
VOL TWO / ISSUE FIVE
NewScientist
THE COLLECTION
15
IDEAS
YOU NEED TO
UNDERSTAND
HUMAN ORIGINS THEORY OF EVERYTHING
ARTIFICIAL INTELLIGENCE RELATIVITY
SECRETS OF SLEEP AND MANY MORE...
A SERIES OF INSTANT EXPERT GUIDES
WRITTEN BY LEADING SCIENTISTS
£9.99
05
9 772054 638003

And not just any beer.

We're talking about the very best beers from the most exciting breweries around the globe.

Until 20th December 2015, Beer Hawk are offering a choice of three exclusive offers for New Scientist readers.

You can **save up to 33%** on our beers and free delivery options are available to most UK addresses.

------- OFFER 1 -------

Save 33% on your first Beer Club delivery

Use the code: **NSxmasBC**

------- OFFER 2 -------

Save 20% on Discovery Cases

Use the code: **NSxmas20**

------- OFFER 3 -------

Save £5 on Christmas Gifts

Use the code: **NSxmas5**

Claim your exclusive discount now at **www.beerhawk.co.uk/new-scientist**

TERMS AND CONDITIONS

You or anybody you buy beer for must be 18 years or over. Offer valid until midnight 20th December 2016. All goods are subject to availability. Prices are only valid at the time of printing, with special offers valid for a maximum of 4 weeks after the posting date unless otherwise stated. In the unlikely event of beers becoming unavailable, a substitute of similar style and of equal or greater value will be supplied. Please see www.beerhawk.co.uka/terms-and-conditions for our full terms and conditions. This offer is brought to you by Beer Hawk Ltd, a limited company registered in England and Wales with registration number 08118833. Unit 1 Saltergate Business Park, Harrogate, North Yorkshire, HG3 2BX.

NewScientist
THE COLLECTION

VOL TWO / ISSUE FIVE
15 IDEAS YOU NEED TO UNDERSTAND

NEW SCIENTIST
THE COLLECTION
110 High Holborn,
London WC1V 6EU
+44 (0)20 7611 1202
enquiries@newscientist.com

Editor-in-chief Graham Lawton
Editor Jeremy Webb
Art editor Craig Mackie
Picture editor Kirstin Jennings
Subeditor Chris Simms
Graphics Dave Johnston
Production editor Mick O'Hare
Project manager Henry Gomm
Publisher John MacFarlane

New Scientist The Collection is published four times per year by Reed Business Information Ltd
ISSN 2054-6386

Printed in England by Precision Colour Printing, Telford, Shropshire, and distributed by Marketforce UK Ltd
+44(0)20 3148 3333
Display advertising
+44(0)20 7611 1291
displayads@newscientist.com

Cover image
Chris Nurse

Knowledge is power

THERE are few more satisfying things in life than getting your head around a profound idea. And if you want to understand yourself and the world around you, there is nowhere better to look for a profound idea than science.

The best way to get to grips with a novel concept is to talk to an expert, but there's rarely one around when you're in need. This issue of *New Scientist: The Collection* is the next best thing: we've done the hard work for you.

The articles in this collection started life as editions of Instant Expert, a monthly supplement in *New Scientist* magazine. Every month we'd ask a leading scientist to write an introductory guide to their specialist subject. We chose scientists who not only knew the topic inside out but who could communicate it with verve and style.

Over the years, these guides have grown into a formidable repository of knowledge. We've now selected and updated 15 of the most significant and popular ideas. I should like to thank our authors for reviewing their guides to ensure that they are bang up to date.

Chapter 1 surveys topics of universal significance. We begin with the much sought-after theory of everything – a single, exquisite idea from which all physical reality will flow. Next we dive into general relativity, Albert Einstein's 100-year-old tour de force, which is still delivering surprises. We finish the opening chapter with a tour of the unseen universe: if we ignore visible light, what do gamma rays, X-rays, and radio waves reveal about the structure of the cosmos?

Chapter 2 turns the spotlight on ourselves, asking what the hubristically named "wise person" *Homo sapiens* knows about its own origins, including the evolution of our lineage and the emergence of language.

Chapter 3 delves deeper into ourselves with a look at the human mind. We lay bare the workings of the brain, the nature of memory and of that strange, alluring capacity we call intelligence. Finally, we probe the purpose and value of sleep, what happens when it goes wrong and how we might cope with a 24/7 lifestyle.

The subject matter of Chapter 4 is the stuff of disaster movies: earthquakes, hurricanes and other natural catastrophes. What causes them, can we predict them and can we lessen their destructive impact? We finish the chapter by examining mass extinctions, those brief but lethal periods in Earth's past when vast swathes of species were wiped out in a short space of time.

Finally, in Chapter 5, we turn our attention to three ideas that seem certain to play a role in our future. Superconductivity may have been discovered a century ago, but we have only recently started to make good use of it, and its heyday is yet to arrive. Artificial intelligence and quantum information are both burgeoning fields that are finally starting to live up to their great promise. Self-driving cars and instantaneous language translation are just the start. Harnessing the weirdness of the quantum world could change our lives immeasurably.

The 15 ideas in this collection are among the most significant products of our quest to understand and master the universe. Prepare to be enlightened – and profoundly satisfied.

Jeremy Webb, Editor

CONTENTS

NewScientist
THE COLLECTION

CONTRIBUTORS

Michael Duff
is a physicist at Imperial College London, UK

Pedro Ferreira
is a physicist at the University of Oxford, UK

Michael Rowan-Robinson
is a physicist at Imperial College London, UK

Tim White
is a palaeoanthropologist at the University of California, Berkeley, US

W. Tecumseh Fitch
is a cognitive scientist at the University of Vienna, Austria

Michael O'Shea
is a neuroscientist at the University of Sussex, UK

Linda Gottfredson
is a psychologist at the University of Delaware, US

Jonathan K. Foster
is a neuropsychologist at Curtin University, Western Australia

Derk-Jan Dijk
is director of the Surrey Sleep Research Centre, UK

Raphaëlle Winsky-Sommerer
is a sleep researcher at the University of Surrey, UK

Susan Hough
is a seismologist at the US Geological Survey in Pasadena, California, US

Jeff Master
is a meteorologist based in Ann Arbour, Michigan, US

Michael J. Benton
is a palaeontologist at the University of Bristol, UK

Stephen Blundell
is a physicist at the University of Oxford, UK

Peter Norvig
is a computer scientist based in Mountain View, California, US

Vlatko Vedral
is an information scientist at the University of Oxford, UK

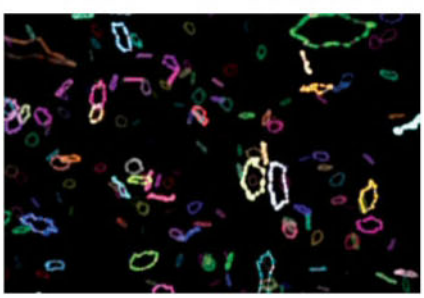
More detailed biographies are available at the end of each article

The articles here were first published in *New Scientist* between July 2010 and July 2013. They have been updated and revised.

VOLUME TWO / ISSUE FIVE

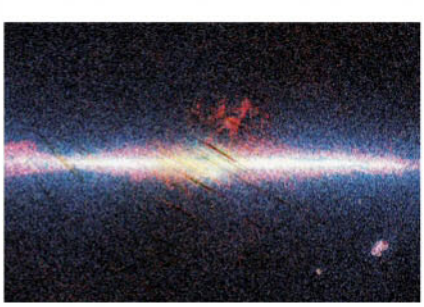
15 IDEAS YOU NEED TO UNDERSTAND

1 Secrets of the universe



7 **Theory of everything**
The idea from which all reality flows

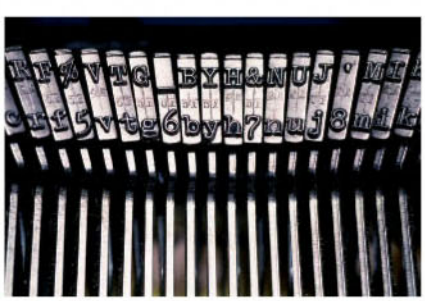
15 **General relativity**
Meet gravity, curved space and black holes



23 **Unseen universe**
Beyond visible light: the cosmos revealed

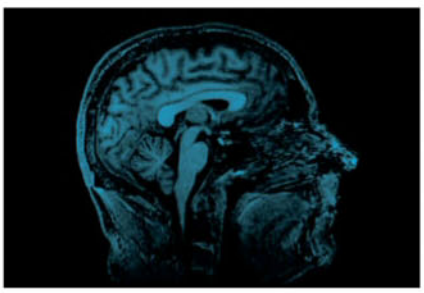
2 Becoming human

31 **Human origins**
Findings from 7 million years of evolution

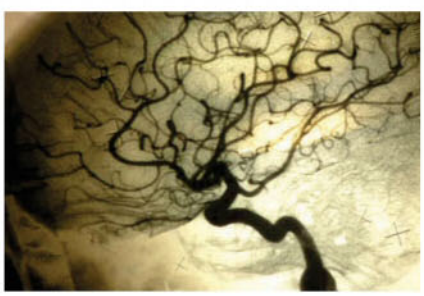


39 **The evolution of language**
A journey from silence to eloquence

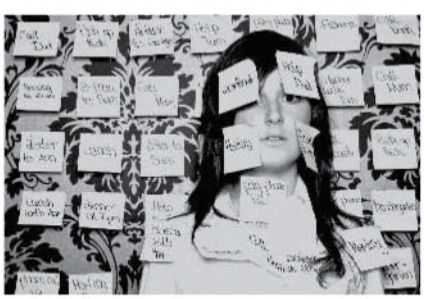
3
Your amazing mind



47 The human brain
What really makes you tick



55 Intelligence
What is it, and can you enhance it?



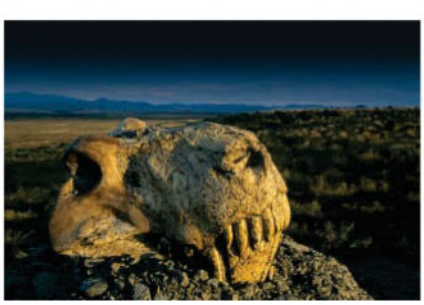
63 Memory
How the hippocampus keeps us human

71 Sleep
Why all complex creatures do it

4
Global hazards

79 Earthquakes
Will we ever be able to predict them?

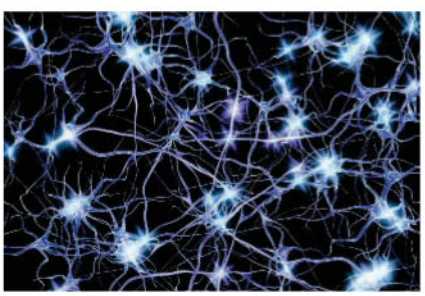
87 Extreme weather
Chasing the roots of severe storms



95 Mass extinctions
What kills half of Earth's species at a stroke

5
Shaping the future

105 Superconductivity
A century old, its full promise still lies ahead



113 Artificial intelligence
Enter a world of machines that think like us

121 Quantum information
Weirdness that could transform all our lives

Porsche recommends Mobil 1 and MICHELIN

If history is any indication, you're looking at the future of sports cars.

The new 911.
Ever ahead.

Discover more at porsche.co.uk/911.

Official fuel economy figures for the 911 Carrera 4S Coupé in mpg (l/100km): urban 22.8 – 27.4 (12.4 – 10.3), extra urban 41.5 – 42.8 (6.8 – 6.6), combined 31.7 – 35.8 (8.9 – 7.9). CO_2 emissions: 204 – 180 g/km. The mpg and CO_2 figures quoted are sourced from official EU-regulated tests, are provided for comparability purposes and may not reflect your actual driving experience.

PORSCHE

COMPLETE YOUR COLLECTION

Missed a copy of *New Scientist: The Collection*? All past issues are available to buy through our online store: **newscientist.com/thecollection**

NewScientist

CHAPTER ONE

SECRETS OF THE UNIVERSE

THEORY OF EVERYTHING

MICHAEL DUFF

INSTANT
EXPERT

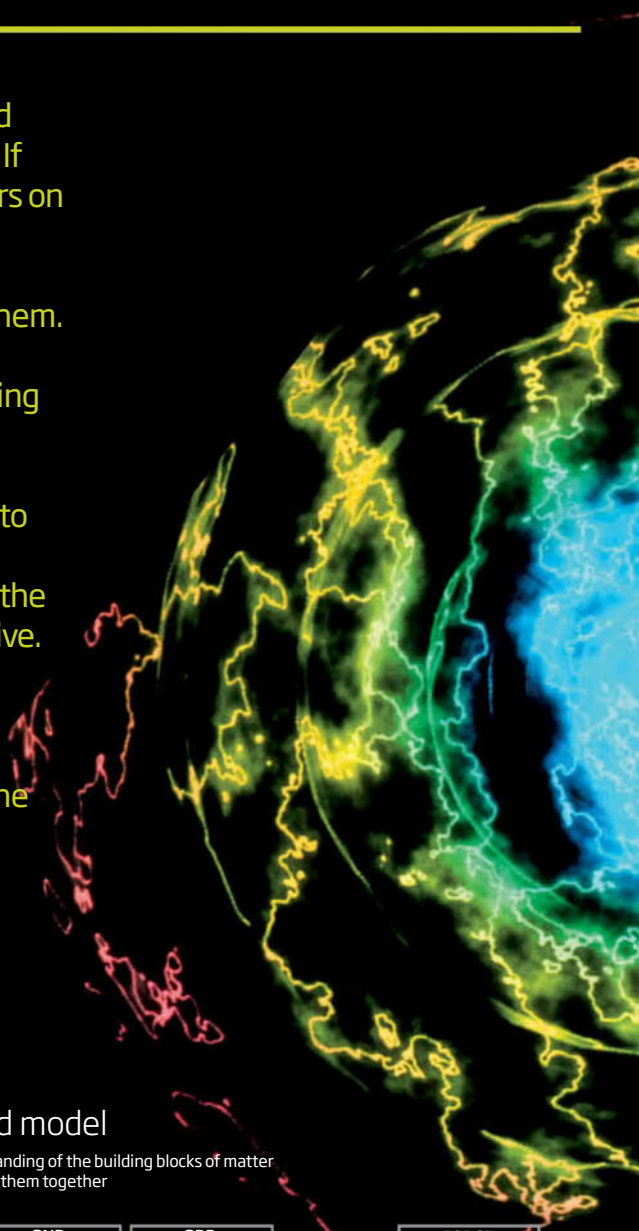
THE BIG QUESTIONS

Theoretical physicists like to ask big questions. How did the universe begin? What are its fundamental constituents? And what are the laws of nature that govern those constituents? If we look back over the 20th century, we can identify two pillars on which our current theories rest.

The first is quantum mechanics, which applies to the very small: atoms, subatomic particles and the forces between them. The second is Einstein's general theory of relativity, which applies to the very large: stars, galaxies and gravity, the driving force of the cosmos.

The problem we face is that the two are mutually incompatible. On the subatomic scale, Einstein's theory fails to comply with the quantum rules that govern the elementary particles. And on the cosmic scale, black holes are threatening the very foundations of quantum mechanics. Something has to give.

An all-embracing theory of physics that unifies quantum mechanics and general relativity would solve this problem, describing everything in the universe from the big bang to subatomic particles. We now have a leading candidate. Is it the much anticipated "theory of everything"?



BUILDING BLOCKS

At the end of the 19th century, atoms were believed to be the smallest building blocks of matter. Then it was discovered that they have a structure: a nucleus made of protons and neutrons, with electrons whizzing around it. In the 1960s, the atom was divided even further when it was theorised, then confirmed by experiments, that protons and neutrons are composed of yet smaller objects, known as quarks.

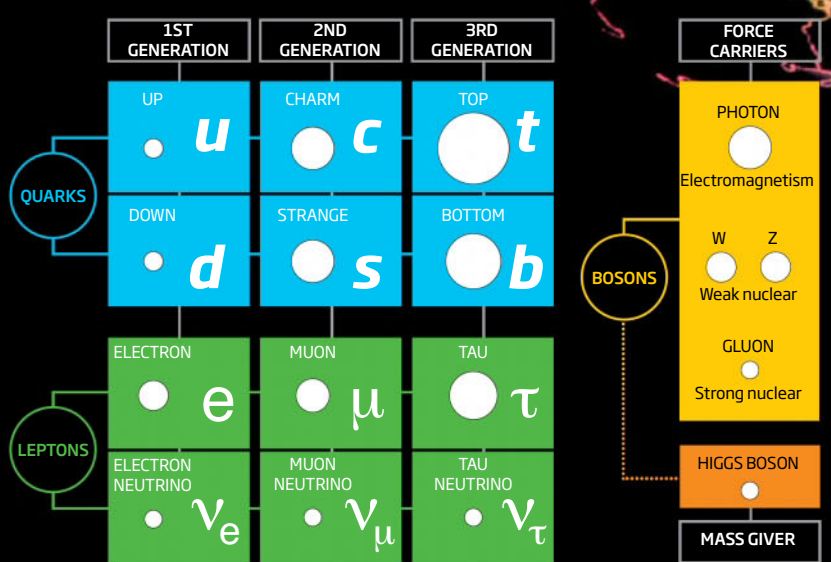
Do these layers of structure imply an infinite regression? All the theoretical and experimental evidence gathered so far suggests not: quarks really are the bottom line. We now believe that quarks are fundamental building blocks of matter along with a family of particles called the leptons, which includes the electron (see table, right).

More or less everything we see in the world around us is made from the lightest quarks and leptons. The proton consists of two up quarks and one down quark, while a neutron is made of two downs and one up. Then there is the electron along with the electron neutrino, an extremely light particle involved in radioactivity.

Nature is not content to stop there. There are two more "generations" of quarks and leptons which are like the first, but heavier. In addition, all these particles have antimatter partners which have the same mass but opposite charge.

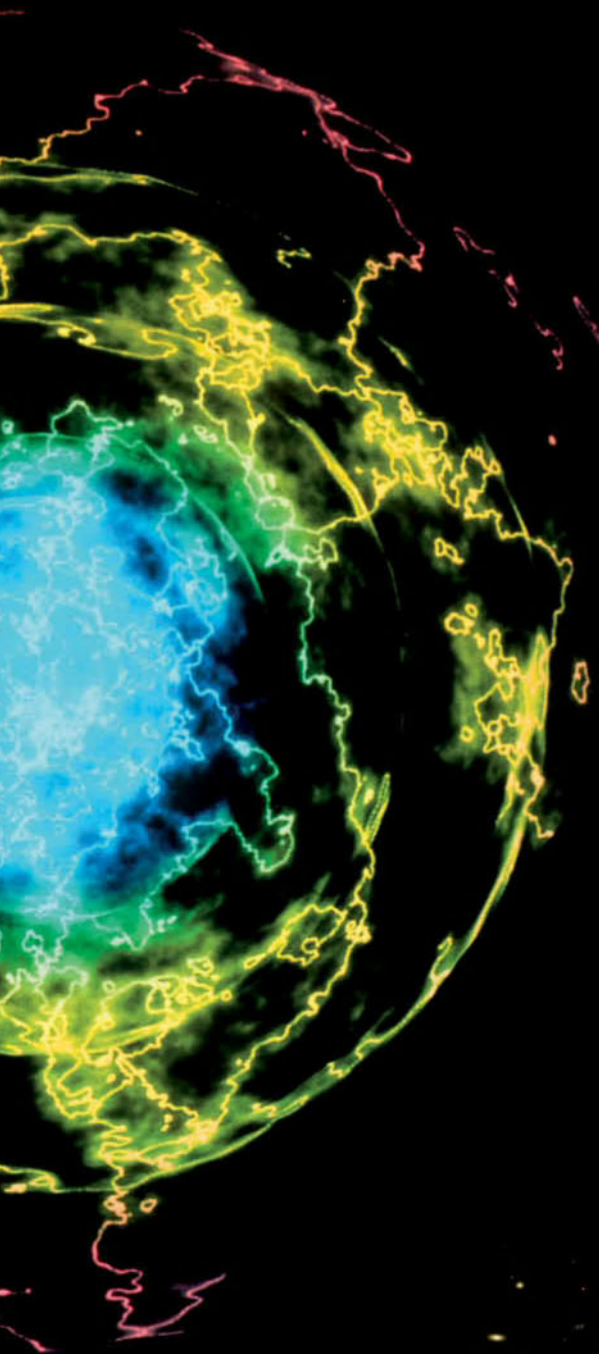
The standard model

This is our best understanding of the building blocks of matter and the forces that glue them together



RIGHT: ANDREW FRUCHTER AND THE ERO TEAM (SYLVIA BAGGETT (STSCI)/RICHARD HOOK (ST-ECF)/ZOLTAN LEVAY (STSCI)/NASA

FOUR FUNDAMENTAL FORCES



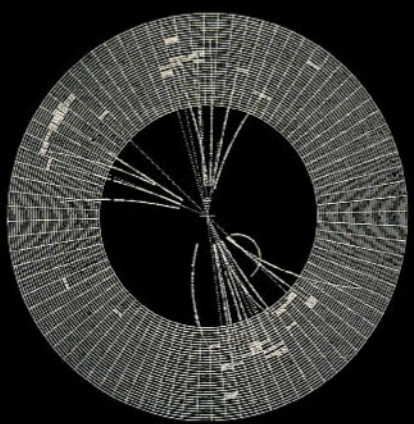
What holds the quarks and leptons, or building-block particles, together? As far as we can tell, there are four fundamental forces: gravity, which keeps Earth going round the sun; electromagnetism, responsible for phenomena such as light and electricity; the weak nuclear force, responsible for radioactivity; and the strong nuclear force, which binds neutrons and protons in the atomic nucleus.

