientific theories are up to the task. We attempt to understand the evolution of space and time and all the mass and energy within it by applying Albert Einstein's general theory of relativity. This theory works extraordinarily well. But when we're dealing with objects that start to approach the infinitesimally small, we need to reach for a completely different structure, called quantum theory. Now, the general theory of relativity can't handle some things in ways that quantum theory can, and vice versa. But when we try to put these two venerable theories together to create some kind of unified theory that could do the work of both, we find that they really don't get along, and the structure becomes difficult to deal with.

So far, all the fixes and work-arounds remain rather speculative, and there is no consensus on what a quantum theory of gravity should look like.

And there's another problem. Insofar as our extrapolations from the present day universe can be trusted, they tell us that the energy of a hot big bang must have been much, much greater than any energy we could ever hope to re-create on Earth in a particle collider. So, even if we could one day build a theory that could be applied with some confidence, we will simply never be able to build an apparatus to perform the experiments and make the observations that would be required to test such a theory's predictions.* We have no alternative but to rely on what we can discover from the observable universe, and use our theories to infer what *might* have happened in the distant past.

What this means is that the very beginning of the universe (if this is indeed the right word) is beyond the reach of science for the foreseeable future and, quite possibly, for all time. Of course, this doesn't stop us from speculating, and there are many contemporary theories that provide various origin-of-the-universe stories. In some of these, the universe emerges 'from nothing' in a quantum fluctuation.† Or the universe is simply one of a large number (possibly an infinite number) of expanded bubbles of spacetime in a 'multiverse' of possibilities.‡ Or the big bang results from the collapse of the universe that went before, as the cosmic reset button is pressed once again in a cycle that has lasted for all eternity.

There is no empirical evidence for any of these different ideas. It's perhaps possible that some of these theories can be developed to the point where they can predict subtle physical phenomena that might be detectable in our own universe, using Earth-bound or satellite-borne instruments (although, frankly,

* Which is probably just as well. I'm not entirely sure what would happen if the conditions that prevailed during the big bang were ever to be re-created on Earth.

† 'Nothing' is a philosophically loaded concept, best avoided if you're unprepared to argue semantics.

‡ If you're unsure what I mean by 'spacetime' stay tuned—a definition is coming.

I'm inclined to think that this possibility is remote). Even then, as I've said, the prediction of phenomena observable today still allows us only to infer what might have happened during or before the big bang. Choosing what to believe about this moment will still require something of an act of faith.

The Austrian philosopher Ludwig Wittgenstein once famously cautioned: 'Whereof one cannot speak, thereof one must be silent'.[1] That's probably good advice, but I've promised you a book about origins, so in this chapter I'm going to try to walk the fine line between accepted science—what we do know and can prove—and speculative theorizing: what we can only make moderately educated guesses at, based on scientific principles that have at least some validity. I'll hang warning signs in the appropriate places, so we don't inadvertently stumble and fall down a metaphysical rabbit hole.

Our creation story begins with the origin of space, time, and energy, and it is here that we meet our first challenge, even before we can properly start to tell the tale. For what are space, time, and energy? How should we conceive of these things?

THE NATURE OF SPACE AND TIME

I'm sitting at the desk in my study, typing these words on a keyboard which is wirelessly connected to a docked laptop, watching my sentences take shape on a large-screen monitor. If I take my eyes away from the monitor and look around me, I see a room with the architecturally favoured number of walls—four. Two of these, to my left and behind me, are decorated with shelves on which sits a modest collection of books. Against another wall to my right I have a sofa-bed which is used occasionally when I have sleep-over guests (and which today, unusually, is not piled with yet more books).

Like you, I have no hesitation in concluding that the things in this room are objects in space.