But what exactly is a force? The modern view is based on a theoretical framework called quantum field theory. This says that the forces between building-block particles are carried or "mediated" by another set of particles. The most familiar of these is the photon, the mediator of the electromagnetic force. When two electrons repel one another, they do so by swapping a photon. This idea was backed up so well by experiments that theorists invented other force-carrying particles: the gluon for mediating the strong force, and the W and Z particles for the weak force.

Sure enough, the existence of the gluon and W and Z was confirmed in the 1970s and 80s. When you put all this together with the recently discovered Higgs boson, whose job it is to give particles their mass, you get the standard model of particle physics.

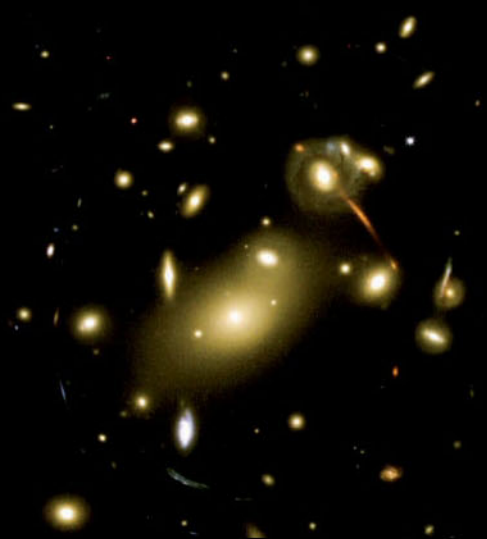
The standard model is a remarkably robust mathematical framework which makes very definite predictions about particle physics that have so far withstood all experimental tests. For example, the fact that quarks and leptons come in three generations is not put in by hand but is required by mathematical consistency; the standard model would not work if one member of the family was missing. For this reason, theory demanded the existence of the top quark, which was duly discovered in 1995.

Many regard the standard model as one of the greatest intellectual achievements of the 20th century. Yet it cannot be the final word because vital questions remain unanswered.



The gluon's discovery was a boost for the standard model

GRAVITY



Distorted space-time bends light from distant galaxies

One glaring omission from the standard model is gravity; where does that fit in? According to Albert Einstein's view of gravity, apples fall to the ground and Earth orbits the sun because space-time is an active and malleable fabric. Massive bodies like the sun bend space-time. A planet that orbits a star is actually following a straight path through a curved space-time. This means we have to replace the Euclidean geometry we learned at school with the curved geometry developed by the 19th-century mathematician Bernhard Riemann.

Einstein's description of gravity has been confirmed by watching light from a distant star being bent around the sun during a total solar eclipse. This is a very different picture of a force from that given by the standard model of particle physics, which says that forces are carried by particles. Extending this idea would suggest that gravity is mediated by a force-carrying particle known as the graviton.

BACKGROUND IMAGE: ARSCIMED/SPL; ABOVE: DESY/NASA

THE ROAD TO UNIFICATION

Many attempts have been made to reconcile Einstein's theory of gravity with the quantum description of the other three forces of nature. The latest and most ambitious is called M-theory and it contains three radical ingredients: extra dimensions of space-time, supersymmetry, and extended objects called superstrings and membranes.

1. EXTRA DIMENSIONS

One of the earliest attempts at unifying the forces of nature was made in the 1920s, when German physicist Theodor Kaluza melded Einstein's gravitational theory with the electromagnetic theory of James Clerk Maxwell.

The universe we live in appears to have four dimensions. Space has three – right-left, forwards-backwards and up-down – and the fourth is time. Kaluza rewrote Einstein's theory as if there were five space-time dimensions. This gives the gravitational field some extra components which he thought could be interpreted as Maxwell's electromagnetic field. Amazingly, he showed that these extra components precisely matched Maxwell's equations. So electromagnetism comes for free if you are willing to buy a fifth dimension for gravity.

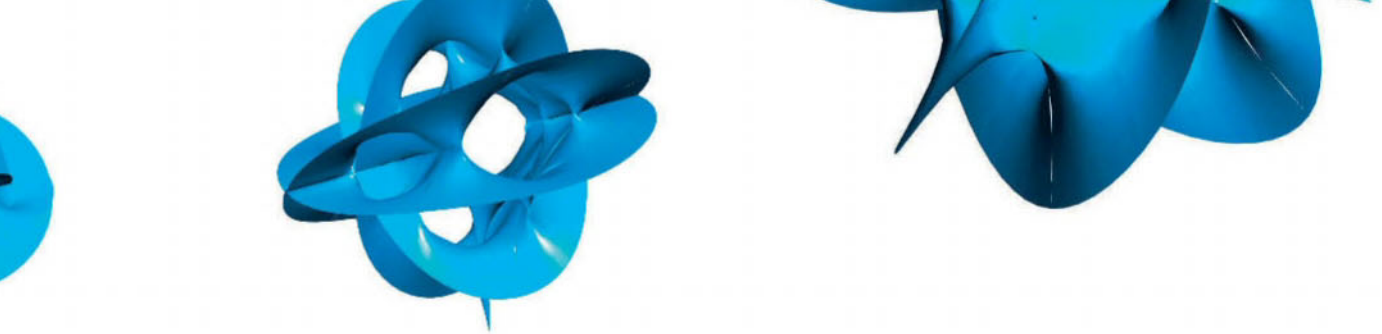
Why can't we see a fifth dimension? In 1926, Swedish physicist Oskar Klein came up with an answer. He supposed that the fifth dimension is not like the other four, but is instead curled up into a circle that is too small to see.

To see how this works, consider a simpler analogy: an ant on a tightrope. As well as walking along the tightrope, the ant can choose to walk around its circumference at any point. Only the ant is aware of the additional circular dimension. Viewed from a distance much, much larger than the ant's size, the rope looks very different: it is essentially a one-dimensional line and the extra dimension is hidden.

This is how Klein envisaged Kaluza's five-dimensional universe and his calculations even showed how small the extra dimension should be curled up. At 10^{-35} metres across, the fifth dimension is too small to probe even with the most powerful particle accelerators, which act as windows into the subatomic realm. Hence we have the impression that we live in a four-dimensional world.

Kaluza and Klein's idea lay dormant for many years. In some ways it was ahead of its time, partly because we knew so little about the weak and strong forces. It was revived by the arrival of supersymmetry.

ALEX WILD



2. SUPERSYMMETRY

The quarks and leptons that make up matter seem very different to the particles that carry nature's forces. So it came as a great surprise in the 1970s when theorists showed that it is possible to construct equations which stay the same when you swap the two around.

This suggests the existence of a new symmetry of nature. Just as a snowflake's underlying symmetry explains why it can look the same even after you rotate it, so the equivalence of particles is down to a new symmetry, called supersymmetry.

One prediction of supersymmetry is that every particle in the standard model has a supersymmetric partner, thereby doubling the number of particle species. Enormous energies are required to make a supersymmetric particle, which may be why no one has found one yet. Experiments at the powerful Large Hadron Collider at the CERN particle physics laboratory near Geneva, Switzerland, are looking for them. Finding one would rank among the biggest scientific discoveries of all time.

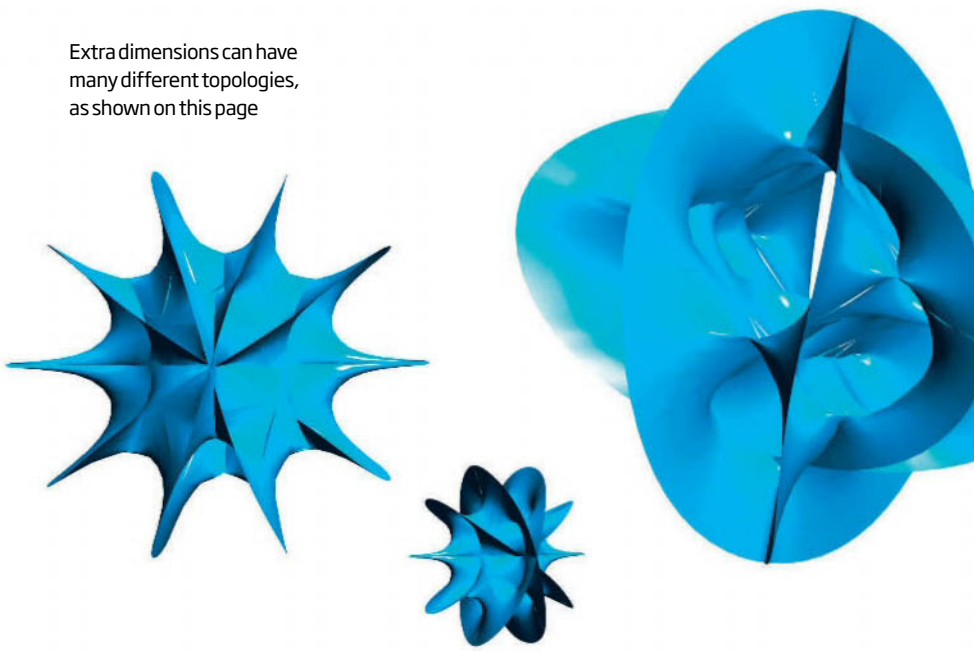
But there is a reason why theorists are so enamoured with supersymmetry despite 40 years without experimental evidence: it predicts gravity. According to the mathematics of supersymmetry, the act of turning an electron into its supersymmetric partner and back again is identical to moving it through space-time.

This means supersymmetry offers a connection between the properties of quantum particles and space-time, making it possible to incorporate gravity, too. The resulting theory that incorporates the gravitational force and supersymmetry is known as supergravity.

The mathematics of supergravity has an unexpected consequence: space-time can have no more than 11 dimensions. In the early 1980s this prompted a revival of the Kaluza-Klein idea, with up to seven curled-up dimensions. Could these extra dimensions describe the strong, weak and electromagnetic forces?

At first supergravity looked extremely promising, but problems crept in. For a start, 11-dimensional supergravity has trouble describing how quarks and electrons interact with the weak nuclear force. Even more serious is a problem that has dogged all other attempts to reconcile gravity and quantum field theory: when you use supergravity's equations to calculate certain quantum-mechanical processes, the answer is infinity. This makes no sense and is a sure sign that supergravity is at best only an approximation to a viable theory of everything. For these reasons, attention turned to a rival approach called superstring theory.

Extra dimensions can have many different topologies, as shown on this page

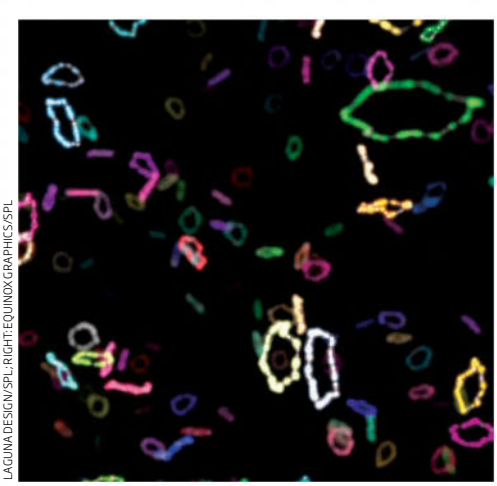


3. THE SUPERSTRING REVOLUTION

In superstring theory, the fundamental building blocks of matter are not point-like particles. Instead they are one-dimensional strings that live in a universe with 10 space-time dimensions. Just like violin strings, they can vibrate in various modes, each one representing a different elementary particle. Certain string vibrations can even describe gravitons, the hypothetical carriers of the gravitational force.

To begin with, superstring theory looked like a theorist's dream. The six extra dimensions could be curled up in such a way as to avoid the problems with the weak force encountered by 11-dimensional supergravity. Also, superstring theory looked just like general relativity when the graviton energy was set sufficiently small. But the most important feature was that the infinities and anomalies that had plagued previous attempts to apply quantum field theory to general relativity no longer existed.

Here, for the first time, was a consistent way to unify gravity with quantum mechanics. Theorists went wild. But after the initial euphoria, doubts began to set in.



LAGUNA DESIGN/SPL; RIGHT: EQUINOX GRAPHICS/SPL

Point-like particles have given way to strings

THEORY OF EVERYTHING

Our leading candidate for a theory of everything is known as M-theory. It grew from a merger of the two seemingly different approaches: 11-dimensional supergravity and 10-dimensional superstring theory. Could this be the final theory of everything?

BRANE POWER

Superstring theory had some serious shortcomings. One problem is that there is not one, but five, mathematically consistent superstring theories, each competing for the title of the theory of everything. We faced an embarrassment of riches.

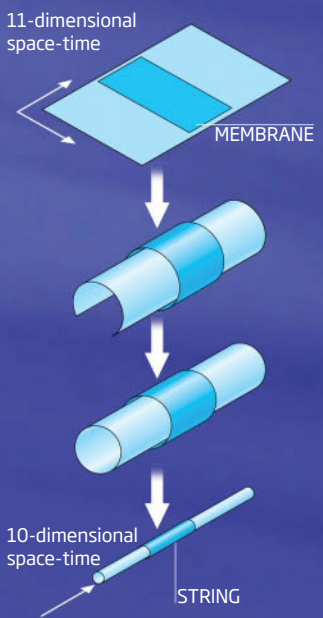
A second puzzle soon became apparent, too. Supersymmetry says that the universe has a maximum of 11 dimensions, yet the mathematics of superstring theory states there should be 10. What gives? And there was a related question: why stop at one-dimensional strings? Why not two-dimensional membranes which might take the form of a sheet or the surface of bubble?

It turns out that supersymmetry and membranes do go together. Just as superstrings live in 10 dimensions, it was calculated in 1987 that "supermembranes" can live in an 11-dimensional space-time dictated by supergravity.

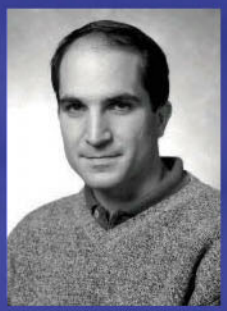
Moreover, if the 11th dimension is curled up, as Kaluza and Klein's early work suggested it could be, then it is possible to wrap the membrane around it. If curled up tightly enough, this wrapped membrane would look like a string in 10 dimensions.

Despite these attempts to revive 11 dimensions with the new ingredient of membranes, most string theorists remained sceptical. For many years there were two camps: string theorists with their 10-dimensional theory, and the membrane theorists working in 11 dimensions. It wasn't clear whether they were on the same page or not.

A membrane in 11 dimensions can be rolled up to appear as a string in 10 dimensions. The two are equivalent



In 1990, Edward Witten won the Fields medal, the mathematics equivalent of the Nobel prize. This shows just how closely mathematics and string theory tie together



Juan Maldacena's work showed that the physics inside a region of space can be described by what happens on its boundary. While his idea originated in M-theory, it has gone on to revolutionise many areas of theoretical physics, making Maldacena one of today's most influential physicists

THE M-THEORY REVOLUTION

All the work on strings, membranes and 11 dimensions was brought together in 1995 by Edward Witten, the string-theory guru at the Institute for Advance Study in Princeton, under one umbrella called M-theory. M, he says, stands for magic, mystery or membrane according to taste.

Witten showed that the five different string theories and 11-D supergravity were not rival theories at all. They were merely different facets of M-theory. Having one unique theory was a huge step forward. It also turned out that M-theory and its membranes were able to do things strings alone could not.

Take black holes, for example, which are excellent laboratories for testing our theories. In 1974, Stephen Hawking showed that black holes are not entirely black – instead they can radiate energy due to quantum effects. This means that black holes have temperature and another thermodynamic property called entropy, which is a measure of how disorganised a system is.

Hawking showed that a black hole's entropy depends on its area. Yet it should also be possible to work out its entropy by accounting for all the quantum states of the particles making up a black hole. However, all attempts to describe a black hole in this way had failed – until M-theory came along. Amazingly, M-theory exactly reproduces Hawking's entropy formula. This success gave us confidence that we were on the right track.

In 1998, Juan Maldacena, also of the Institute for Advanced Study, used membranes to explore what would happen inside a hypothetical universe with many dimensions of space and gravity. He showed that everything happening on the boundary of such a universe is equivalent to everything happening inside it: ordinary particles interacting on the boundary's surface correspond precisely to how membranes interact on the interior. When two mathematical approaches describe the same physics in this way, we call it a duality.

This duality is remarkable because the world on the surface of the universe looks so different to the world inside. If Maldacena's idea is applied to our universe, it could mean that we are just shadows on the boundary of a higher-dimensional universe.

Maldacena's paper has been cited over 11,000 times. This is partly because his idea has found

applications in unexpected areas of physics, including superconductivity and fluid mechanics, regardless of whether M-theory is the theory of everything or not.

More recently, my colleagues and I have found yet another area of physics to which M-theory can be applied: the black-hole/qubit correspondence. A classical bit is the basic unit of computer information and takes the value 0 or 1. A quantum bit, or qubit, can be both 0 and 1 at the same time. Only when we measure it do we fix which one it is, and the outcome cannot be predicted with certainty. This gives rise to the phenomenon of entanglement between two or more qubits, where measuring one qubit affects the other no matter how far apart they are. Einstein called this effect "spooky action at a distance".

For reasons we do not fully understand, the mathematics that describes qubit entanglement is exactly the same as that which governs certain black holes in M-theory. It turns out that these black holes fall into 31 classes, depending on their mass, charge and entropy. We recently used this to predict that four qubits can be entangled in 31 different ways. This can, in principle, be tested in the lab and we are urging experimentalists to find ways of doing just that.

BACKGROUND: PLAINPICTURE/AMANAIMAGES; LEFT: WILFRIED HOFFACKER/PLAINPICTURE

Events at the boundary of a universe reveal what is happening inside

A LANDSCAPE OF UNIVERSES

DANMCCOY/SCIENCEFACTION/CORBIS

Particles may be more like bubbles in a world with extra dimensions

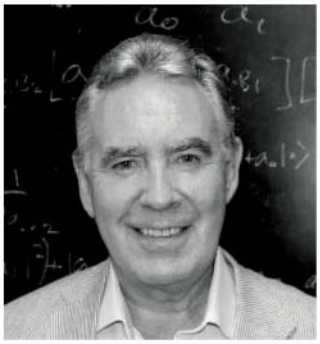
The geometrical and topological properties of the curled-up extra dimensions dictate the appearance of our four-dimensional world, including how many generations of quarks and leptons there are, which forces exist, and the masses of the elementary particles. A puzzling feature of M-theory is that there are many (possibly infinitely many) ways of curling up these dimensions, leading to a "multiverse" – a number of different universes. Some may look like ours, with three generations of quarks and leptons and four forces; many will not. But from a theoretical point of view they all seem plausible.

The traditional view is that there is one universe and a unique set of fundamental laws. The alternative view, which is gaining credibility, says that there are multiple universes out there with different laws of physics, and one of these universes just happens to be the one we are living in. Each of these universes must be taken seriously.

So is M-theory the final theory of everything? In common with rival attempts, falsifiable predictions are hard to come by. Some generic features such as supersymmetry or extra dimensions might show up at collider experiments or in astrophysical observations, but the variety of possibilities offered by the multiverse makes precise predictions difficult.

Are all the laws of nature we observe derivable from fundamental theory? Or are some mere accidents? The jury is still out.

In my opinion, many of the key issues will remain unresolved for quite some time. Finding a theory of everything is perhaps the most ambitious scientific undertaking in history. No one said it would be easy.



IMPERIAL COLLEGE LONDON

Michael Duff
Michael Duff is Emeritus Professor of Theoretical Physics at Imperial College London.

ANSWERING THE CRITICS

The job of theoretical physicists is twofold: first, to explain what our experimental colleagues have discovered; and second, to predict phenomena that have not yet been found. The history of scientific discovery shows that progress is achieved using both methods.

Quantum theory, for example, was largely driven by empirical results, whereas Einstein's general theory of relativity was a product of speculation and thought experiments, as well as advanced mathematics.

Speculation, then, is a vital part of the scientific process. When Paul Dirac wrote down his equation describing how quantum particles behave when they travel close to the speed of light, he wasn't just explaining the electron, whose properties had been well established in experiments. His equation also predicted the hitherto undreamed-of positron, and hence the whole concept of antimatter.

Such speculation is not a flight of fancy. It is always constrained by the straightjacket of mathematical consistency and compatibility with established laws. Even before it was tested experimentally, Einstein's theory of general relativity had to pass several theoretical tests. It had to yield special relativity and Newtonian mechanics in those areas where they were valid, as well as predict new phenomena in those where they were not.

It is a common fallacy that physics is only about what has already been confirmed in experiments. Commentators have unfairly compared the study of cosmic strings – macroscopic objects that may have been formed in the early universe – to UFOs and homeopathy, on the grounds that cosmic strings have yet to be observed. Others have stated that until M-theory is backed by empirical evidence, it is no better than "faith".

Yet support for superstrings and M-theory is based on their ability to absorb quantum mechanics and general relativity, to unify them in a mathematically rigorous fashion, and to suggest ways of accommodating and extending the standard models of particle physics and cosmology. No religion does that.