But what, precisely, *is* space? I can move through it, but I can't see it and I can't touch it. Space is not something that we perceive directly. We perceive objects (such as monitors, books, and sofa-beds) and these objects have certain relations one with another that we call spatial relations: this here on the left, that over there on the right. But space itself does not form part of the content of our direct experience. We interpret the objects as existing in a three-dimensional space as a result of a synthesis of electrical signals in our brains translated into visual perceptions by our minds.

Similarly, I move through time (in one direction, at least) but I can't see it and I can't reach out and touch it. Time is not a tangible object. My sense of time would seem to be derived from my sense of self and the objects around me changing their relative positions (this *was* on the left, *now* it's on the right), or changing their nature, from one type of thing into another.

Does the space in this room exist independently of the objects in it? Does time exist independently of the things that happen here? In other words, are space and time 'absolute' things-in-themselves?

In developing the theory of mechanics that he described in his great work *The Mathematical Principles of Natural Philosophy*, first published in 1687, Isaac Newton was willing to acknowledge the essential relativity of space and time, in what he called our 'vulgar experience'. He was willing to accept that objects move towards or away from each other, changing their relative positions in space and in time. This is relative motion, which can be defined simply in terms of the relationships between the objects.

But Newton's theory required an absolute motion which, he argued, must imply an absolute space and time that forms a kind of container within which objects exist and things happen. Take all the objects out of the universe and, Newton's theory demanded, the empty container would remain: there would still be 'something'.

Einstein begged to differ. He pondered this question while working as a 'technical expert, third class' at the Swiss Patent Office in Bern more than two hundred years later, in 1905. He concluded that absolute space and time cannot exist. This conclusion follows from Einstein's special theory of relativity.

This theory is based on two fundamental principles. The first, which became known as the *principle of relativity*, states that observers who find themselves in states of relative motion at different (but constant) speeds *must* observe precisely the same fundamental laws of physics.

This seems perfectly reasonable. Suppose I make a set of physical measurements here on Earth which allows me to deduce some underlying physical law. You make the same measurements on board a distant spaceship moving away from the Earth at high speed. The conclusions we draw from both sets of measurements must surely be the same. There can't be one set of physical laws for me and another set for space travellers. Otherwise they wouldn't be laws.

We can turn this on its head. If the laws of physics are the same for all observers, then there is no measurement we can make which will tell us which observer is moving relative to the other. To all intents and purposes, you may actually be stationary, and it is me who is moving away at high speed. We cannot tell the difference using physical measurements.

The second of Einstein's principles concerns the speed of light. At the time that he was working on special relativity, physicists had rather reluctantly concluded that the speed of light is constant, completely independent of the speed of the source of the light. If I measure the speed of the light emitted by a flashlight held stationary on Earth and you measure it again using the same flashlight on board a spaceship moving at high speed, we expect to get precisely the same answers.

Instead of trying to figure out *why* the speed of light is independent of the speed of its source, Einstein simply accepted this as an established fact. He assumed the speed of light to be a universal constant and proceeded to work out the consequences.

One immediate consequence is that there can be no such thing as absolute time.

Here's why. Suppose you observe a remarkable occurrence. During a heavy thunderstorm you see two bolts of lightning strike the ground simultaneously, one to your left and one to your right. You're standing perfectly still, so the fact that it takes time for the light from each of these lightning bolts to reach you is of no real consequence. Light travels very fast so, as far as you're concerned, you see both bolts at the instant they strike.

However, I see something rather different. I'm travelling at very high speed—half the speed of light, in fact—from left to right. I pass you just as you're making your observations. Because I'm moving so fast, the time taken for the light from the lightning bolts to reach me now has measureable consequences. By the time the light from the left-hand bolt has caught up with me, I've actually moved quite a bit further to the right, and so the light has further to travel. But the light from the right-hand bolt has less ground to cover because I've moved closer to it. The upshot is that I see the right-hand bolt strike first (Figure 1).

You see the lightning bolts strike simultaneously. I don't. Who is right?

We're both right. The principle of relativity demands that the laws of physics must be the same irrespective of the relative motion of the observer, and we can't use physical measurements to tell whether it is you or me who is in motion.

We have no choice but to conclude that there is no such thing as absolute simultaneity. There is no definitive or privileged frame of reference in which

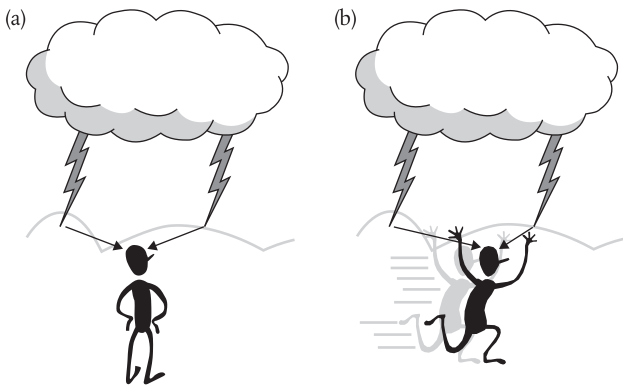

FIGURE 1 The stationary observer in (a) sees the lightning bolts strike simultaneously, as the light from both travels so fast as to appear instantaneous. But the observer in (b), who is moving at a considerable fraction of the speed of light, sees something different. He's moving at half the speed of light from left to right, so by the time the light from the left-hand bolt catches up with him, he's moved a bit further to the right. The light from the right-hand bolt now has less far to travel. Consequently the observer in (b) sees the right-hand bolt strike first.

we can declare that these things happened at precisely the same time. They may happen simultaneously in this frame or they may happen at different times in a different frame, and all frames are equally valid. Consequently, there can be no 'real' or absolute time. Something's got to give. We perceive events differently because time is relative.

Einstein developed a similar set of arguments to show that space is relative, too. The bizarre consequences of special relativity are reasonably well known. Demanding that the laws of physics appear the same for all observers in a universe in which the speed of light is fixed means that time intervals (durations) can dilate and spatial intervals (distances) can contract. This means that durations and distances will be measurably different for different observers travelling at different speeds.

But all is not lost. Time dilation and distance contraction are like two sides of the same coin. They're linked by the speed of the observer making the measurements relative to the speed of light. If we now combine space and time together in a four-dimensional *spacetime*, then intervals measured in this spacetime are unaffected by relativity.[2] In spacetime intervals, time dilations are compensated for by distance contractions, and vice versa.

Does this mean that, although space and time are relative, spacetime is absolute? Some contemporary physicists think so. Others disagree. What's important for us to realize is that we must abandon our simplistic, common-sense notions of an independent space and time and accept that in our universe these are inextricably connected.

MASS AND ENERGY

Einstein's 1905 research paper on special relativity was breathtaking in its simplicity yet profound in its implications. But he wasn't quite finished. He continued to think about the consequences of the theory and just a few months later he published a short addendum in the same journal.

In this second paper he considered the situation in which a moving object emits two bursts of light in opposite directions. The initial energy of the object is entirely in the form of its energy of motion (which is called kinetic energy). Each burst of light carries the same amount of energy away from the object, ½*E*. As the total energy must be conserved, the object's kinetic energy must therefore fall by a total amount *E*. This makes perfect sense. The light carries energy away, and the energy must come from somewhere.

He then imagined what two different observers might measure, with one observer moving along in the 'rest frame' of the object (keeping pace with the object so that it appears as though at rest) and the second observer moving at a constant speed relative to this frame. These observers measure the difference in the energy of the object before and after emission of the light. He found, perhaps not altogether surprisingly, that the different observers get different results.

After a little bit of algebraic manipulation, he arrived at a mathematical expression which allowed him to draw an extraordinary conclusion. The energy of the light bursts comes from the object's kinetic energy of motion. This can be calculated from the mass of the object and its speed.[3] From the difference in the two sets of results, Einstein deduced that the energy carried away is derived not from the object's speed (as might be anticipated), but from its *mass*.