By the same token, some alternative ideas purporting to be theories of everything have had to be rejected even before their predictions could be tested – not on the grounds of faith but because they were mathematically erroneous. What separates theoretical speculation from faith is that we modify or reject theories in the light of new evidence and discovery.

The most effective way for critics of M-theory to win their case would be to come up with a better alternative. So far nobody has.

RECOMMENDED READING AND LINKS

"The membrane at the end of the universe" by Michael Duff and Christine Sutton, *New Scientist*, 30 June 1988, p 67

"The theory formerly known as strings" by Michael Duff, *Scientific American*, February 1998, p 64

"The illusion of gravity" by Juan Maldacena, *Scientific American*, November 2005, p 56

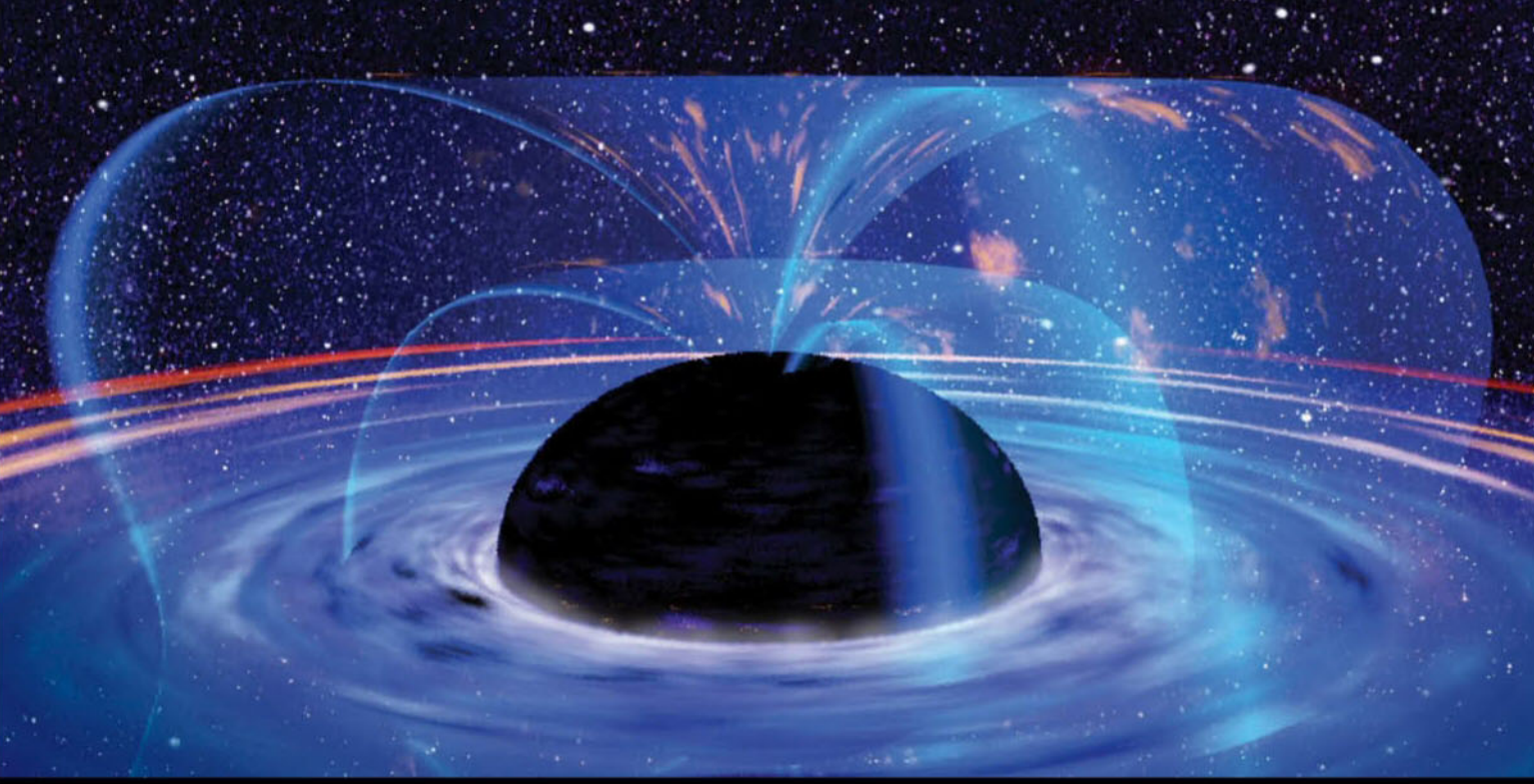
The Elegant Universe: Superstrings, hidden dimensions and the quest for the ultimate theory by Brian Greene (Vintage, 2005)

The Grand Design: New answers to the ultimate questions of life by Stephen Hawking and Leonard Mlodinow (Bantam Press, 2010)

"Four-bit entanglement from string theory" by Leron Borsten and others, arxiv.org/abs/1005.4915

The official string theory website superstringtheory.com

Cover image
Equinox Graphics/SPL



GENERAL RELATIVITY

PEDRO FERREIRA

INSTANT EXPERT

HISTORY OF GENERAL RELATIVITY

Albert Einstein's general theory of relativity is one of the towering achievements of 20th-century physics. Published in 1915, it explains that what we perceive as the force of gravity in fact arises from the curvature of space and time.

Einstein proposed that objects such as the sun and Earth change this geometry. In the presence of matter and energy, space-time can evolve, stretch and warp, forming ridges, mountains and valleys that cause bodies moving through it to zigzag and curve. So although Earth appears to be pulled towards the sun by gravity, there is no such force. It is simply the geometry of space-time around the sun telling Earth how to move.

The general theory of relativity has far-reaching consequences. It not only explains the motion of the planets; it can also describe the history and expansion of the universe, the physics of black holes and the bending of light from distant stars and galaxies.

GRAVITY BEFORE EINSTEIN

In 1686, Isaac Newton proposed an incredibly powerful theory of motion. At its core was the law of universal gravitation, which states that the force of gravity between two objects is proportional to each of their masses and inversely proportional to the square of their distance apart. Newton's law is universal because it can be applied to any situation where gravity is important: apples falling from trees, planets orbiting the sun, and many, many more.

For more than 200 years, Newton's theory of gravity was successfully used to predict the motions of celestial bodies and accurately describe the orbits of the planets in the solar system. Such was its power that in 1846 the French astronomer Urbain Le Verrier was able to use it to predict the existence of Neptune.

There was, however, one case where Newton's theory didn't seem to give the correct answer. Le Verrier measured Mercury's orbit with exquisite precision and found that it drifted by a tiny amount – less than one-hundredth of a degree over a century – relative to what would be expected from Newton's theory. The discrepancy between Newton's theory and Mercury's orbit was still unresolved at the beginning of the 20th century.

EINSTEIN'S INSIGHT

In 1905, at the age of 26, Albert Einstein proposed his special theory of relativity. The theory reconciled the physics of moving bodies developed by Galileo Galilei and Newton with the laws of electromagnetic radiation. It posits that the speed of light is always the same, irrespective of the motion of the person who measures it. Special relativity implies that space and time are intertwined to a degree never previously imagined.

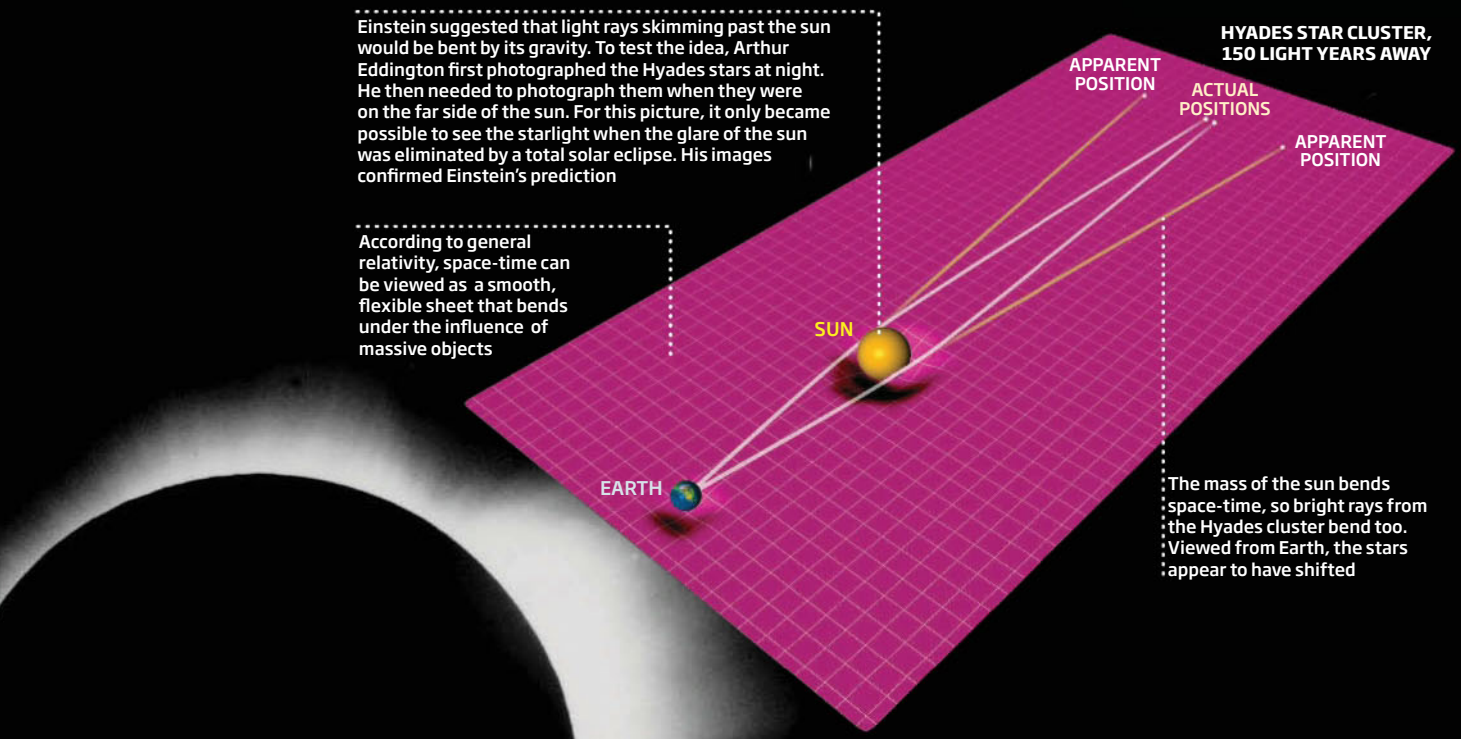
Starting in 1907, Einstein began trying to broaden special relativity to include gravity. His first breakthrough came when he was working in a patent office in Bern, Switzerland. "Suddenly a thought struck me," he recalled. "If a man falls freely, he would not feel his weight… This simple thought experiment… led me to the theory of gravity." He realised that there is a deep relationship between systems affected by gravity and ones that are accelerating.

SCIENCE PHOTO LIBRARY

The next big step forward came when Einstein was introduced to the mathematics of geometry developed by the 19th-century German mathematicians Carl Friedrich Gauss and Bernhard Riemann. Einstein applied their work to write down the equations that relate the geometry of space-time to the amount of energy that it contains. Now known as the Einstein field equations, and published in 1915, they supplanted Newton's law of universal gravitation and are still used today, a century later.

Using general relativity, Einstein made a series of predictions. He showed, for example, how his theory would lead to the observed drift in Mercury's orbit. He also predicted that a massive object, such as the sun, should distort the path taken by light passing close to it: in effect, the geometry of space should act as a lens (see diagram top right).

Einstein also argued that the wavelength of light emitted close to a massive body should be stretched, or red-shifted, as it climbs out of the warped space-time near the massive object. These three predictions are now called the three classical tests of general relativity.



"Space tells matter how to move and matter tells space how to curve"
John Archibald Wheeler

PHYSICIST, SUPERSTAR

In 1919, the English astronomer Arthur Eddington travelled to the island of Príncipe off the coast of west Africa to see if he could detect the lensing of light predicted by general relativity. His plan was to observe a bright cluster of stars called the Hyades as the sun passed in front of them, as seen from Earth. To see the starlight, Eddington needed a total solar eclipse to blot out the glare of the sun.

If Einstein's theory was correct, the positions of the stars in the Hyades would appear to shift by about 1/2000th of a degree.

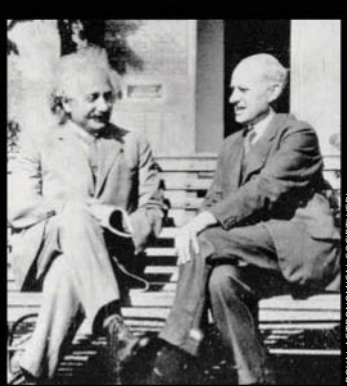
To pinpoint the position of the Hyades in the sky, Eddington first took a picture at night from Oxford. Then, on 29 May 1919, he photographed the Hyades as they lay almost directly behind the sun during the total eclipse that Príncipe experienced that day. Comparing the two measurements, Eddington was able to show that the shift was as Einstein had predicted and too large to be explained by Newton's theory.

Following the eclipse expedition, there was some controversy that Eddington's analysis had been biased towards general relativity. Matters were put to rest in the late 1970s when the photographic plates were analysed again and Eddington's analysis was shown to be correct.

Eddington's result turned Einstein into an international superstar: "Einstein's theory triumphs" was the headline of *The Times* of London. From then on, as more consequences of his theory have been discovered, general relativity has become entrenched in the popular imagination, with its descriptions of expanding universes and black holes.

In 1959, the American physicists Robert Pound and Glen Rebka measured the gravitational red-shifting of light in their laboratory at Harvard University, thereby confirming the last of the three classical tests of general relativity.

Images of the 1919 solar eclipse like the one above proved that gravity bends starlight


ROYAL ASTRONOMICAL SOCIETY/SPL

By 1930, Albert Einstein and Arthur Eddington were famous for their work on general relativity

HOW GENERAL RELATIVITY SHAPES OUR UNIVERSE

Einstein's general theory of relativity has revealed that the universe is an extreme place. We now know it was hot and dense and has been expanding for the past 13.8 billion years. It is also populated with incredibly warped regions of space-time called black holes that trap anything falling within their clutches.

BLACK HOLES

Shortly after Einstein proposed his general theory of relativity, a German physicist called Karl Schwarzschild found one of the first and most important solutions to Einstein's field equations. Now known as the Schwarzschild solution, it describes the geometry of space-time around extremely dense stars – and it has some very strange features.

For a start, right at the centre of such bodies, the curvature of space-time becomes infinite – forming a feature called a singularity. An even stranger feature is an invisible spherical surface, known as the event horizon, surrounding the singularity. Nothing, not even light, can escape the event horizon. You can almost think of the Schwarzschild singularity as a hole in the fabric of space-time.

In the 1960s, the New Zealand mathematician Roy Kerr discovered a more general class of solutions to Einstein's field equations. These describe dense objects that are spinning, and they are even more bizarre than Schwarzschild's solution.

The objects that Schwarzschild and Kerr's solutions describe are known as black holes. Although no black holes have been seen directly, there is overwhelming evidence that they exist. They are normally detected through the effect they have on nearby astrophysical bodies such as stars or gas.

The smallest black holes can be found paired up with normal stars. As a star orbits the black hole, it slowly sheds some of its material and emits X-rays. The first such black hole to be observed was Cygnus X-1, and there are now a number of well-measured X-ray binaries with black holes of about 10 times the mass of the sun.

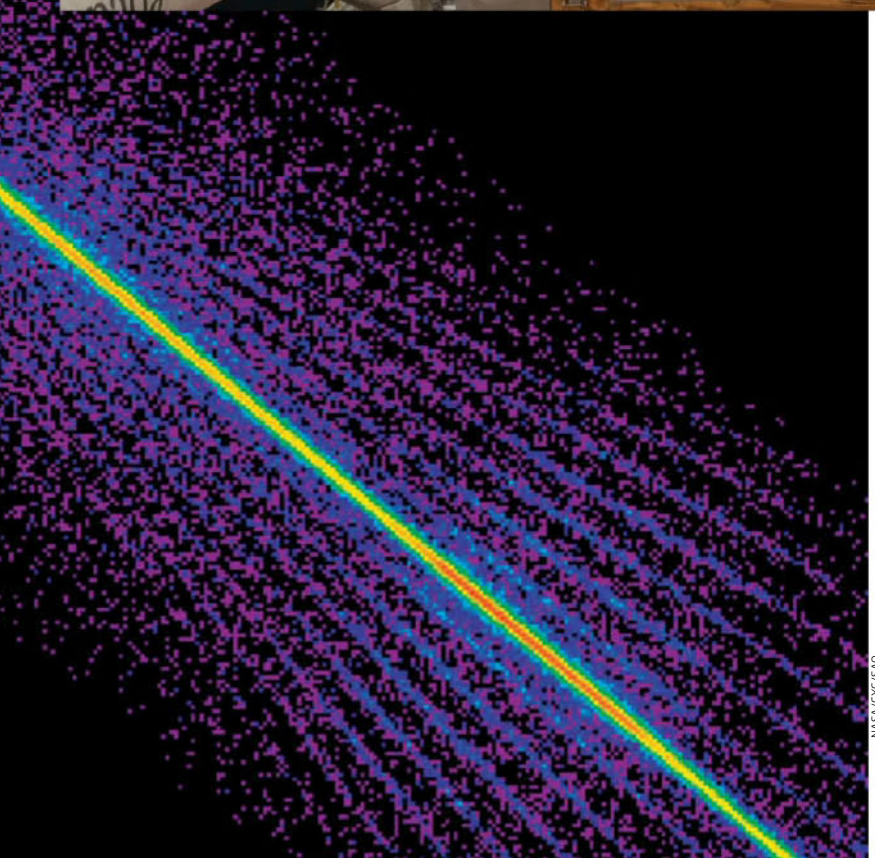
Evidence for much larger black holes came in the 1960s when a number of very bright and distant objects were observed in the sky. Known as quasars, they arise from the havoc black holes seem to create at the cores of galaxies. Gas at the centre of a galaxy forms a swirling disc as it is sucked into the black hole. Such is the strength of the black hole's pull that the swirling gas emits copious amounts of energy that can be seen many billions of light years away. Current estimates place these black holes at between a million and a billion times the mass of the sun. As a result, they are called supermassive black holes.

The evidence now points to there being a supermassive black hole at the centre of every galaxy, including our own. Indeed, observations of the orbits of stars near the centre of the Milky Way show that they are moving in very tightly bound orbits. These can be understood if the space-time they live in is deeply distorted by the presence of a supermassive black hole with more than 4 million times the mass of the sun.

Despite their names, British physicist Stephen Hawking has pointed out that black holes may not be completely black. He argues that, near the event horizon, the quantum creation of particles and antiparticles may lead to a very faint glow. This glow, which has become known as Hawking radiation, has not been detected yet because it is so faint. Yet, over time, Hawking radiation would be enough to remove all the energy and mass from a black hole, causing all black holes to eventually evaporate and disappear.

"No black holes have been seen directly yet, though there is overwhelming evidence that they exist"

NASA



NASA/CXC/SAO

Images from the Hubble Space Telescope (above) and Chandra X-ray Observatory have firmed up our relativity-based ideas about the universe

THE EXPANDING UNIVERSE

One of general relativity's most striking predictions arises if we consider what happens to the universe as a whole.

Shortly after Einstein published his theory, Russian meteorologist and mathematician Alexander Friedmann and Belgian priest Georges Lemaître showed that it predicted that the universe should evolve in response to all the energy it contains. They argued that the universe should start off small and dense, and expand and dilute with time. As a result, galaxies should drift away from each other.

Einstein was initially sceptical of Friedmann and Lemaître's conclusion, favouring a static universe. But a discovery by the American astronomer Edwin Hubble changed his mind.

Hubble analysed how galaxies recede from the Milky Way. He found that distant galaxies move away faster than those that are relatively nearby. Hubble's observations showed that the universe was indeed expanding. This model of the cosmos later became known as the big bang.

Over the past 20 years, a plethora of powerful observations by satellites and large telescopes have further firmed up the evidence for an expanding and evolving universe. We have obtained an accurate measure of the expansion rate of the universe and of the temperature of the "relic radiation" left over from the big bang, and we have been able to observe young galaxies when the universe was in its infancy. It is now accepted that the universe is about 13.8 billion years old.

FRONTIERS OF GENERAL RELATIVITY

General relativity predicts that the universe is full of exotic phenomena. Space-time can tremble like the surface of a pond and it seems to be full of a mysterious form of energy that is pushing it apart. It is also conceivable for space-time to be so warped that it becomes possible to travel backwards in time.

GRAVITATIONAL WAVES

According to general relativity, even empty space-time, devoid of stars and galaxies, can have a life of its own. Ripples known as gravitational waves can propagate across space in much the same way that ripples spread across the surface of a pond.

One of the remaining tests of general relativity is to measure gravitational waves directly. To this end, experimental physicists have built the Laser Interferometer Gravitational-Wave Observatory (LIGO) at Hanford, Washington, and Livingston, Louisiana. Each experiment consists of laser beams that are reflected between mirrors placed up to 4 kilometres apart. If a gravitational wave passes through, it will slightly distort space-time, leading to a shift in the laser beams. By monitoring time variations in the laser beams, it is possible to search for the effects of gravitational waves.