If the total energy carried away by the light is *E*, Einstein concluded that the mass (*m*) of the object must diminish by an amount *E* divided by c^2, where *c* is the speed of light. It doesn't matter what kind of object we might be referring to: this is a general result, universally applicable. The inertial mass of an object (a measure of its resistance to acceleration) is also a measure of the amount of energy it contains.

Today we would probably rush to rearrange the equation in Einstein's paper to give the iconic formula $E = mc^2$. But Einstein himself didn't do this. Although he was uncertain that this was something that could ever be tested experimentally, he was prepared to speculate that the conversion of mass to energy might one day be observed in radioactive substances, such as radium.

The special theory of relativity blurs our commonsense conceptions of a physical reality of space and time, matter, and energy. Space and time are relative, they are defined by the things they contain and events that happen, though spacetime *might* be absolute. Mass is energy, and from energy can spring mass. These modifications of our common conceptions are important to acknowledge if we are to understand precisely what it was that originated in the big bang.

But there is yet a further modification we need to make. Spacetime and mass-energy do not themselves exist completely independently of one another. They are locked together in an elegant dance described by Einstein's general theory of relativity.

GRAVITY AND GEOMETRY

Gravity is a familiar 'everyday' kind of force. When I drop something, it falls to the ground. It does this because it experiences the force of gravity. We struggle against this force every morning when we get out of bed. We fight its effects every time we lift a heavy weight. When we stumble to the ground and graze a knee, it is gravity that causes the hurt. Such is its familiarity that it's tempting to assume that science must have long ago answered all our questions about it.

And, indeed, we learn in school of Newton's law of universal gravitation. Bodies of material substance are attracted to one another, with the force of attraction increasing with the product of their masses and inversely with the square of the distance between them.[4] But, although Newton's law was a great achievement, there are some real problems with its interpretation, problems that were obvious in Newton's time but for which he could offer no solutions.

In the mechanical universe described by Newton's laws of motion, we interpret force to be something that is exerted or imparted by objects impinging on each other. A stone does not move unless we kick it or throw it, thereby accelerating it to some final speed as it sails through the air. But precisely what is it that grasps the Moon as it swoons in Earth's gravitational embrace? How does the Moon push the afternoon tide up against the shore? When a cocktail glass slips from a guest's fingers, what grabs it and forces it to shatter on the wooden floor just a few feet below?

Newton was at a loss. His force of gravity seems to imply some kind of curious action-at-a-distance. Objects influence each other over great distances through empty space, with nothing obviously transmitted between them. Critics accused him of introducing 'occult elements' in his theory of mechanics.

Part of the solution to this riddle would come to Einstein during an otherwise average day at the Patent Office in November 1907, by which time he had been promoted to 'technical expert, second class'. As he later recalled: 'I was sitting in a chair in my patent office at Bern. Suddenly a thought struck me: If a man falls freely, he would not feel his weight'.[5]

In this stunningly simple observation, Einstein realized that our local experiences of gravity and of acceleration are the same. He called it the equivalence principle. Working out what this meant would take him another eight years and would require another extraordinary connection, between gravity and geometry.

The geometry we learn about in school is Euclidean geometry, named for the ancient Greek mathematician Euclid of Alexandria. In this geometry, parallel lines never cross, the angles of a triangle add up to 180 degrees, and the circumference of a circle is twice its radius multiplied by π. This is a geometry associated with a kind of three-dimensional space that mathematicians call 'flat'. We learn about Euclidean geometry because the spacetime of our universe happens to be a flat spacetime.

In a flat space the shortest distance between two points is obviously the straight line that we can draw between them. But what is the shortest distance between London and Sydney, Australia? We could look up the answer: 10,553 miles. But this distance is not, in fact, a straight line. The surface of the Earth is curved, and the shortest distance between two points on such a surface is actually a curved path called a *geodesic*.