No one has yet detected a gravitational wave directly, but we do have indirect evidence that they exist. When pulsars orbit very dense stars, we expect them to emit a steady stream of gravitational waves, losing energy in the process so that their orbits gradually become smaller. Measurement of the decay of binary pulsars' orbits has confirmed that they do indeed lose energy and the best explanation is that these pulsars are losing energy in the form of gravitational waves.

Pulsars are not the only expected source of gravitational waves. The big bang should have created gravitational waves that still propagate through the cosmos as gentle ripples in space-time. These primordial gravitational waves are too faint to be detectable directly, but it should be possible to see their imprint on the relic radiation from the big bang – the cosmic microwave background. Experiments are now under way to search for these signs.

Gravitational waves should also be emitted when two black holes collide. As they spiral in towards each other, they should emit a burst of gravitational waves with a particular signature. Provided the collision is sufficiently close and violent, it may be possible to observe them with instruments on Earth.

A more ambitious project is the Evolved Laser Interferometer Space Antenna (eLISA), made up of a trio of satellites that will follow Earth in its orbit around the sun. They will emit precisely calibrated laser beams towards each other, much like LIGO. Any passing gravitational wave will slightly distort space-time and lead to a detectable shift in the laser beams. NASA and the European Space Agency hope to launch eLISA in the next decade.

The LIGO detectors (left) are looking for gravitational waves ringing through space

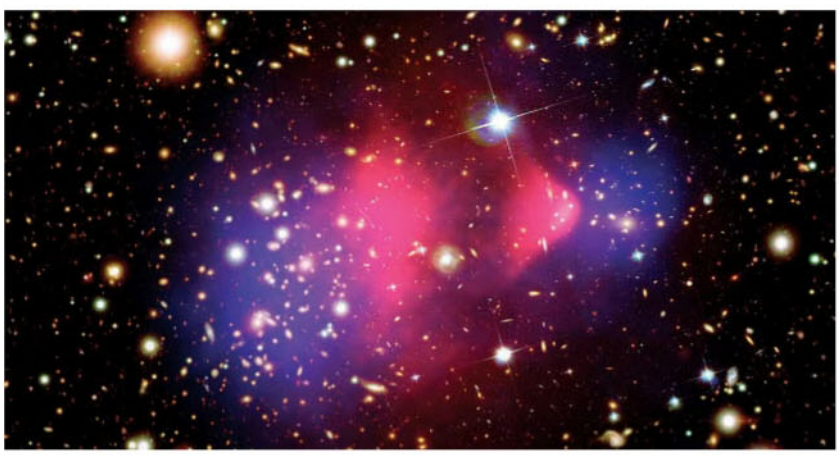
"It is possible to build tunnels linking different parts of space and different parts of time – in theory, at least"

TIME TRAVEL

Einstein's theory allows for the intriguing possibility of time travel. One suggested way of achieving this involves the construction of tunnels called wormholes that link different parts of space at different times. It is possible to build wormholes – in theory. But unfortunately they would require matter with negative energy, and other unnatural physical circumstances, not only to open them up but also to allow them to be traversed. Another possibility is to create a large region of space that rotates, or use hypothetical objects called cosmic strings.

The possibility of time travel can lead to physical paradoxes, such as the grandfather paradox in which the time traveller goes back in time and kills her grandfather before he has met her grandmother. As a result, one of her parents would not have been conceived and the time traveller herself would not exist. It has been argued, however, that physical paradoxes such as these are, in practice, impossible to create.

THE DARK UNIVERSE



The galaxy group called the Bullet cluster (right) gives good evidence for dark matter (added in blue)

HENZE/NASA; FAR LEFT: DAVE BULLOCK; RIGHT: CXC/CFA/M.MARKEVITCH ET AL.; MAGELLAN/U.ARIZONA/D.CLOWE ET AL.; ESO WFI; NASA/STSCIW

The expanding universe predicted by general relativity has become firmly entrenched in modern science. As our ability to observe distant galaxies and map out the cosmos has improved, our picture of the universe has revealed some even more exotic features.

For a start, astronomers have been able to measure how fast distant spiral galaxies spin, and this shows that the outskirts of galaxies are rotating far too quickly to be reined in by the mass of the stars and gas at their centres. More matter is needed in galaxies to generate enough gravity to prevent galaxies from flying apart.

The popular explanation is that galaxies contain large quantities of other forms of matter – known as "dark matter" because it does not emit or reflect light. Dark matter is thought to clump around galaxies and clusters of galaxies in gigantic balls known as halos. Dark matter halos can be dense enough to significantly distort space-time and bend the path of any light rays that pass close by. This gravitational lensing has been observed in a number of clusters of galaxies, and is one of the strongest pieces of evidence for the existence of dark matter.

But that's not all. Cosmologists have been able to figure out how fast the universe expanded at different times in its history. This is done by measuring the distance to exploding stars called supernovae, and how quickly they are receding due to the expansion of space-time. The ground-breaking results from these observations, which emerged around two decades ago, is that the expansion of the universe seems to be speeding up.

One explanation for this accelerating expansion is that the universe is permeated by an exotic form of energy, known as dark energy. Unlike ordinary matter and dark matter, which bend space-time in a way that draws masses together, dark energy pushes space apart, making it expand ever more quickly over time.

If we add together all the forms of matter and energy in the universe we end up with a striking conclusion: only 4 per cent of the universe is made up of the form of matter we are familiar with. Around 24 per cent is invisible dark matter and 72 per cent is dark energy.

This result emerged from the marriage of the general theory of relativity and modern astronomy and it has become a prime focus of physics. Experimenters and theorists are directing their efforts at trying to answer the burning questions: what exactly are dark matter and dark energy? And why do they have such strange properties?

ISABELLE MOURIÉS ADLER

Pedro Ferreira

Pedro Ferreira is professor of astrophysics at the University of Oxford. He works on the origin of large-scale structures in the universe, on the general theory of relativity and on the nature of dark matter and dark energy. He is author of *The Perfect Theory* (Abacus), a biography of general relativity.

THE BIG UNSOLVED PROBLEM
QUANTUM GRAVITY

General relativity is only one of the pillars of modern physics. The other is quantum mechanics, which describes what happens at the atomic and subatomic scale. Its modern incarnation, quantum field theory, has been spectacularly successful at describing and predicting the behaviour of fundamental particles and forces.

The main challenge now is to combine the two ideas into one overarching theory, to be known as quantum gravity. Such a theory would be crucial for explaining the first moments of the big bang, when the universe was dense, hot and small, or what happens near the singularity at the cores of black holes, where the effects of quantum physics may compete with those of general relativity.

Although there is as yet no final theory of quantum gravity, there are several candidate theories being actively explored. One is string theory, which describes the fundamental constituents of matter not as point-like particles but as microscopic vibrating strings. Depending on how they vibrate, the strings will be perceived as different particles – including the graviton, the particle thought to carry the gravitational force.

Another possibility is that space-time is not smooth but built up of discrete building blocks that interact with each other. As a result, if we were able to peer at its fine structure, it might look like a frothy space-time foam. In such theories, what we perceive as the space-time that bends and warps smoothly in the presence of matter is merely an emergent phenomenon masking more radical behaviour on small scales.

The quest for the theory of quantum gravity is arguably the biggest challenge facing modern physics. One of the difficulties is that it only really manifests itself at extremely high energies, well beyond our experimental reach. Physicists now face the task of devising experiments and astronomical observations that can test candidate theories of quantum gravity in the real world.

RECOMMENDED READING

The State of the Universe by Pedro G. Ferreira (Phoenix)

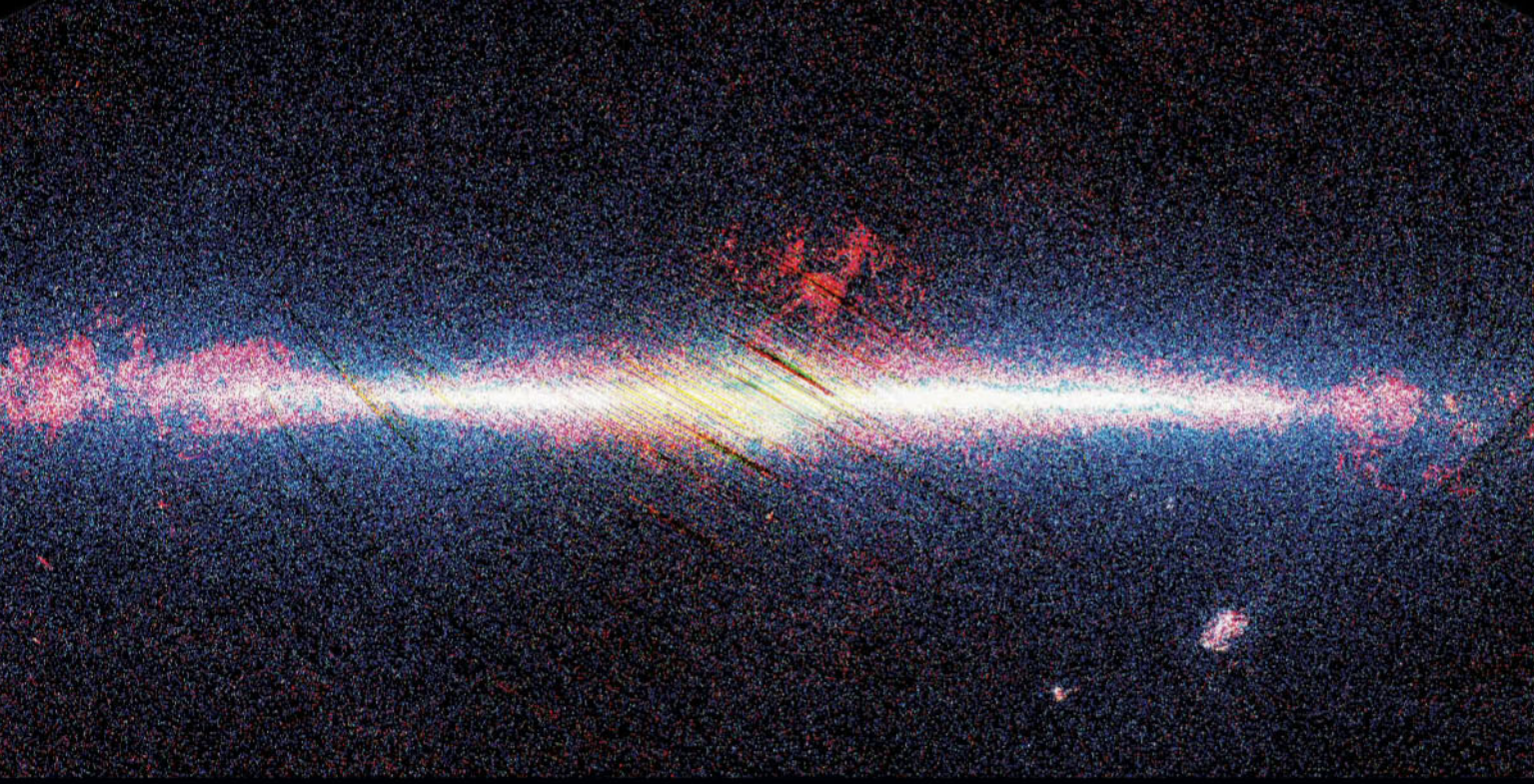
Black Holes and Time Warps: Einstein's Outrageous Legacy by Kip Thorne (Papermac)

Gravity: An Introduction to Einstein's General Theory of Relativity by James B. Hartle (Addison-Wesley)

Time Travel in Einstein's Universe by Richard Gott (Phoenix)

Was Einstein Right? Putting General Relativity to the Test by Clifford Will (Basic Books)

Cover image
XMM-Newton/ESA/NASA



UNSEEN UNIVERSE

MICHAEL ROWAN-ROBINSON

INSTANT
EXPERT

INFRARED ASTRONOMY

As we look into a clear night sky, we see just a fraction of what the universe contains: mainly stars in our galaxy radiating in the narrow visible wavelength band between 390 and 750 nanometres.

Optical telescopes extend that vision to far-off galaxies, but it is only in the past century or so, as we have begun to observe the broad sweep of invisible electromagnetic wavelengths, that the full drama of the cosmos has been unveiled.

The first invisible radiation to be detected was in the infrared range, at wavelengths from 750 nanometres up to a millimetre. It was discovered in 1800 when British astronomer William Herschel used a prism to split sunlight and saw the mercury of a thermometer placed beyond the red end of the spectrum begin to rise.

Infrared astronomy took off in the 1960s. It studies objects in the universe at temperatures between 10 and 1000 kelvin: asteroids, comets, interstellar dust and newly forming stars and galaxies.

DUST TO DUST

The most significant source of the infrared light that reaches Earth is the interstellar medium. This mixture of gas and dust pervades the space between stars in galaxies and has a temperature of 10 to 50 kelvin. It radiates only in the infrared, and dims the visible light from distant stars, reddening their colour.

The first direct image of the interstellar dust came in 1983 courtesy of the Infrared Astronomical Satellite (IRAS), a space telescope funded by the US, the Netherlands and the UK. It was a signal moment in astronomy. Observing interstellar dust allows us to glimpse the full cycle of stellar life and death, including the formation of new stars and planetary systems from the dust – sometimes in violent bouts as distant galaxies collide – long before these stars become visible to optical telescopes. A striking example lies in the pair of merging galaxies known as the Antennae, around 45 million light years from us: their brightest infrared regions (below right) are dark at visible wavelengths (above right).

Infrared observations also reveal dying stars blowing off clouds of dust and gas, replenishing the interstellar medium. The dust is mainly silicates and amorphous carbon – sand and soot. The production of this dust is crucial to our existence: every carbon atom in our bodies was created in the core of a star, was ejected as that star died, and drifted around in the interstellar medium before being sucked into our solar system.

NASA/ESA/HHT/STScI/AURA

Star-forming regions of the Antennae galaxies show up in this infrared Herschel image (right)

STAR INSTRUMENT: HERSCHEL

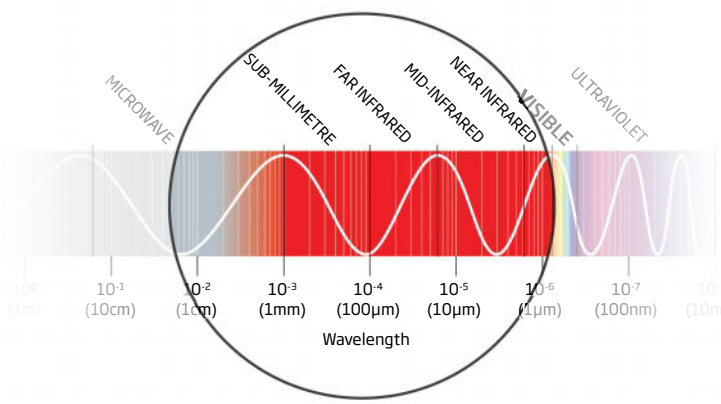
Most infrared wavelengths are absorbed by water and carbon dioxide in the atmosphere, with only a few narrow spectral "windows" of infrared reaching the ground. Infrared telescopes must therefore be situated at the top of mountains or, better still, in space.

Top dog in the infrared pack has been the European Space Agency's Herschel telescope, which operated from 2009 to 2013. It was the largest telescope ever launched into orbit, and carried a spectrometer and two cameras that covered wavelengths between 70 and 500 micrometres. All this equipment had to be cooled to temperatures close to absolute zero to prevent the telescope's own infrared emissions affecting the measurements.

The interpretation and follow-up of Herschel data carries on but the telescope has already delivered some spectacular images of filamentary interstellar dust clouds in which stars may be forming, as well as galaxies with unexpectedly large amounts of very cold dust missed by earlier studies.

The Herschel space telescope (right) was named after the founding father of infrared astronomy (below)

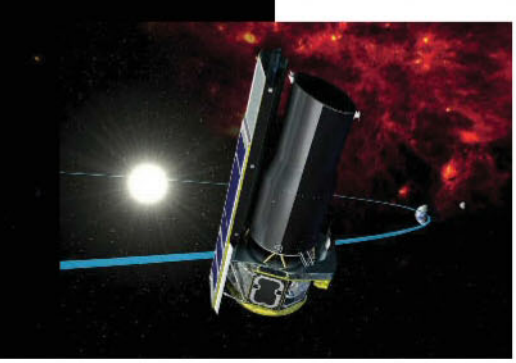
ML ATZ/AFP/GETTY IMAGES; LEFT: ABBOTT, LEMUEL FRANCIS (C.1760-1803) VICTORIA ART GALLERY BATH AND NORTH EAST SOMERSET COUNCIL/BRIDGEMAN ART LIBRARY



OTHER WORLDS

The first dedicated infrared space telescope, IRAS, found discs of dust and debris around some bright stars, pointing the way to searches for planetary systems. Infrared surveys have since detected many debris discs and planets in the process of forming.

Most fully formed extrasolar planets are found by optical telescopes looking at small changes in a star's velocity as a planet orbits it, or tiny drops in brightness as a planet crosses in front of the star. NASA's Kepler mission has found over 1000 exoplanets in this way. Infrared instruments, such as NASA's Spitzer Space Telescope (left), complement this approach. They look for "hot Jupiters", close-orbiting massive planets, as they pass in front of stars. An infrared instrument on the European Southern Observatory's Very Large Telescope gave the first direct image of an extrasolar planet. In orbit around a brown dwarf star, it is five times the mass of Jupiter.



NASA/JPL

GALACTIC ORIGINS

Because infrared observations spy out stars as they form and die, we can use them to look back in time, tracing how stars and galaxies formed throughout cosmic history almost as far back as the big bang.

When NASA's Cosmic Background Explorer (COBE) space mission, launched in 1999, measured the total background radiation at millimetre and sub-millimetre wavelengths, it found a strong contribution from distant galaxies. It turns out that more than half of the energy emitted by far-off stars at optical and ultraviolet wavelengths is absorbed by dust and re-emitted in the infrared before it reaches us, making infrared essential for our understanding of the universe.

The infrared is also important for finding out how galaxies first arose. The universe is expanding, which means most galaxies are receding from us and the radiation they emit undergoes a Doppler shift to longer wavelengths. This "red shift" means visible light from the most distant galaxies known, emitted in the first billion years after the big bang, is stretched to infrared wavelengths by the time it reaches us.

RADIO AND MICROWAVE ASTRONOMY

Radio and microwave telescopes study the longest electromagnetic wavelengths – anything longer than about a millimetre. Some of these emissions are produced by the coldest objects in the cosmos, such as the 2.7-kelvin background radiation from the big bang.

Most, however, are generated as "synchrotron radiation", given off when electrons spiral through magnetic fields at close to the speed of light. Identifying the sources of this radiation has revealed some of the universe's most extreme objects, such as pulsars and quasars.

THE COSMIC MICROWAVE BACKGROUND

TED THAI/TIME LIFE/GETTY

In 1965, while trying to make the first microwave observations of the Milky Way, Arno Penzias and Bob Wilson of Bell Labs in Holmdel, New Jersey, (below) found their instruments plagued by unexplained noise coming from all directions in the sky. This turned out to be one of the most important astronomical discoveries of the 20th century: the radiation left over from the big bang, known as the cosmic microwave background or CMB.

This radiation has a spectrum exactly like that of a body with a temperature of 2.73 kelvin, a spectacular confirmation of what the big bang theory predicts. Its strength is virtually identical no matter where you look: disregarding a systematic 1 in 1000 variation caused by our galaxy's motion through the cosmos, its intensity varies by no more than 1 part in 100,000.

These tiny fluctations are nonetheless important, as they provide information about the abundance of different types of mass and energy in the universe. Measurements of the CMB by the Wilkinson Microwave Anisotropy Probe (WMAP) suggest just 4 per cent of the universe is ordinary matter, while some 24 per cent is dark matter, presumed to be made of unknown particles, and 72 per cent is the even more perplexing dark energy.

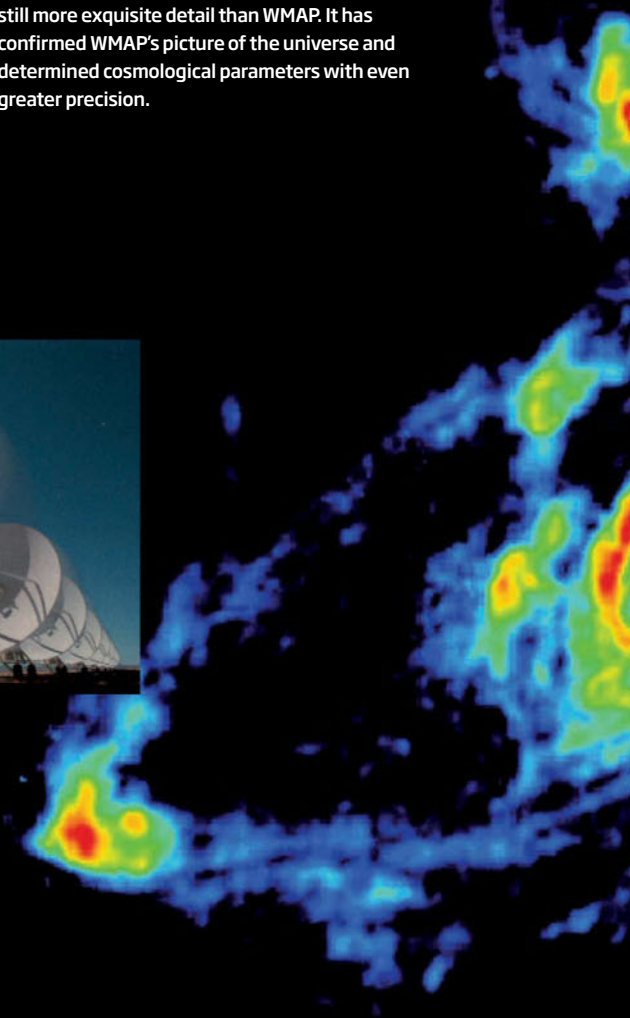
The European Space Agency's Planck Surveyor mission, launched in 2009 on the same rocket as the Herschel infrared telescope, mapped the CMB in still more exquisite detail than WMAP. It has confirmed WMAP's picture of the universe and determined cosmological parameters with even greater precision.

STAR INSTRUMENT: THE VERY LARGE ARRAY

The classic image of the radio telescope is of an overblown television satellite dish. Famous examples include the steerable telescopes at Jodrell Bank in the UK, the Parkes Observatory in New South Wales, Australia, and the National Radio Astronomy Observatory at Green Bank, West Virginia. The largest single dish of them all is the fixed 305-metre-diameter dish at Arecibo in Puerto Rico, which famously featured in the James Bond film *GoldenEye*.

Even such a monster cannot pinpoint a radio source in the sky to the desired accuracy, however. To make high-resolution observations, you need a dish hundreds of thousands of times bigger than the radio wavelengths you are observing. This is done by combining the signals from many scattered dishes using a technique called aperture synthesis. The prime example of such an instrument is the Very Large Array in New Mexico, which consists of 27 dishes spread along three arms of a "Y", each 10 kilometres long. It can locate a radio source in the sky to an accuracy of around a 1/10,000th of a degree.

Signals from the antennas of New Mexico's Very Large Array (above) are combined to make detailed radio images, like this one (right) of swirling hydrogen gas in the M81 galaxy group



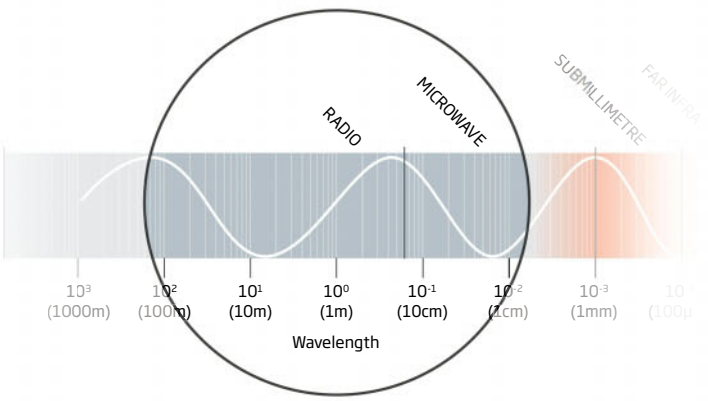
PULSARS

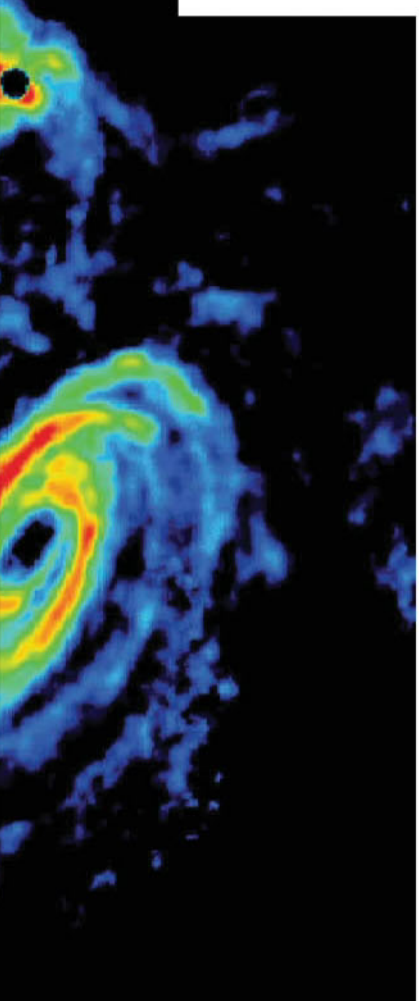
HENCOUP ENTERPRISES LTD/SCIENCE PHOTO LIBRARY

Discoverers of the first pulsar: Jocelyn Bell and Anthony Hewish

In 1967, Jocelyn Bell and Antony Hewish were studying emissions from quasars (see below) with a new radio antenna on the edge of Cambridge, UK, when Bell noted a pulsing radio signal repeating every second or so. It was the first of a new class of radio sources known as pulsars. These rapidly rotating neutron stars, the remnants of massive supernovas, have stupendous magnetic fields that can reach 10 gigateslas; Earth's field, by comparison, is a puny 50 microteslas. As they spin, pulsars emit synchrotron radiation in jets that sweep through space like a lighthouse beam, resulting in the pulsing signal seen by our telescopes.

Radio telescopes have found thousands of pulsars with periods ranging from a millisecond to several seconds. In 1974, the orbit of a pulsar in a binary system with an ordinary, non-pulsing neutron star was seen to be slowing down exactly as it would if it were emitting gravitational waves – the only indirect evidence we have so far of this key prediction of Einstein's general theory of relativity (see General Relativity, page 15).





ROGER RESSMEYER/CORBIS; FAR LEFT: NRAO/AUI

GALACTIC INTERACTIONS

Regular galaxies are suffused with hydrogen gas. As hydrogen atoms emit radio waves with a wavelength of 21 centimetres, radio telescopes can map this gas. Often it extends far beyond a galaxy's visible boundary and can even link objects that appear separate. An example is the M81 group of galaxies around 12 million light years away (pictured left). In an optical telescope, these galaxies seem distinct, but radio observations show a web of hydrogen connects them, through which they tug at each other gravitationally.

We can get a wealth of information on the internal dynamics of galaxies by looking at other spectral lines from interstellar gas molecules, for example in the microwave band, which lies between the radio and the infrared. Such observations reveal that dense molecular clouds have a rich chemistry, much of it based on carbon: more than 140 molecules have been identified, with carbon monoxide the most abundant after hydrogen.

QUASARS

The first isolated celestial source of radio waves, Cyg A in the constellation Cygnus, was identified as a distant galaxy in 1954. By 1962, astronomers at the University of Cambridge had listed over 300 radio sources in the northern sky.

A few of these were remnants of supernovae in our galaxy, including an object – now known to be a pulsar – at the heart of the Crab nebula, the remains of a supernova explosion seen by Chinese astronomers in AD 1054. Most, however, were within distant galaxies. Some were associated with objects that looked like stars, and became known as quasi-stellar radio sources, or quasars. What these luminous, compact objects were was long controversial. Today we believe them to be supermassive black holes at the centre of distant galaxies, with masses ranging from a million to a billion times that of the sun.

We now suspect that most galaxies, including our own, have a black hole at their heart, and that in radio galaxies and quasars this black hole is swallowing up the surrounding gas. As the gas spirals in towards the hole, magnetic field lines in the gas get wound up too, accelerating electrons and producing radio waves. More than 200,000 quasars are now known.

X-RAY AND GAMMA-RAY ASTRONOMY

X-rays and gamma rays are the most energetic electromagnetic waves, with wavelengths of a fraction of a nanometre or less.

Observations at these wavelengths show the universe at its hottest and most violent. This is a realm of gamma-ray bursts, of gas at temperatures of hundreds of millions of degrees swirling around the remnants of dead stars, and of fascinating objects such as white dwarfs, neutron stars and black holes.

DEATH STARS

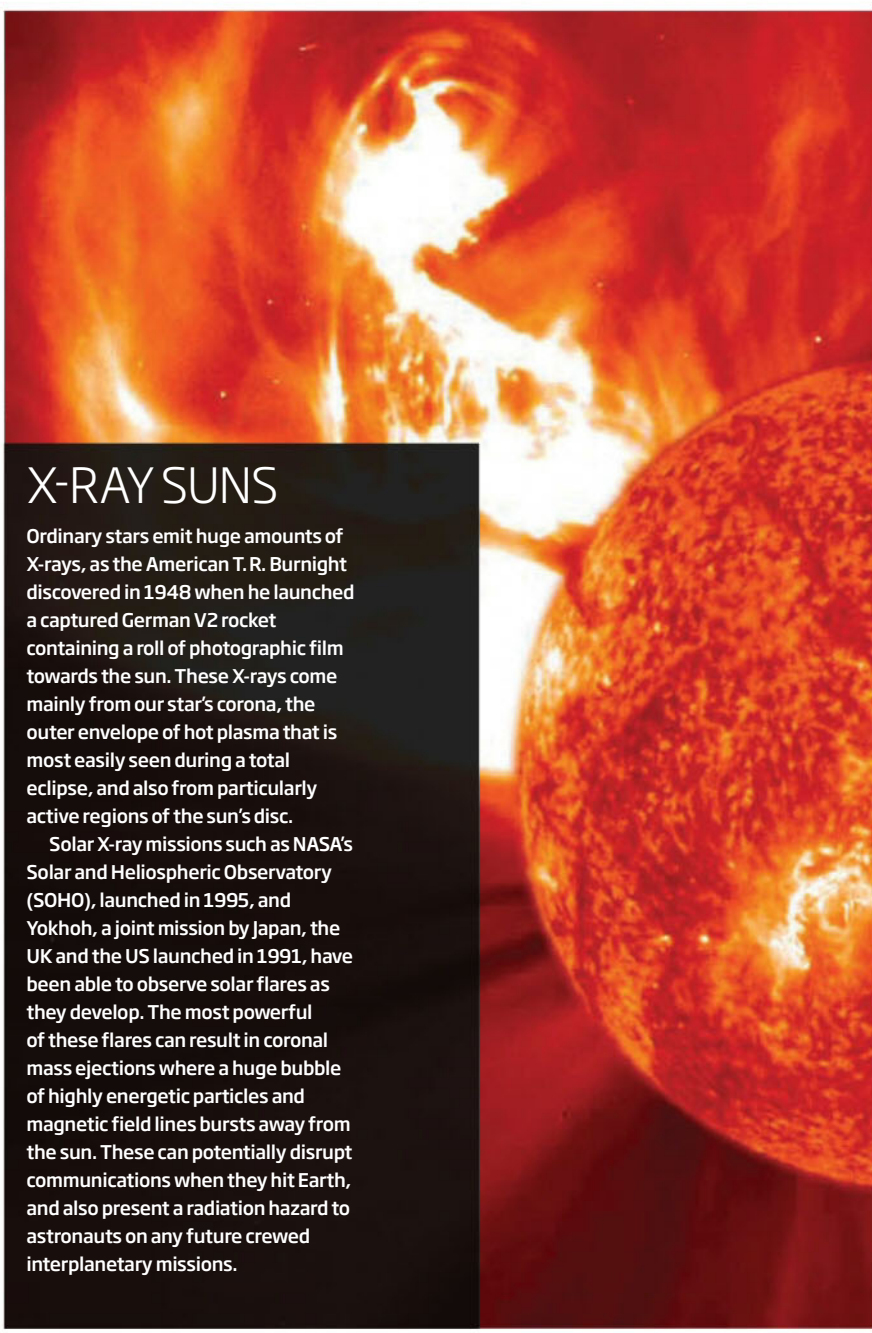
Cosmic X-rays are absorbed by oxygen and nitrogen in Earth's atmosphere, so X-ray telescopes must be put into orbit. The first compact X-ray source, Sco X-1 in the constellation of Scorpio, was found during rocket observations of the moon in 1962. In 1970, the first dedicated X-ray satellite, NASA's Uhuru, was launched.

Many X-ray sources are binary star systems in which gas being shed by a dying star spirals into its companion – a dead, compact remnant of what was once a star. As it does so, it heats up and emits X-rays.

In Sco X-1 the companion object is a neutron star, the remnant of a star 10 times the mass of our sun. Other systems have larger, white-dwarf companions. But measurements in 1971 of the unseen companion's orbital wobble in one X-ray source, Cyg X-1 in the constellation Cygnus (pictured below), showed it was too heavy for a white dwarf or neutron star. It had to be a black hole – the first observational evidence of the existence of such a body (see page 18).

X-rays are also emitted from the hot inner edges of discs of material accreting around supermassive black holes in active galactic centres and quasars (see page 27). Surveys by NASA's Chandra X-ray observatory and the European Space Agency's XMM-Newton satellite, both launched in 1999, have pinpointed thousands of such sources. One X-ray spectral line from highly ionised iron has been particularly informative: in some cases, it provides evidence of distortion due to the effects of general relativity.

NASA/CXC

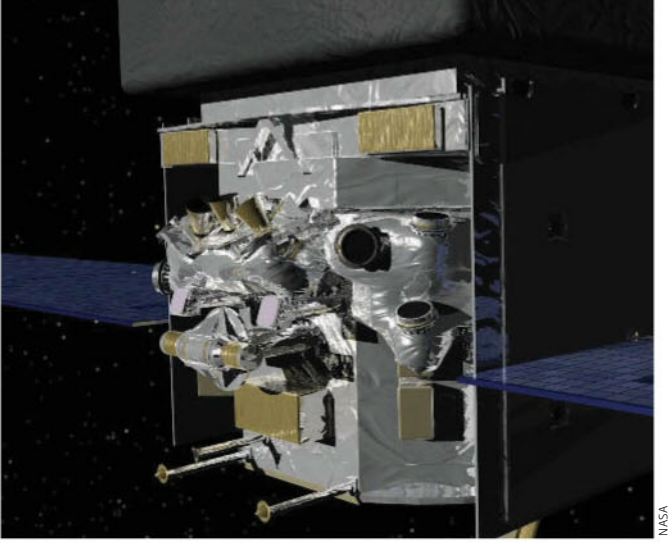


SOHO/NASA/ESA

X-RAY SUNS

Ordinary stars emit huge amounts of X-rays, as the American T. R. Burnight discovered in 1948 when he launched a captured German V2 rocket containing a roll of photographic film towards the sun. These X-rays come mainly from our star's corona, the outer envelope of hot plasma that is most easily seen during a total eclipse, and also from particularly active regions of the sun's disc.

Solar X-ray missions such as NASA's Solar and Heliospheric Observatory (SOHO), launched in 1995, and Yokhoh, a joint mission by Japan, the UK and the US launched in 1991, have been able to observe solar flares as they develop. The most powerful of these flares can result in coronal mass ejections where a huge bubble of highly energetic particles and magnetic field lines bursts away from the sun. These can potentially disrupt communications when they hit Earth, and also present a radiation hazard to astronauts on any future crewed interplanetary missions.



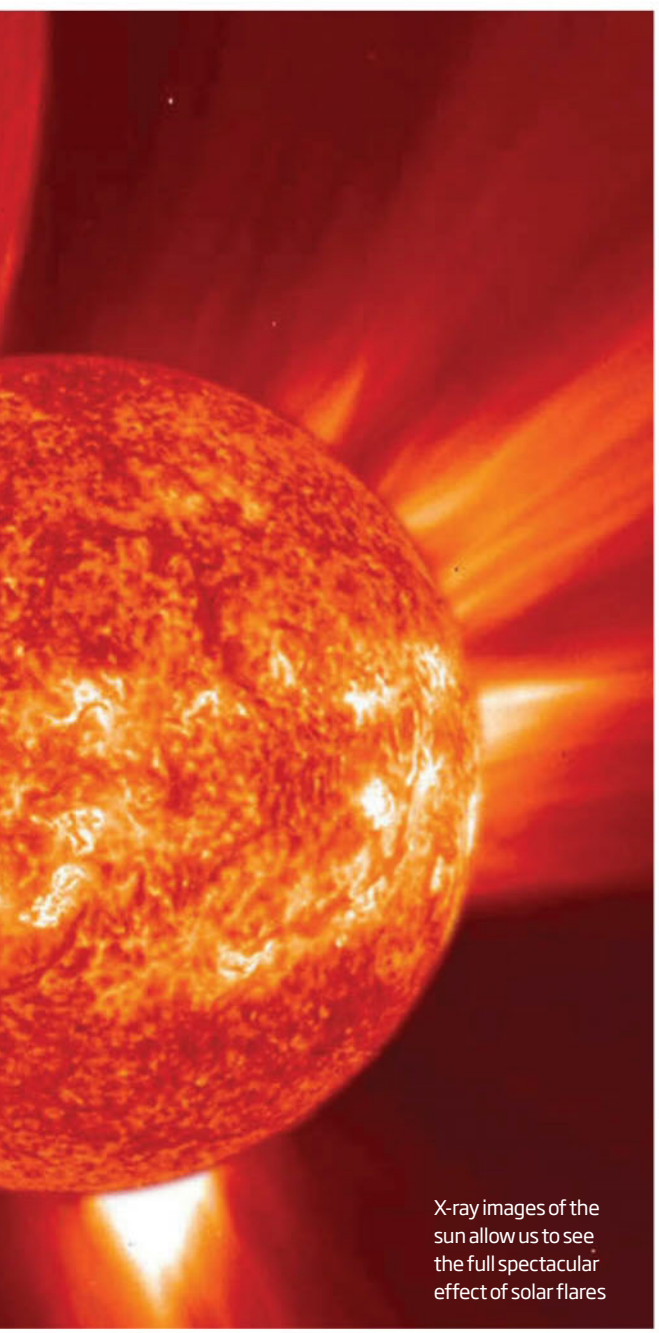
NASA

STAR INSTRUMENT: FERMI

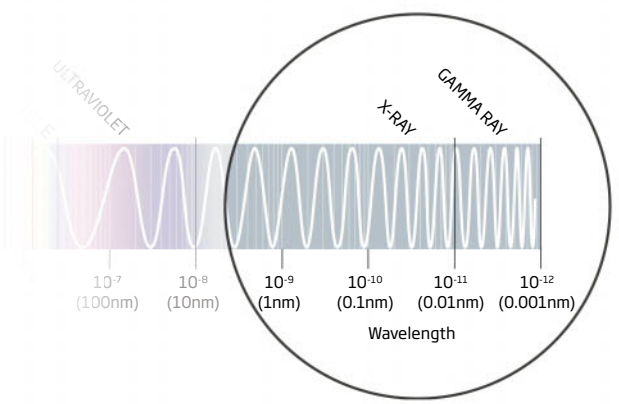
The international Fermi gamma-ray space telescope was launched in 2008. It has carried out a survey of the whole sky as well as studying gamma-ray bursts (see below), pinpointing their locations to within 1/60th of a degree.

Most of the gamma-ray sources are powered by supermassive black holes at the centre of galaxies, but Fermi has also studied pulsars, supernova remnants and the general background of gamma rays that emanates from all corners of the cosmos and whose origin is not fully understood.

Fermi might also detect interactions between the postulated dark-matter particles known as WIMPs, if they exist. And it will perform other tests of fundamental physics that can be carried out at these ultra-high energies, such as measuring whether the speed of light is the same at all wavelengths.



X-ray images of the sun allow us to see the full spectacular effect of solar flares



BURST ASTRONOMY

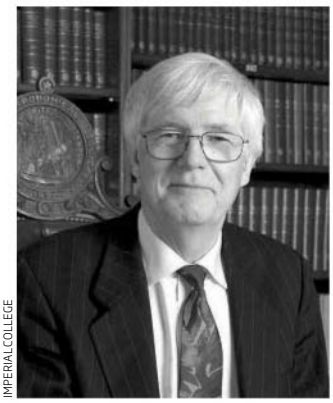
Gamma rays have wavelengths shorter than 0.01 nanometres and are emitted during radioactive decay, or by particles travelling near the speed of light. The first gamma-ray burst was detected in 1967 by satellites monitoring atmospheric nuclear weapons testing.

Most bursts probably occur when a massive, fast-spinning star collapses to form a black hole, sending out a narrow beam of intense radiation, while shorter bursts may be emitted when two neutron stars merge. Bursts typically last a few seconds, with a longer-lived X-ray and optical afterglow, but can release as much energy as our sun will radiate in a 10-billion-year lifetime. They are visible even from the edge of the visible universe: rays have been seen from a galaxy 13 billion light years away, meaning they were emitted just 600 million years after the big bang.

As with X-rays, gamma rays are absorbed by Earth's atmosphere. A dedicated space mission, NASA's SWIFT telescope, has studied more than 800 bursts since it was launched in 2004, while ground-based instruments such as HESS in Namibia, MAGIC in the Canary Islands and VERITAS in Arizona keep an eye out for light from showers of short-lived subatomic particles created when energetic gamma rays collide with atoms in Earth's atmosphere.

The HESS telescope spies gamma-ray bursts from the Namibian veld

HESS



IMPERIAL COLLEGE

Michael Rowan-Robinson

Michael Rowan-Robinson is professor of astrophysics at Imperial College London. He works principally on infrared and sub-millimetre astronomy, and cosmology. He contributed to the IRAS, ISO and Spitzer infrared space missions, and is now involved with both the Herschel and Planck projects. He has been writing for *New Scientist* for over 40 years.

THE FUTURE OF THE UNSEEN UNIVERSE

The coming years will see more of the invisible universe revealed by existing instruments and new probes spanning all wavelengths.