Now comes a rather breathtaking leap of imagination. What if the spacetime near a large object isn't 'flat'? What would happen if it were to be curved? Einstein realized that he could get rid of the action-at-a-distance implied by Newton's gravity by replacing it with curved spacetime. An object with a large mass-energy warps the spacetime around it, and objects straying close to it follow the shortest path determined by this curved spacetime.

American physicist John Wheeler summarized the situation rather succinctly some years later: 'Spacetime tells matter how to move; matter tells spacetime how to curve'.[6] In general relativity, gravity is not a force that matter exerts on matter. It is a force that matter (or, strictly speaking, mass-energy) exerts on spacetime itself. Objects do experience a mutual gravitational attraction, but the attraction is *indirect*, mediated by the curvature of the spacetime between them.

These are subtle, but real, effects. The general theory of relativity correctly accounts for some peculiarities in the orbit of the planet Mercury that originate in the curvature of spacetime near the Sun, something that Newton's theory of gravity fails to predict correctly. And, although light from a distant star that passes close to the Sun on its way to Earth follows a straight-line path, the curvature of spacetime near the Sun makes the path appear to bend. When this phenomenon was first demonstrated during a total eclipse in 1919, Einstein became a household name.

On 24 April 2004, an exquisitely delicate instrument called Gravity Probe B was launched into polar orbit, 642 kilometres above the Earth's surface. The satellite housed four gyroscopes, designed to measure the effects of spacetime curvature around the Earth. To eliminate unwanted torque, the satellite was rotated once every 78 seconds and thrusters were used to keep it pointing towards the star IM Pegasi in the constellation of Pegasus (Figure 2).

Two effects were measured. The curvature of spacetime causes the gyroscopes to precess* in the plane of the satellite's orbit (that is, in a north–south direction), an effect known as *geodetic drift*. The second effect is *frame-dragging*. As the Earth rotates on its axis, it drags spacetime around with it in the plane perpendicular to the plane of the satellite orbit (in a west–east direction). This gives rise to a second precession of the gyroscopes.

The results were announced at a press conference on 4 May 2011. Although an unexpected wobble in the gyroscopes had resulted in some significant uncertainty, the measurements of both geodetic drift and frame-dragging provided a very powerful experimental vindication of general relativity.

Should it be needed, we can cite one last piece of firm evidence. In June 1916, Einstein speculated that small fluctuations in a gravitational field would, like

* This results in a shift in the axis of rotation of the gyroscopes which, though very small, can be measured with great accuracy. I remember playing with a toy gyroscope when I was young. You would set it spinning with a pull string, sit back and watch (these were simpler times). It taught me about precession, although I didn't know that was what I was seeing at the time.