The workhorse of astronomy today, the Hubble space telescope, will continue to operate until at least 2018, at which time its successor, the James Webb Space Telescope, should be ready for launch. The JWST will operate mainly in the infrared, covering wavelengths from 500 nanometres to 24 micrometres. Its main aim will be to obtain images of Earth-sized planets and to detect the very first galaxies at the edge of the observable universe. Towards 2025, SPICA, a joint Japanese-European infrared space telescope, should also be well advanced, together with a slew of giant ground-based optical and near-infrared telescopes – the European Extremely Large Telescope, the Thirty-Metre Telescope and the Giant Magellan Telescope.

The Atacama Large Millimeter Array (ALMA) spans wavelengths from 0.4 to 3 millimetres and came on stream in Chile in 2011. It is probing star-forming regions in our galaxy and others with exacting angular resolution and sensitivity.

Even ALMA will be surpassed in scale, though, by an international radio telescope known as the Square Kilometre Array (SKA). To be sited in South Africa and Australia, it will connect a dense central square kilometre of radio antennas with receiving stations up to 3000 kilometres away. Ambitions for SKA are mind-blowing: it will study cosmic evolution and the nature of dark matter and dark energy through observations of hydrogen gas in a billion galaxies, and perform fundamental measurements to test our understanding of gravity and detect gravitational waves.

At the X-ray end of the spectrum, NASA and the European Space Agency are investigating the feasibility of an X-ray Observatory, called Athena. If it goes ahead, Athena will peer through dust and obscuring clouds of gas to discover and map supermassive black holes back at times when galaxies were first forming, and uncover the history of matter and energy, both visible and dark. It will also investigate when and how the elements were created and how they became dispersed in the intergalactic medium.

RECOMMENDED READING

Night Vision by Michael Rowan-Robinson (Princeton University Press, to be published late 2010)

Finding the Big Bang by P. J. E. Peebles, L. A. Page Jr and R. B. Partridge (Cambridge University Press)

Cover image: JAXA/ESA

CHAPTER TWO

BECOMING HUMAN

HUMAN ORIGINS

TIM WHITE

INSTANT
EXPERT

SEARCH FOR OUR ORIGINS, FROM DARWIN TO TODAY

Charles Darwin's only remark about human evolution in his seminal work *On The Origin of Species* was that "light will be thrown on the origin of man and his history". In his autobiography, Darwin justifies his brevity: "It would have been useless and injurious… to have paraded, without giving any evidence, my conviction with respect to his origin." His boldest statement was in *The Descent of Man*, where he concluded: "It is somewhat more probable that our early progenitors lived on the African continent than elsewhere." Today, thanks to a range of discoveries and technologies, we can tell in amazing detail the story that Darwin only guessed at.

THE BIG PICTURE

Twelve million years ago, Earth was a planet of the apes. Fossil evidence shows there were many ape species spread across the Old World, from Namibia to Germany to China. About 7 million years ago, a long-gone African species whose fossils have yet to be found was the last common ancestor shared by humans and our closest living relatives, the chimpanzees. By 6 million years ago, a daughter genus had evolved primitive bipedality and smaller canines. Some 2 million years later, its descendants had extended their range across Africa. After another million years, one of the species in the genus *Australopithecus* sparked a technological revolution based on stone tool manufacture that helped to push later hominids beyond Africa and across Europe and Asia.

The genus *Homo* is the group of species that includes modern people as well as the first hominids to have left Africa. The first species of the genus to do this, *Homo erectus*, rapidly spread from Africa into Eurasia by 1.8 million years ago, reaching Indonesia and Spain, though this was still long before the ice ages began. Many cycles of cold and nearly a million years later, another African descendent of *Homo erectus* – one that would eventually vaingloriously name itself *Homo sapiens* – again ventured beyond the continent. It has now reached the moon, and perhaps soon, will stand on a neighbouring planet.

Not bad for a two-legged primate.

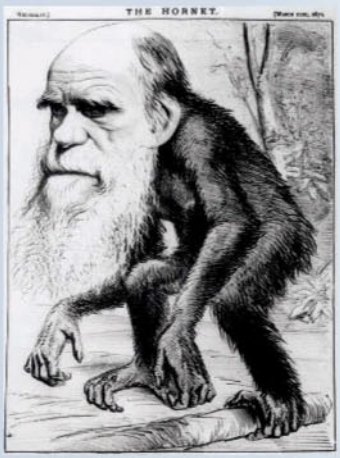
Reconstruction of the skull of *Ardipithecus ramidus*

NOT FROM CHIMPS

Nineteenth-century sceptics illustrated what many people saw as the implausibility of human evolution with a cartoon depicting Darwin's head atop the body of a knuckle-walking chimpanzee. Even though Darwin was clear from the start that we had not evolved from living chimpanzees, similar ideas, and the "missing link" concept, have stuck with us.

Darwin's champion, Thomas Huxley, concluded from his own anatomical studies of African apes that they were our closest living relatives, a conclusion vindicated when molecular studies showed – and continue to show – how genetically close these animals are to us. Ironically, Darwin was almost alone in calling for restraint in the use of modern primates as "stand-in", proxy ancestors.

The recent discovery of human ancestors that were quite unlike chimps, dating from soon after the two lines split, has shown that his caution was well founded, and how living chimps have evolved a great deal in relation to the common ancestor that we once shared with them.



ARDIPITHECUS: THE WOODLAND HOMINID

We still lack enough fossils to say much about the very earliest hominids. The key features of the fossils that have been found suggest that they walked on two legs. We know their social system was different from that of any other living or fossil ape because the canines of males were much smaller and blunter than those of non-human apes, and so did not function as weapons.

African fossils of these earliest hominids from about 6 million years ago have been given different names: *Sahelanthropus tchadensis*, found in Chad; *Orrorin tugenensis* from Kenya; and *Ardipithecus kadabba* from Ethiopia. None of these resembles modern apes, and all share anatomical features only with later *Australopithecus*.

Before these fossils were found, many researchers had predicted that we would keep finding *Australopithecus*-like hominids all the way back to the fork between hominids and the evolutionary line leading to modern chimpanzees. The discovery of a skeleton of *Ardipithecus ramidus* from Ethiopian deposits dated at 4.4 million years upset all of those expectations because it is so different from even the most primitive *Australopithecus*.

The partial skeleton, nicknamed "Ardi", suggests that our last common ancestor with chimpanzees was not a halfway-house between a chimpanzee and a human, but rather a creature that lacked many of the specialisations seen in our closest cousins, such as knuckle-walking, a fruit-based diet, male-male combat and acrobatic arboreality. *A. ramidus* was a mosaic organism: partly bipedal, omnivorous with small canines, relatively little difference between the sexes and a preference for woodland habitats. Ardi represents the first phase of hominid evolution.

TIM WHITE/A VENERABLE ORANGOUTANG, FROM 'THE HORNET' / PRIVATE COLLECTION / THE BRIDGEMAN ART LIBRARY / AGNES GARAUDEL

WHAT'S IN A NAME?

Ever since Darwin, all non-human primates more closely related to humans than to our closest living relatives, the chimpanzees, have been placed in the zoological family Hominidae. The finding that humans and African apes are genetically very similar has met with calls to change this classification, grouping apes and humans into a single family. This means that "hominids" would then include chimpanzees and gorillas, while humans and their ancestors would be classified at the subfamily or tribal level as "hominins".

Whatever arbitrary name we choose to apply to our branch of the primate tree, the branch itself dates back to around 7 million years ago, when a species of ape whose fossils we have not yet found split into two branches. Because of this, I prefer the stability and clarity of continuing to classify all the members of the human clade (on our side of the last common ancestor we shared with chimps) as "hominid".

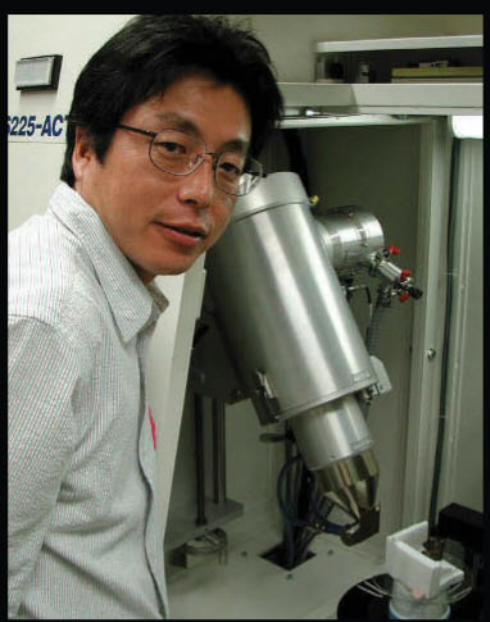
EVOLVING TECHNOLOGY

Technology has transformed our search for human origins. The majority of methods used to date the rocks that hold fossilised bones and artefacts are based on radioactive decay. For example, radioisotopic dating of the layers of volcanic ash sandwiching the remains of "Ardi", the partial skeleton of *Ardipithecus ramidus* (see left), show that the sediments in which the skeleton was found were laid down 4.4 million years ago.

Using micro-computed tomography (micro-CT), we can peer inside fossils without damaging them. In the case of Ardi, 5000 micro-CT "slices" through the fragments of her squashed and scattered skull allowed a team at the University of Tokyo in Japan to assemble a virtual model and then "print" the skull on a 3D printer.

Other technologies that have had a huge impact include differential GPS to map our finds with sub-metre accuracy and to pinpoint the location of ancient stone tool quarries, satellite imagery to identify surface outcrops of ancient sediments and image-stabilised binoculars to examine those outcrops from afar.

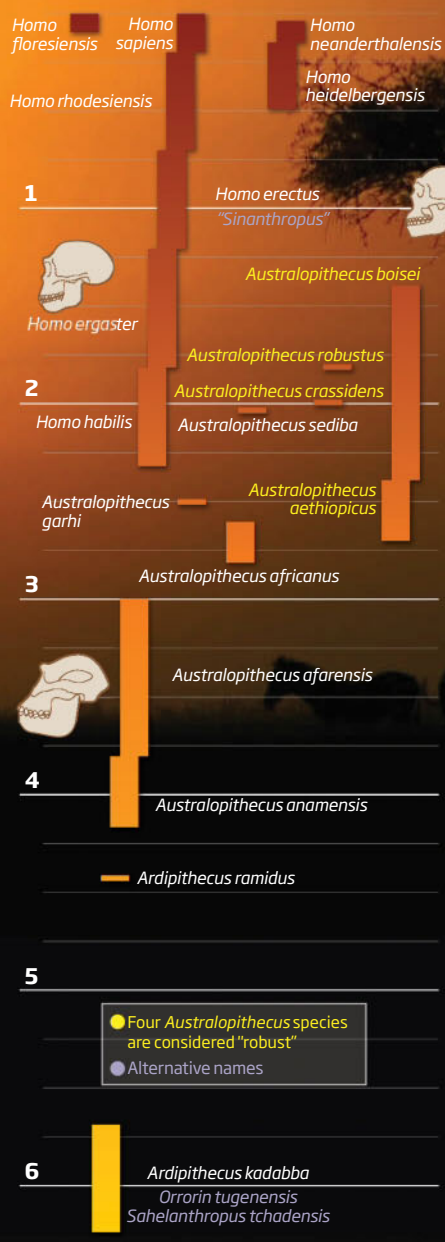
We use mass spectrometers to examine the soil around any animals we find and also measure the isotopic composition of their tooth enamel. This helps us determine their environment and diet. We use digitisers to capture and analyse the shape of fossils. We can even match the chemical fingerprints of rocks thousands of kilometres apart. For example, we have matched volcanic ash from the Middle Awash, our study area in Ethiopia's Afar Depression, to ash outcrops in other sites in Africa and to volcanic layers in deep-sea cores from the Gulf of Aden. The archaeopalaeontology tool kit has come a long way from little hammers and brushes.



Gen Suwa, who scanned and restored "Ardi" at the University of Tokyo

AFRICAN ORIGINS

A mountain of evidence has accumulated showing that our ancestors emerged in Africa. What is less clear-cut is what spurred their evolution. The answer lies in the environments in which our predecessors lived, and the influence of technology, which hugely expanded their ecological niche.



CACTUS OR BUSH?

The late American palaeontologist Steven Jay Gould wrote a classic essay in 1977 in which he predicted that the hominid family tree would prove to be "bushy". Today, it is common to see lists of more than 25 different hominid species, and Gould's prediction is often declared to be fulfilled.

Not so fast. Many of these species are "chronospecies", which evolve from one to the other, such as the earliest two *Australopithecus* species, *A. afarensis* and *A. anamensis*. These names are merely arbitrary divisions of a single evolving lineage.

A modern biologist addressing the question of species diversity counts the number of related species existing at any one time. When we do the same thing across the hominid fossil record, what we get is not a bush but something like a saguaro cactus, with only a few branches or species lineages. Indeed, the greatest diversity among hominid species appears to be at around 2 million years ago, when as many as four different lineages "briefly" co-existed in Africa.

The key question turns out to be not how many species there were *per se*, but rather why species diversity has been so limited on our branch of the evolutionary tree compared with other mammals like fruit bats or South American monkeys? The reason is probably that our ancestors' niche kept broadening, as a woodland omnivore 6 million years ago expanded ecologically into more open environments, and then again as technology further extended its capability and horizons.

ROBUST *AUSTRALOPITHECUS*

Look at the skull on the right. Would you say it looked robust? That's what palaeontologist Robert Broom thought when it was found in South Africa in the late 1930s, naming this hominid *Paranthropus robustus*. It has oversized molars, tiny canines and incisors, a massive lower jaw, dished face, small brain and, usually, a bony crest atop its skull. It came to be known as "robust" *Australopithecus*, and it appears in the fossil record more than 2.5 million years ago, in eastern Africa, with its last members some 1.2 million years ago. By that late date, our genus, *Homo*, had been on the scene for more than a million years. There are many mysteries about robust *Australopithecus* still to be solved, but one thing is clear: this side of 2.5 million years ago, our lineage was not alone.

TECHNOLOGICAL PRIMATE

Hominids are frustratingly rare in the fossil record, but at some time around 2.6 million years ago they began to leave calling cards, in the form of stone artefacts.

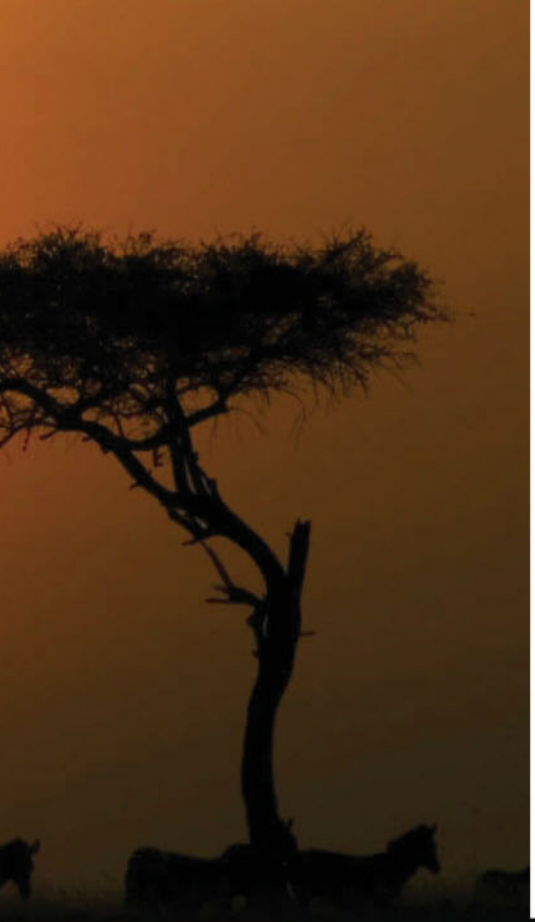
At the adjacent archaeological sites of the Gona and Middle Awash in Ethiopia, there is now abundant and unambiguous evidence of the earliest stone tools made by hominids. The fossilised bones of large mammals bear definite traces of marks made by sharp instruments.

The production of sharp-edged stone flakes enabled hominids to eat large amounts of meat and marrow previously unavailable to primates. At the same time, the selective pressures associated with such activities – particularly for a bipedal primate operating cooperatively under the noses of abundant predators, from hyenas to sabre-toothed cats – would lead to dramatic anatomical change as the braincase enlarged in *Homo*.

Stone technology greatly widened our ancestors' ecological niche, as well as their geographic range, enabling *H. erectus* to reach Europe and Indonesia more than 1.5 million years ago.

TIM WHITE

Stone tool from Gona, Ethiopia, made about 2.6 million years ago



LUCY: DAVID L. BRILL; BACKGROUND: JAMES HAGER/ROBERT HARDING/GETTY; SKULL: JAVIER TRUEBA/MSF/SPL

Skull of *Australopithecus robustus* (left), from Swartkrans, South Africa

"Lucy", the fossil of *Australopithecus afarensis*, is 3.2 million years old (right)

THE SAVANNAH HYPOTHESIS

Many modern palaeoanthropologists invoke climate change as the motor for our evolution. But they are hardly the first to recognise the impact of the environment. Long before relevant fossils were found, an early proponent of evolution, Jean-Baptiste Lamarck, saw grasslands as pivotal in the evolution of our ancestors from tree dwellers to bipeds. He was followed by Raymond Dart in the 1920s, who argued correctly that the fossil child he named *Australopithecus* was adapted to open environments.

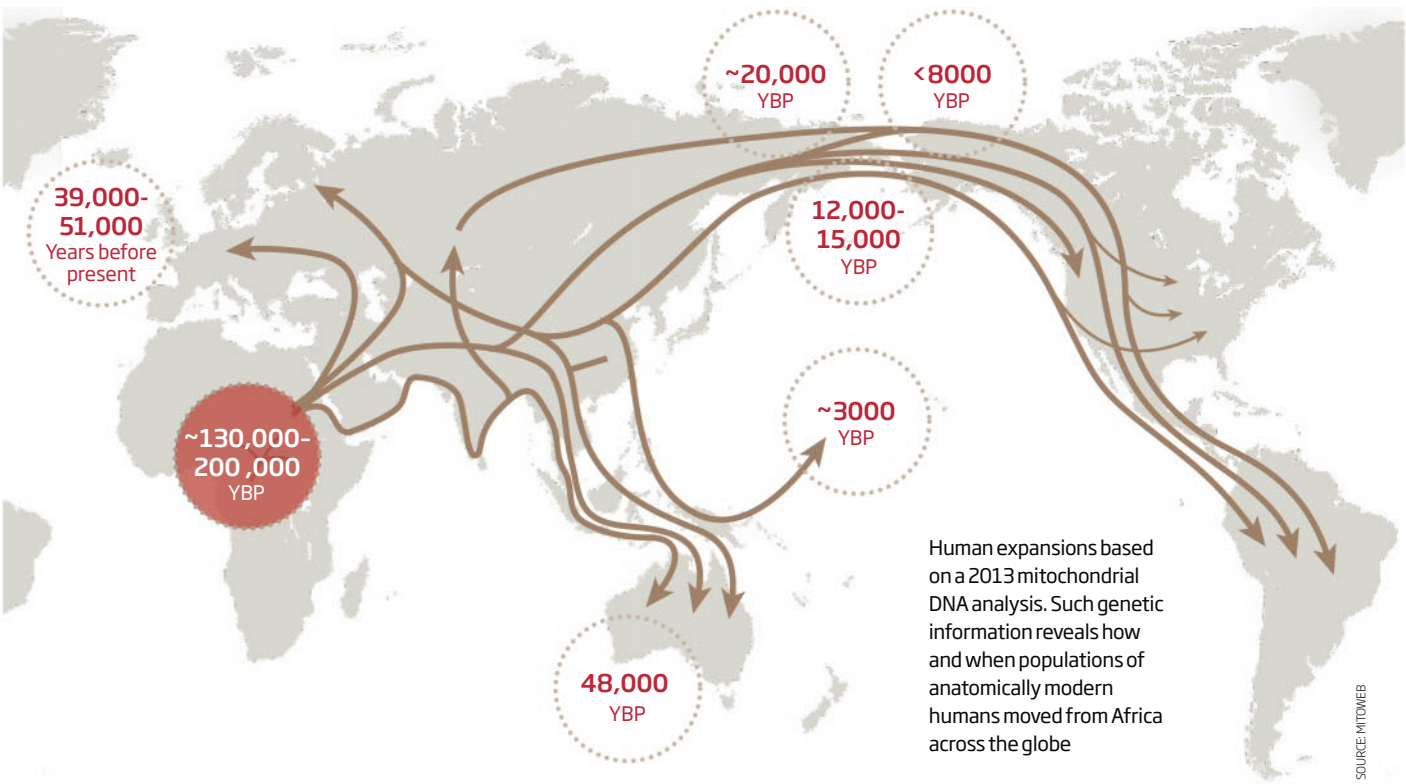
But the popularity of the "savannah hypothesis" began to wane in the 1990s, when *Ardipithecus* fossils were found in contexts suggesting a woodland habitat. Today independent lines of evidence suggest that the earliest hominids were indeed woodland creatures: climbing adaptations; diet as deduced from the shape, wear and isotopic composition of teeth; and the thousands of plants, insects, snails, birds and mammals that also prefer such habitats and are abundant in the same localities.

***Australopithecus* came next, though, and does appear to have been associated with more open landscapes. It has been known since the 1940s that the hip, knee and foot of *Australopithecus* were adapted to bipedality. However, it was the discovery of the "Lucy" fossils (see left) in Ethiopia and fossilised footprints in Tanzania during the 1970s that established this genus as representative of the evolutionary phase from which later hominids evolved. By 3 million years ago, *Australopithecus* species had managed to spread from north to south across much of Africa.**

To 20th-century anthropologists, *Australopithecus* seemed like an unstable transition between ape and human. Now, however, in the light of the *Ardipithecus* discoveries, this genus is seen as a long-lasting phase of our evolution. As well as gaining the means for habitual two-footed walking, robust forms became adapted to heavy chewing and developed relatively large back teeth with thicker enamel (see "Robust *Australopithecus*", left). It seems likely that some contemporary but less robust species eventually gave rise to the *Homo* genus.

RISE OF THE MODERN MIND

The initial hominid expansion from Africa occurred about 2 million years ago, long before the Neanderthals had evolved in Europe. The direct ancestors of our species spread out from Africa much later, after they had already become anatomically and behaviourally much more human. In Asia and Europe they would encounter populations of hominid species from earlier migrations that had evolved their own differences. These species became extinct, while the new hominids from Africa went on to evolve relatively superficial features that today characterise the geographically diverse populations of our species.



Human expansions based on a 2013 mitochondrial DNA analysis. Such genetic information reveals how and when populations of anatomically modern humans moved from Africa across the globe

SOURCE: MITOMAP

OUT OF AFRICA, TWICE

The first hominid expansion from Africa came about 2 million years ago, as revealed by stone tools and an outstanding collection of hominid fossils at the site of Dmanisi in Georgia. This expansion has sometimes been called "Out of Africa, Part 1", but the implication that hominids ever deserted Africa is manifestly incorrect. This continent continued to be the crucible of our evolution. Even the emigrant *Homo erectus* and its hand-axe technology are ubiquitous in Africa, with evidence of the species' occupation from the Cape to near Cairo.

Darwin predicted that Africa would one day yield fossils to illuminate human evolution. Today, he would be delighted to learn we have found fossils not only from the first two phases of human evolution, but also within our own genus, *Homo*. The earliest is *Homo habilis*, makers of stone flakes and cores that dominated technology for almost a million years. Next came *H. erectus*. What is clear is that our ancestors continued to evolve in Africa as more northerly latitudes were repeatedly buried in thick ice.

By 160,000 years ago, African hominids were nearly anatomically modern, with faces a little taller than ours, and skulls a little more robust. Their brain sizes were fully modern. In Ethiopia, at a locality called Herto by the local Afar people, the crania of two adults and a child represent some of the best evidence of the anatomy of these early people, who lived by a lake. Among their activities was the butchery of hippopotamus carcasses with their sophisticated stone tool kits.

Herto humans were also doing things that we would recognise as distinctively human: they were practicing mortuary rituals. Fine cut marks and polishing on a child's cranium show that it was defleshed when fresh, and then repeatedly handled.

Examination of the DNA of people today shows we all carry inside us a kind of "living fossil" that opens a window on our past. Whether modern human DNA samples are taken in the Arctic or the Congo, our DNA is remarkably similar to each other's, especially when compared with the variation seen in most other mammals. And the variation observed is greatest among African populations.

What this means is that we are a recent species, and that the ancestors of all modern people were Africans.

NEANDERTHAL FATES

REUTERS/NIKOLA SOLIC

Ever since their discovery in the mid-19th century, the place of Neanderthals in human evolution has been a mystery. Early evolutionists adopted them as evidence of human evolution, but as more and more fossils were recovered from around the Mediterranean, it became clear that these forms were peculiar hominids. With the excavation of further sites in Europe, the archaeological record showed a rapid technological change just as they disappeared.

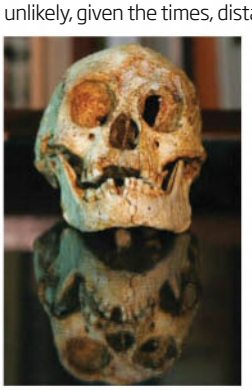
Debate about whether Neanderthals were ancestors or cousins persisted for decades, but fossil discoveries and genetics have finally solved this problem. Early anatomically near-modern and modern people lived in Africa long before the Neanderthals perished about 35,000 years ago. Genomic studies suggest that there was slight interbreeding between them, with leakage – of at most a few per cent – of genes from Neanderthals into human populations. They were our evolutionary close cousins, but the equivalent of a separately evolving species.

HOW TO SPOT A 'HOBBIT'

The discovery of the remains of diminutive humans on the Indonesian island of Flores, east of Java, captured the world's imagination in 2003. The remains were named *Homo floresiensis* and the people were, almost inevitably, nicknamed "hobbits".

Three hypotheses have been put forward to explain these Flores Island fossil hominids, which date from between 90,000 and 18,000 years ago. One is that their heads were abnormally small as a result of a congenital condition. However, no good match has been found between this microcephaly and a modern developmental disturbance of this kind.

The second hypothesis sees a very early occupation of Flores by hominids who were small, with small brains. In other words, far-flung *Australopithecus* or very early *Homo*. This also seems unlikely, given the times, distances, geographies and anatomies.



REUTERS/BEAWIHARTA

The third, most likely, scenario is that nearby *H. erectus* or *H. sapiens* became established there, rapidly evolving into hobbits via the well-known phenomenon of island dwarfing. All researchers agree that more evidence is needed to solve the mystery of *H. floresiensis*.

ADVANCE OF CIVILISATION

It has long been dogma that it was only when people domesticated plants and animals – otherwise known as agriculture – that they were able to settle down and begin to build cities and create monumental architecture. This plausible version of technological evolution is now being strongly challenged by discoveries in the Middle East.

At the site of Göbekli Tepe in south-east Turkey, an 11,000-year-old site has recently been uncovered. It boasts T-shaped limestone pillars of various heights, up to 6 metres, carved with images of animals. They were erected here in a monumental circular arrangement, 20 metres in diameter. This structure pre-dates the domestication of plants and animals. The people who built it still lived by hunting and gathering.

This and other sites discovered in the region challenge the notion that agriculture was the catalyst for what we loosely call "civilisation". Could it be that symbolism, ritual and religion came first and that these were the cause, rather than the consequence, of domestication and agriculture?

BERTHOLD STEINHILBER/LAIF/CAMERA PRESS

Göbekli Tepe in Turkey, with its limestone pillars and carved images

LANGUAGE, SYMBOLISM, AGRICULTURE... AND BEYOND

When did humans acquire language? It is a question that anthropologists and linguists still puzzle over. Some suggest it was very late, only after we had become *H. sapiens* and begun to spread beyond Africa sometime after the emergence of anatomically modern humans around 60,000 years ago. The founding of basic languages may have accelerated trade and, as the writer Matt Ridley argues in *The Rational Optimist*, trade is to culture as sex is to biology.

The first evidence of symbolic behaviour comes as 100,000-year-old South African shell beads and ochre incised with designs. Around 10,000 years ago, in the Fertile Crescent of Mesopotamia and the Levant, came the planned sowing and harvesting of plants.

Those of us old enough to remember the Apollo spacecraft or the dial telephone have witnessed so much technological innovation within a couple of generations that we find it difficult to appreciate that this speed of change is exceptional. For thousands of years, Babylonian farmers continued to do the same things, with the same tools, as their great-grandfathers had done.

TIM WHITE

Tim D. White
Professor of integrative biology and director of the Human Evolution Research Center at the University of California, Berkeley.

LESSONS FROM AN AFRICAN VALLEY

Set in Africa's north-east corner, the Afar rift holds a unique record of hominid history. Seasonal floods have carried sediments into this basin, where they have been accumulating for millennia. Every year, we risk our Middle Awash field camp being inundated. And it happens. When the floodwaters evaporate, the millimetre of silt left behind becomes the youngest stratum in a succession of sediment layers that is today about 1.5 kilometres thick.

These layers have been accumulating for 6 million years as rivers and lakes came and went. Near the bottom of this succession of rocks is *Ardipithecus* – not a chimpanzee, and not a human – but from the earliest portion of our branch of the hominid family tree.

In 4.2 million-year-old strata we found the earliest *Australopithecus*, followed by remains of the "Lucy" species in sandstones that are 3.4 million years old. Above this, in sediments 2.5 million years old, are traces of the butchery of large mammals accompanied by some of the earliest stone tools.

One million years ago, this valley was populated by hand-axe-making *Homo erectus*, which evolved into *Homo rhodesiensis* and then into the nearly anatomically modern *Homo sapiens idaltu*. In some of the youngest strata, we find fossils so anatomically modern that they could be lost among the 7 billion of us on Earth today. In these layered rocks we also find an unparalleled record of stone-tool technology.

As Darwin would surely appreciate, this evidence is overwhelming – the mammalian species we call *H. sapiens* has deep evolutionary roots in Africa.

Why does it matter? Human evolutionary history has important lessons for our species. We now know that all of our closest relatives have gone extinct, leaving only more distant African apes. The perspective that this knowledge provides is both timely and essential to the bipedal, large-brained, innovative, technological primate whose grasping hands now hold the power to determine our future on planet Earth.

Given the facts, it would not be wise to gamble on the widely held but risky notion that our future will be guided to good ends by divine intervention. Having evolved the capacity to influence the global future, it is high time the species begins to act sapiently.

RECOMMENDED READING

The Complete World of Human Evolution by Chris Stringer and Peter Andrews (Thames & Hudson)

The First Human: The race to discover our earliest ancestors by Ann Gibbons (Anchor)

Science, Evolution and Creationism by National Academy of Sciences (National Academies Press)

Evolution vs. Creationism: An introduction by Eugenie C. Scott (University of California Press)

Why Evolution is True by Jerry A. Coyne (Penguin)

Evolution Since Darwin: The first 150 years by Michael Bell, Douglas Futuyma, Walter Eanes and Jeffrey Levinton (Sinauer Associates)

Cover image: Tim White

THE EVOLUTION OF LANGUAGE

W. TECUMSEH FITCH

INSTANT EXPERT

HOW DO WE STUDY LANGUAGE EVOLUTION?

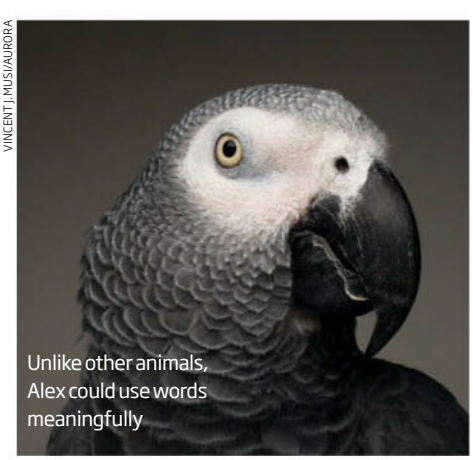
Humans have wondered about the origins of our unique capacity for language since the beginnings of history, proffering countless mythic explanations. Scientific study began in 1871 with Darwin's writings on the topic in *The Descent of Man*. For nearly a century afterwards, however, most writing on the subject was highly speculative and the entire issue was viewed with distrust by reputable scholars.

Recently, we have moved towards specific, testable hypotheses. Because language does not fossilise, only indirect evidence about key past events is available. But the situation is no worse off in this respect than cosmology or many other mature empirical sciences and, as with these other disciplines, scientists studying the evolution of language now combine many sources of data to constrain their theories.

One of the most promising approaches compares the linguistic behaviour of humans with the communication and cognition of other animals, which highlights shared abilities and the characteristics that make human language unique. The comparisons allow us to build theories about how these individual traits might have evolved.

Linguists define language as any system which allows the free and unfettered expression of thoughts into signals, and the complementary interpretation of such signals back into thoughts. This sets human language apart from all other animal communication systems, which can express just a limited set of signals. A dog's barks, for example, may provide important information about the dog (how large or excited it is) or the outside world (that an intruder is present), but the dog cannot relate the story of its puppyhood, or express the route of its daily walk.

For all its uniqueness, human language does share certain traits with many animal communication systems. A vervet monkey, for example, produces different calls according to the predators it encounters. Other vervets understand and respond accordingly – running for cover when a call signals an "eagle", for example, and scaling the trees when it makes a "leopard" call. This characteristic, known as functional referentiality, is an important feature of language. Unlike human languages, however, the vervets' system is innate rather than learned. This makes their system inflexible, so they cannot create a new alarm call to represent a human with a gun, for example. What's more, vervets do not seem to intentionally transmit novel information: they will continue producing leopard calls even when their whole group has moved to the safety of the trees. Thus, although the vervet communication system shares one important trait with human language, it still lacks many other important features.



VINCENT J. MUSI/AURORA

Unlike other animals, Alex could use words meaningfully

Building time lines

Similarly, the honeybees' complex dance routine offers some parallels with human language. By moving in certain ways, bees can communicate the location of distant flowers, water and additional hive sites to their hivemates – a system that is clearly functionally referential. More importantly, the bees are also communicating about things that aren't present. Linguists call this characteristic "displacement", and it is very unusual in animal communication – even vervets can't do this. Nonetheless, since bees can't communicate the full range of what they know, such as the colour of a flower, their system cannot be considered a language.

Looking at shared traits helps biologists to work out how those traits might have first evolved. Different animals might exhibit the same features simply because a common ancestor had the trait, which then persisted throughout the course of evolution. Such traits are called "homologies". Obvious examples include hair in mammals or feathers in birds. Alternatively, similar traits can evolve independently without being present in a common ancestor, a process called convergent evolution. The emergence of wings in both birds and bats is an example of this kind of evolution, as is the displacement seen in the bee's dance and human language.

Homologies allow us to build a time line of when different features first evolved. The fact that fish, mammals, birds, reptiles and amphibians all have skeletons, for example, suggests that bones evolved before lungs, which most fish lack but the other groups all share. Comparing creatures with convergent traits, by contrast, helps identify the common selection pressures that might have pushed the different species to evolve the trait independently.

This "comparative approach" has been instrumental in understanding where our abilities to learn, understand and produce new words came from. Together, these distinct traits allow free expression of new thoughts, so they are fundamental to human language, but they are not always present in other types of animal communication. Which other creatures share these abilities, and why?

The ability to learn to understand new signals is the most common. Typical dogs know a few words, and some unusual dogs like Rico, a border collie, could remember hundreds of names for different objects. The bonobo Kanzi, who was exposed from an early age to abundant human speech, can also understand hundreds of spoken words, and even notice differences in word order. This suggests that

THREE KINDS OF EVOLUTION

Language develops through time at three different rates, all of which have sometimes been termed "language evolution". The fastest process is ontogeny, in which an initially language-less baby becomes an adult native speaker. Then there's glossogeny: the historical development of languages. This guide to language evolution, however, focuses on human phylogeny: the biological changes that occurred during the last 6 million years of our lineage through which our species *Homo sapiens* evolved from an initially language-less primate.

Kanzi, a bonobo, can understand the meaning of hundreds of spoken words

learning to understand new signals is widespread and broadly shared with most other mammals – a homology. But neither Kanzi nor Rico ever learned to produce even a single spoken word, as they lack the capacity for complex vocal learning.

Many other species do have this ability, however. Almost everyone has seen a talking parrot, but there are more unusual examples. Hoover, an orphaned seal raised by fishermen, learned to produce whole English sentences with a Maine accent, and scientists have uncovered complex vocal learning in a wide variety of other species, including whales, elephants and bats. The fact that close relatives of these animals lack vocal learning indicates that this trait is an example of convergent evolution. Crucially, most animals who learn to speak do not understand the meaning of what they say. Hoover mostly directed his sentences at female seals during the mating period, for example, suggesting that vocal production and meaning recognition are two distinct traits that use different neural machinery.

Only with specific training can animals learn to both produce and appropriately interpret words. Alex, the African gray parrot (pictured, far left) of psychologist Irene Pepperberg, provided one example of a bird that used words for shapes, colours and numbers meaningfully.

In terms of creating a time line for human evolution, the evidence suggests our ability to recognise sounds – the homology – probably arose in a mammalian common ancestor, while our ability to produce complex sounds arose more recently in prehistory. Even more importantly, studying the various convergent examples of vocal learning have uncovered what was necessary for one important aspect of human language: speech.

TOP: PHILASHLEY/STONE+/GETTY/RIGHT: VINCENT J. MUSI/AURORA

THE MECHANICS OF SPEECH

Speech comes so easily to adult humans that it's easy to forget the sheer amount of muscular coordination needed to produce even the most basic sounds. How we came to have this ability, when most other animals find it so difficult, is one of the key questions in language evolution – and one of the few that has yielded to empirical studies.

Speech is just one aspect of human language, and is not even strictly necessary, since both sign language and written language are perfectly adequate for the unfettered expression of thought. However, since it is the normal medium of language in all cultures, it is reasonable to assume that its emergence must have represented a big step in the evolution of language.

Because no other apes apart from us can learn to speak, some change must have occurred after we diverged from our last common ancestor with chimpanzees, about 7 million years ago. The nature of the change has been somewhat unclear. Darwin suggested two possible explanations: either it was a change in our vocal apparatus, or there is a key difference in the brain. In each case, biologists have gained fundamental insights by examining other animals.

Let's start with anatomy. Humans have an unusual vocal tract: the larynx (or voicebox) rests low in the throat. In most other mammals, including chimpanzees, the larynx lies at a higher point, and is often inserted into the nasal passage, creating a sealed nasal airway. In fact, humans begin life this way: a newborn infant can breathe through its nose while swallowing milk through its mouth. But as the infant grows, the larynx descends, and by the age of 3 or 4 this feat is no longer possible.

The reconfigured human vocal tract allows the free movement of the tongue that is crucial to make the many distinct sounds heard in human languages. For a long time, the descended larynx was considered unique to our species, and the key to our possession of speech. Researchers had even tried to place a date on the emergence of language by studying the position of the larynx in ancient fossils.

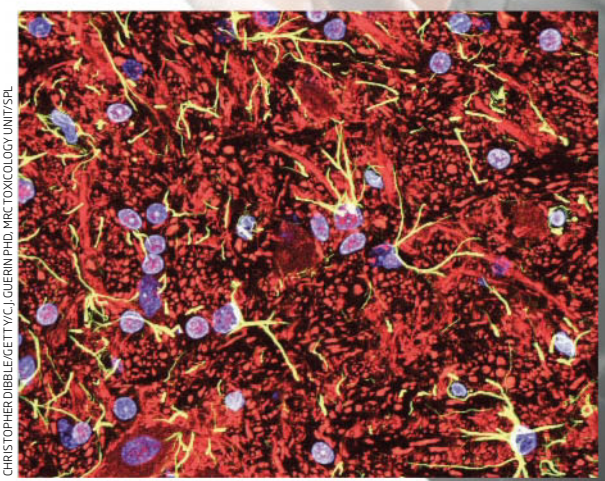
Evidence from two different sources of comparative data casts doubt on this hypothesis. The first was the discovery of animal species with permanently descended larynges like our own. We now know that lions, tigers, koalas and Mongolian gazelles all have a descended larynx – making it a convergent trait. Since none of these species produce anything vaguely speech-like, such changes in anatomy cannot be enough for speech to have emerged.

The second line of evidence is even more damning. X-ray observations of vocalising mammals show that dogs, monkeys, goats and pigs all lower the larynx during vocalisation. This ability to reconfigure the vocal tract appears to be a widespread, and probably homologous, feature of mammals. With its larynx retracted, a dog or a monkey has all the freedom of movement needed to produce many different vocalisations (see diagrams, right). The key changes must therefore have occurred in the brain.

Direct connections

The human brain is enormously complex, and differs in many ways from those of other animals. We expect different neural changes to underlie each of the different components of language, like syntax, semantics and speech. Others presumably underlie abilities like improved tool use or increased intelligence. Determining the specific neural changes that correspond to particular capabilities is often very difficult, and in many cases we don't even have good guesses about what changes were needed.