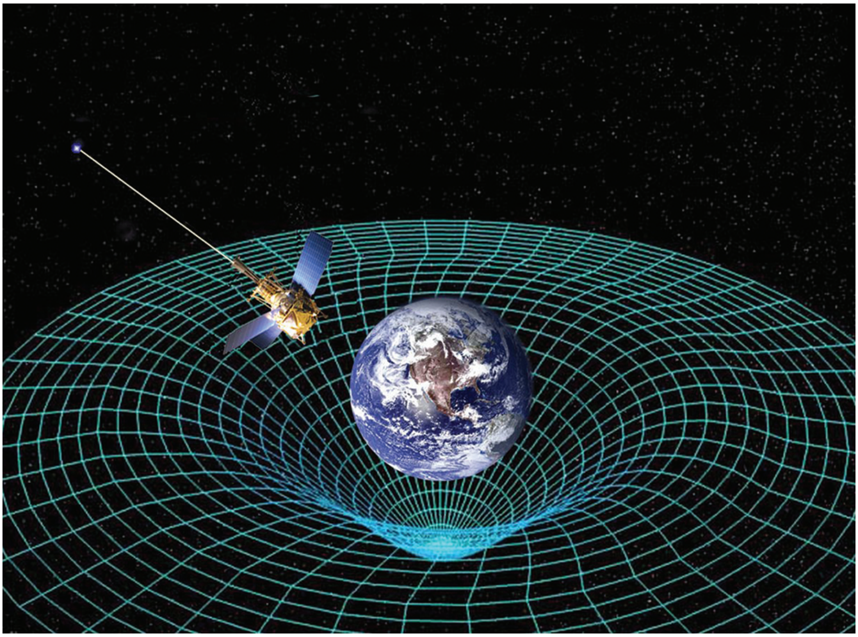

FIGURE 2 Gravity Probe B was launched in April 2004 and measured two phenomena associated with the curvature of spacetime around the Earth. The results were announced in May 2011, and provided a powerful vindication of general relativity. This picture shows the satellite moving in the curved spacetime around the Earth, pointing towards the star IM Pegasi, in the constellation of Pegasus.

the ripples on the surface of a lake, appear as waves. Such *gravitational waves* can only be produced by two large masses rotating around each other in what astronomers call a binary system, stretching and compressing the spacetime between them. It was not until the 1950s and 1960s that physicists thought they might stand a chance of actually detecting gravitational waves, and on 15 September 2015 their patience was finally rewarded.

This was the date on which gravitational waves generated by the merger of two black holes were recorded by an experimental collaboration called LIGO, which stands for Laser Interferometry Gravitational-wave Observatory. The result was announced at a press conference on 11 February 2016. Several such black hole mergers have since been reported, and in 2017 gravitational waves produced by the collision of two neutron stars were detected by LIGO and the Virgo experiment in Italy.

The successful detection of gravitational waves is not only an extraordinary vindication of general relativity, it also opens a new window on events in distant parts of the universe, one that doesn't rely on light or other forms of electromagnetic radiation to tell us what's happening.

THE EXPANDING UNIVERSE

Einstein presented his new general theory of relativity to the Prussian Academy of Sciences in Berlin in 1915. Two years later he applied the theory to the whole universe.

At first glance, this seems impossibly difficult. How can a single set of equations describe the whole universe? The answer is: by making a couple of simplifying assumptions. Einstein had to assume that the universe is uniform in all directions, containing objects that have the same kind of composition. He also had to assume that the universe we observe from our vantage point on Earth is no different from the universe as observed from any and all such vantage points. In other words, observers on Earth occupy no special or privileged position. What we see is a 'fair sample' of the universe as a whole.

What Einstein got was singularly appealing, a universe that is finite but nevertheless 'unbounded', without edges. We know that the ground on which we walk appears to be flat, but if we walk far enough in one direction we also know that we will eventually circumnavigate the Earth, without falling off the edge. So in Einstein's universe, spacetime curves back on itself like the surface of a sphere. At any one point in this universe spacetime looks flat but it is, in fact, gently curved.

But Einstein then quickly ran into a big problem. He had anticipated that the universe that should emerge from his equations would be consistent with prevailing scientific prejudice—a universe that is stable, static, and eternal. What he got instead was a universe that is dynamic—either expanding or contracting depending on the initial assumptions. The field equations suggested that a static universe is impossible.

Gravity is the weakest of nature's forces (we'll meet the others later in this chapter). But it is cumulative and inexorable and acts only in one 'direction'—it serves to pull objects together but it doesn't push them apart. Einstein realized that the mutual gravitational attraction between all the material objects in the universe would cause it to collapse in on itself. This was a troubling result, quite inconsistent not only with prevailing opinion but also arguably with simple observation. After several centuries of astronomy there is no evidence that all the stars in the universe are rushing towards each other in a catastrophic collapse. Quite the opposite, in fact, as we will see.

But there was nothing in his gravitational field equations to stop this from happening. As he later explained: 'We admittedly had to introduce an extension to the field equations that is not justified by our actual knowledge of gravitation'.[7]

The left-hand side of the most general form of this equation describes the curvature of spacetime and hence the strength of the force of gravity that will act on all the mass-energy, which is summarized in the right-hand side of the equation. Einstein chose to modify the equation by subtracting from the left-hand side a term containing a 'cosmological constant', usually given the Greek symbol Λ (lambda).

In essence, this extra term imbues spacetime with a kind of odd, anti-gravitational force, a kind of negative pressure which builds in strength over long distances and counteracts the effect of the curvature caused by all the mass-energy in the universe. By carefully selecting the value of the cosmological constant, Einstein found that he could balance the gravitational attraction that tended to pull everything together, caused by all the mass-energy on the right-hand side, with a spacetime on the left-hand side that has a tendency to push everything apart. The result was perfect balance, a static universe.

It was quite a neat solution. Introducing the cosmological constant didn't alter the way general relativity works over shorter distances, so the successful predictions of the orbit of Mercury and the bending of starlight were preserved. But it was, nevertheless, a rather unsatisfactory 'fudge'. There was no evidence for the cosmological constant, other than the general observation that the universe *seems* to be stable and static.

Freed from prejudices about the kind of universe that *should* result, Einstein's field equations actually yield many different kinds of possible solutions. In 1922, Russian physicist and mathematician Alexander Friedmann offered a number of different solutions of Einstein's original equations. He was not particularly interested in trying to represent our own universe, preferring instead to explore the different possibilities allowed by the mathematics. Consequently, while he retained the cosmological term that Einstein had introduced, he assumed that it could take any value, including zero.

Friedmann discovered a range of different possible model universes, with properties and behaviour that depend on the relationships between the amount of mass and the size of the cosmological constant. He focused his attention on solutions with positive spacetime curvature, showing that they could expand or contract. He was particularly taken with solutions based on an assumed cosmological constant of zero which oscillate back and forth, alternating between expansion and contraction, the period of oscillation depending on the amount of mass.

A universe in which the density of mass-energy is high (lots of objects in a given volume of space) and the rate of expansion is modest is said to be 'closed'. It will expand for a while before slowing, grinding to a halt and then turning in on itself and collapsing. Spacetime in such a universe has a positive curvature. A few years later Friedmann examined universes in which spacetime is negatively curved. Such universes are infinite, they are said to be 'open' and will expand forever.

Tragically, Friedmann died of typhoid fever in 1925. But his expanding-universe solutions were independently rediscovered in 1927 by Belgian theorist (and ordained priest) Abbé Georges Lemaître. Einstein was initially dismissive of the idea of an expanding universe but when, in the early 1930s, the evidence from observational astronomy* suggested rather strongly that the universe is indeed expanding, he accepted that he had been wrong, and expressed regret for fudging his equations.[8]

The evidence in favour of an expanding universe became overwhelming in 1965, with the discovery of the cosmic background radiation. This is the cold remnant of hot radiation that spilled into the universe soon after its birth. We will encounter it in Chapter 3.

But what *causes* the universe to expand? In a paper published in 1933, Lemaître suggested that expansion is triggered because empty spacetime is not, in fact, empty. Einstein had introduced his cosmological term on the left-hand side of his field equation, as a modification of spacetime itself, designed to offset the effects of the curvature caused by all the mass-energy in the universe. But it takes just a moment and a little knowledge of algebra to move the cosmological term from the left-hand (spacetime) side of the equation to the right-hand (mass-energy) side. Now it represents a *positive* contribution to the total mass-energy of the universe. This is not the familiar mass-energy we associate with stars, planets, and people. Rather, it takes the form of an energy of 'empty' spacetime, sometimes called *vacuum energy*.[9]

It seems that Lemaître's paper had little impact at the time. But it will be useful to remember this connection between the rate of expansion of spacetime and vacuum energy, as we'll need it again quite soon.

* We will take a closer look at this evidence in Chapter 4.