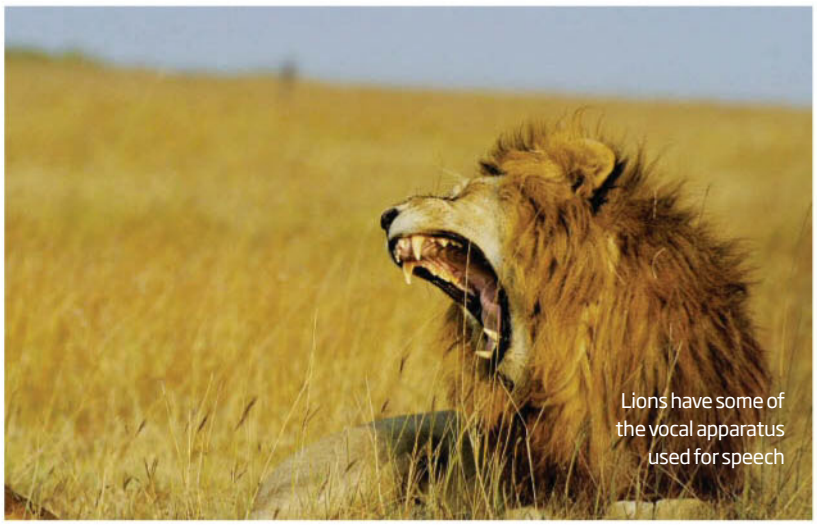
Biologists have been more fortunate when studying the neural machinery of speech, however. Motor neurons that control the muscles involved in vocalisation – in the lips, the tongue and the larynx – are located in the brainstem, and after decades of painstaking research we now know that



CHRISTOPHER DIBBLE/GETTY/C.J. GUERIN PHD, MRC TOXICOLOGY UNIT/SPL

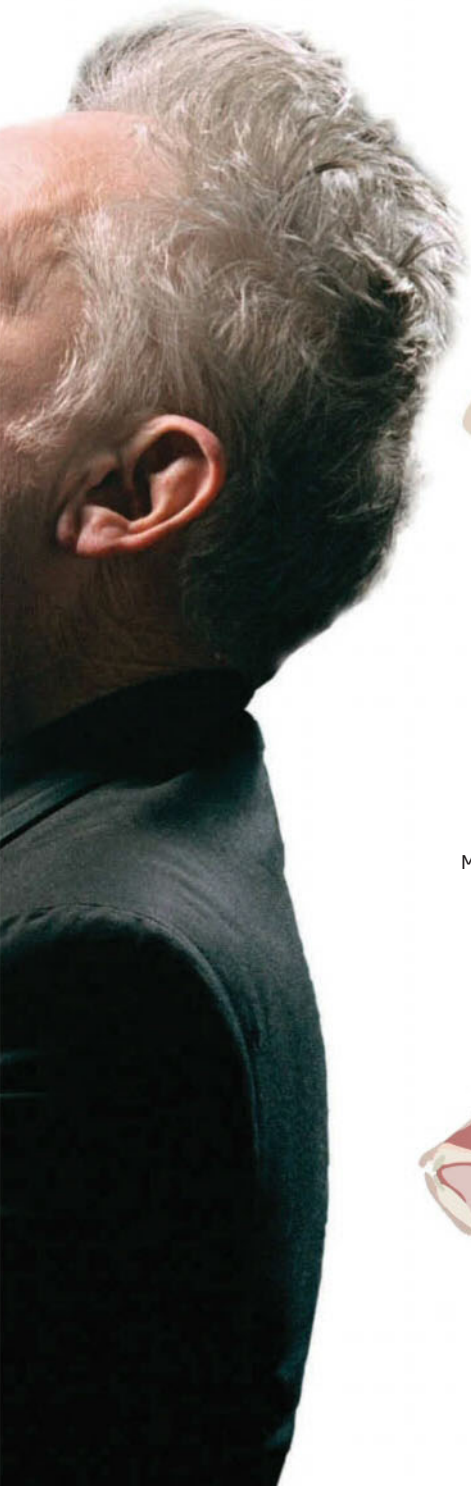
The human motor cortex has direct connections to the brainstem nerve cells that control speech

"The crucial evolutionary changes needed for speech are in the brain, rather than the larynx or tongue"



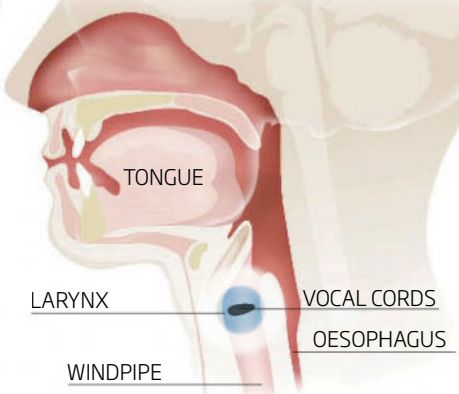
MICHAEL K. NICHOLS/NATIONAL GEOGRAPHIC/GETTY

Lions have some of the vocal apparatus used for speech



LOW LARYNX

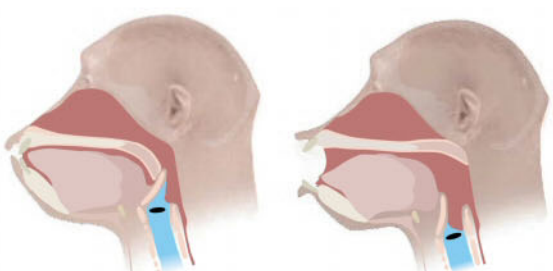
The human larynx at rest is placed lower in the throat than in most other mammals. This allows us to move the tongue more freely, which is important in the production of complex sounds



HIGH LARYNX

MAMMAL LARYNGES TYPICALLY OCCUPY A HIGHER POSITION

X-ray video scans, however, have revealed that many mammals lower the larynx during vocalisation. This sets the tongue free, which would allow the production of complex sounds – suggesting the true source of speech lies in the brain



HIGH LARYNX (at rest)

LOWERED LARYNX (during calling)

humans have direct neural connections, absent in nonhuman primates, between the motor cortex and these brainstem neurons. Could these direct neural connections explain our enhanced ability to control and coordinate the movements necessary for speech? The explanation seemed plausible. Fortunately, we can test the hypothesis with the help of other species that exhibit complex vocal learning.

If direct neural connections are necessary for vocal learning, we predict they should appear in other vocal learning species. For birds at least, this prediction appears to hold true: parrots or songbirds have the connection while chickens or pigeons, which are not vocal learners, lack them. For many vocal learning species, including whales, seals, elephants and bats, we don't know, because their neuroanatomy has yet to be fully investigated, providing untapped sources to test the "direct connections" hypothesis.

An ability to produce the correct sounds for speech is one thing, but complex vocal control in humans also relies on our ability to control the different articulators in the correct, often complicated, sequences. The discovery of the *FOXP2* gene has recently provided insights into the origins of this ability. Modern humans all share a novel variant of this gene which differs from the one most primates have, and disruptions of this gene in people create severe speech difficulties.

But what does it do? Various studies have found that the gene seems to be crucial for memory formation in the basal ganglia and cerebellum, which are involved in coordinating the patterns of movements that are crucial for our complex vocalisations. Recently, fossil DNA recovered from Neanderthals has shown that they shared the modern variant, suggesting that they already possessed complex speech.

Speech is just one component of language, though, and similar questions must be asked about syntax and semantics before we can hope to understand the evolution of language as a whole.

PROTOLANGUAGES

When viewing language as a collection of many distinct components, it becomes clear that the different linguistic traits must have appeared at different periods of human evolution, perhaps for different reasons. But while most theorists agree that early humans passed through multiple stages en route to modern language, there are major differences of opinion concerning the order in which the different components appeared.

A system that possesses some, but not all, components of language can be termed a "protolanguage" – a term introduced by anthropologist Gordon Hewes in 1973. Three potential protolanguages dominate theories of language evolution.

MUSICAL BEGINNINGS

This prominent model of protolanguage was offered by Darwin in 1871, and focused on the origins of vocal learning, a capability assumed (but not explained) by word-based, or lexical, protolanguage. Darwin realised that in most vocal learning species, complex learned vocalisations are not used to communicate detailed information, but rather provide a display of the singer's virtuosity. While the songs of some birds or whales rival human speech in acoustic complexity, they convey only a very simple message, roughly "I'm an adult of your species and want to mate".

Based on this analogy, Darwin suggested that human vocal learning originated in the context of sexual selection, territoriality and mate choice, and initially resembled song more closely than speech. Only later, by this model, did the individual notes and syllables of these vocal displays take on meaning, probably in an initially holistic manner. Since Darwin, many others have taken up the musical protolanguage hypothesis, and it is attracting increasing support today. One virtue of this hypothesis is that it also provides an explanation for music: another universal characteristic of our species. By this model, music is a living reminder of an earlier stage of human evolution, preceding true language.

JIM RICHARDSON/NATIONAL GEOGRAPHIC/GETTY/ROBERT CAPUTO/AURORA

Speech and music are universal characteristics of our species – but did they evolve together?

FIRST WORDS

One highly intuitive model suggests that early humans had words, but did not arrange them into syntactically structured sentences. This model of "lexical protolanguage" parallels language development in children, who start out with single-word utterances, move on to a two-word stage, and then begin forming more complex sentences with syntax.

Linguist Derek Bickerton at the University of Hawaii, Manoa, is one of the main proponents of this model. He once suggested that the addition of syntax in all of its complexity might have occurred quite suddenly, due to a mutation with large effects on brain wiring, quickly catapulting our species into full language. But linguist Ray Jackendoff at Tufts University in Medford, Massachusetts, suggests a much finer and more gradual path to modern syntax, starting with simple word order and progressing steadily to fine points of grammar that make modern languages difficult for adults to learn.

Despite various other disagreements, all proponents of lexical protolanguage agree that language began with spoken words, referring to objects and events. Most also agree that the purpose of protolanguage was the communication of ideas. Although each of these assumptions seems intuitive, they are challenged in other models of the evolution of language.

ELOQUENT GESTURES

Another well-established model of protolanguage suggests that language was originally conveyed by gestures, rather than speech. One avenue of evidence comes from observations of apes, which lack vocal learning and speech, but use manual gestures in an intentional, flexible and informative way. While attempts to teach apes spoken language fail completely, efforts to teach apes to communicate via manual gesture have been much more successful.

Although no ape has ever mastered a full sign language, apes can learn and use hundreds of individual manual gestures communicatively. The visual/manual mode is clearly adequate for full human language, as sign languages convincingly demonstrate. These two features make gestural protolanguage a popular alternative view today.

One prominent version of gestural protolanguage, offered by neuroscientist Michael Arbib at the University of Southern California, Los Angeles, suggests that signs did not, initially, refer to individual objects or actions, but rather to whole thoughts or events. This is an example of what is called a holistic protolanguage, in which whole complex signals map directly onto whole complex concepts, rather than being segmented into individual words. This is precisely how early gestures and utterances are understood and used by infants in early language acquisition. In such holistic models, whole sentences came first, and were only divided up into words during a later "analytic" stage of biological evolution and cultural development. This approach contrasts with the "synthetic" models of protolanguage, which have individual words from the very beginning.

Gestural models face the difficulty of explaining why our species switched to using speech so thoroughly. It may have been due to the need to communicate in darkness, or because hands became occupied by tools. But speech has disadvantages too. Speaking aloud, we cannot safely communicate with our mouths full, or in the presence of dangerous predators, or in loud environments like waterfalls or storms. The selective pressures that might have driven humans to rely so heavily on speech alone remain elusive.

SANDRA ROCHA/KAMERAPHOTO/PICTURETANK

Could gestures have provided the beginnings of language?

APE THAT SPOKE A PROTOLANGUAGE?

The earliest writing, providing clear evidence of modern language, dates from just 6000 years ago, but language in its modern form emerged long before then. Because all modern humans come from an ancestral African population, and children from any existing culture can learn any language, language must have preceded our emigration from Africa at least 50,000 years ago. But can we put a date on the emergence of the first rudimentary protolanguages?

Whether gestural, musical or lexical, protolanguage considerably surpassed modern ape communication in the wild. With all the cognitive challenges, and benefits, this would bring, we would expect these early humans to differ considerably from their forebears in both anatomy and culture. Using this logic, *Homo erectus*, which originated almost 2 million years ago, appears to be the most likely candidate.

H. erectus were larger than their predecessors, and had brain sizes of 900 to 1100 cubic centimetres. These approach the size of our own brains, which average about 1350 cubic centimetres. This suggests a capability for flexible intelligence and culture. Their stone tools were vastly more sophisticated than those of *Australopithecus*, suggesting they may have had more advanced communication, though the tools were less sophisticated than tools made by Neanderthals and modern humans.

Importantly, the *H. erectus* tools appeared to reach a kind of stasis – their iconic Achulean hand axe, which was a symmetrical all-purpose tool, persisted for a million years. This suggests they did not have full language, which would have accelerated cultural and technological change. Hence they might have had some, but not all, of the linguistic capacities modern humans possess – a protolanguage, in other words.

TOM MAIN

W. Tecumseh Fitch

W. Tecumseh Fitch is professor of cognitive biology at the University of Vienna in Austria. He studied evolutionary biology, neuroscience, cognitive science and linguistics at Brown University in Providence, Rhode Island, did his postdoc at MIT and has taught at Harvard University and the University of St Andrews in the UK. Fitch's research focuses on the evolution of cognition and communication in vertebrates and includes human music and language. His book *The Evolution of Language* was published in 2010.

THE FUTURE

Each of the models of human protolanguages clearly has strengths and weaknesses. Contemporary theorists mix and match among the possibilities, and the truth will probably incorporate elements from each of these models. But since each model of protolanguage makes different predictions about when particular new capabilities appeared during the course of human evolution, they are in principle testable.

Genetics provides the most exciting source of new evidence for the origins of language. DNA recovered from early human fossils allows us to estimate when particular mutations tied to particular aspects of language arose, and when studying more recent genetic changes, it is also possible to estimate the timing of evolutionary events by examining variation in modern humans.

Multiple genes have recently been linked to dyslexia, for example. Although dyslexia is identified by difficulties with learning to read, it often seems to result from some more fundamental problems with the way the sounds of language are processed. These genes may therefore be linked to the phonological components of language, which Darwin's model would argue evolved early, but which Michael Arbib's gestural model would predict to be latecomers.

In contrast, genes linked to autism lead to difficulties in understanding others' thoughts and feelings: capacities linked to semantics. By Darwin's model the normal human form of these genes should be latecomers, while in a gestural or lexical model they would have become involved in language at an earlier stage. Determining when human-specific variations of such genes arose in the human lineage may therefore allow us to test hypotheses about protolanguage directly.

So although we may never be able to write a Neanderthal dictionary, there is good reason to think that, as our data improves in the coming decades, we will be able to test ideas about human language evolution. The scientific study of language evolution appears to be coming of age.

RECOMMENDED READING

The Evolution of Language by W. Tecumseh Fitch (Cambridge University Press, 2010)

The Descent of Man, and Selection in Relation to Sex by Charles Darwin (John Murray, 1871)

Baboon Metaphysics by Dorothy L. Cheney and Robert M. Seyfarth (University of Chicago Press, 2007)

Language and Species by Derek Bickerton (Chicago University Press, 1990)

The Symbolic Species by Terrence Deacon (Norton, 1997)

Origins of the Modern Mind by Merlin Donald (Harvard University Press, 1991)

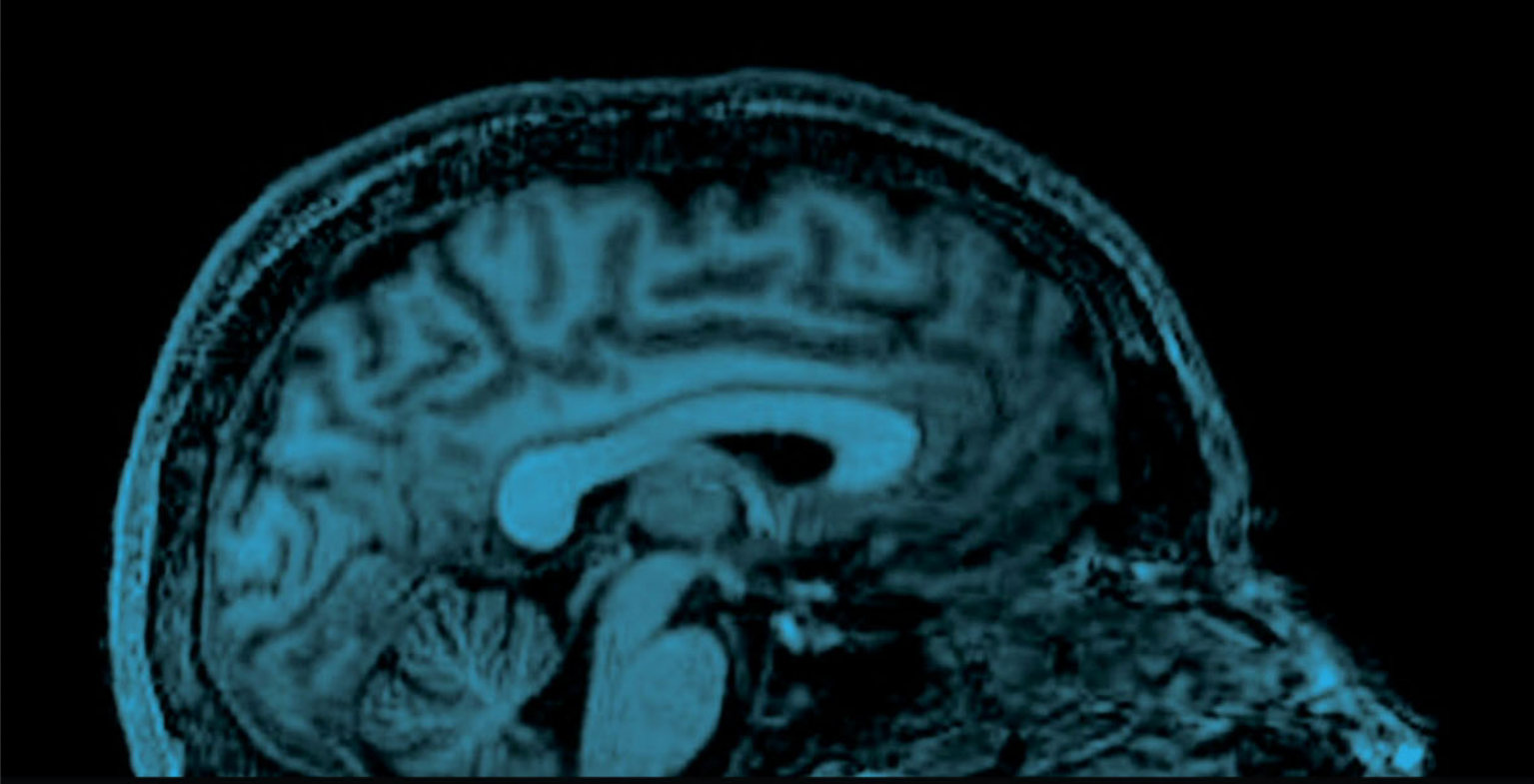
The Singing Neanderthals by Steven Mithen (Weidenfeld & Nicolson, 2005)

"The Language Faculty: What is it, who has it, and how did it evolve?" by Marc Hauser, Noam Chomsky and W. Tecumseh Fitch, (*Science*, 2002, vol 298, p 1569

"Neural systems for vocal learning in birds and humans: a synopsis" by Erich Jarvis,. *Journal of Ornithology*, 2007, vol 148, supplement 1, p 35

"The derived *FOXP2* variant of modern humans was shared with Neanderthals" by Johannes Krause and others, (2007) (*Current Biology*, *2007*, vol 17, p 1908

Cover image: Grant Faint/Getty



CHAPTER THREE

YOUR AMAZING MIND

THE HUMAN BRAIN

MICHAEL O'SHEA

INSTANT
EXPERT

THE BRAIN THROUGH HISTORY

About 250,000 years ago, something quite extraordinary happened. Animals with an unprecedented capacity for thought appeared on the savannahs of Africa. These creatures were conscious and had minds. Eventually, they were smart enough to start questioning the origins of their own intelligence. We are finally close to getting some answers, with a particularly detailed understanding of the brain's building block – the neuron. But it has not been a smooth journey.

AXON ENDING FORMS SYNAPSE WITH NEXT NEURON

THE BEGINNINGS

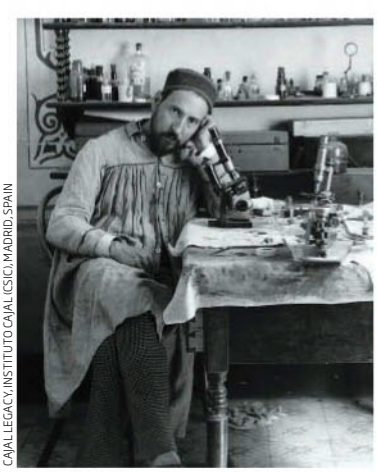
The birth of neuroscience began with Hippocrates some 2500 years ago. While his contemporaries, including Aristotle, believed that the mind resided in the heart, Hippocrates argued that the brain is the seat of thought, sensation, emotion and cognition.

It was a monumental step, but a deeper understanding of the brain's anatomy and function took a long time to follow, with many early theories ignoring the solid brain tissue in favour of the brain's fluid filled cavities, or ventricles. The influential 2nd-century physician Galen was perhaps the most notable proponent of this idea. He believed the human brain had three ventricles, and that each one was responsible for a different mental faculty: imagination, reason and memory. According to his theory, the brain controlled our body's activities by pumping fluid from the ventricles through the nerves to other organs.

Such was Galen's authority that the idea cast a long shadow over our understanding of the brain, and fluid theories of the brain continued to dominate until well into the 17th century. Even such luminaries as the French philosopher René Descartes compared the brain to a hydraulic-powered machine. Yet the idea had a major flaw: a fluid could not move quickly enough to explain the speed of our reactions.

A more enlightened approach came when a new generation of anatomists began depicting the structure of the brain with increasing accuracy. Prominent among them was 17th-century English doctor Thomas Willis, who argued that the key to how the brain worked lay in the solid cerebral tissues, not the ventricles. Then, 100 years later, Luigi Galvani and Alessandro Volta showed that an external source of electricity could activate nerves and muscle. This was a crucial development, since it finally suggested why we respond so rapidly to events. But it was not until the 19th century that German physiologist Emil Du Bois-Reymond confirmed that nerves and muscles themselves generate electrical impulses.

PETER MACDIARMID/GETTY; RIGHT: SPL



CAJAL LEGACY, INSTITUTO CAJAL (CSIC), MADRID, SPAIN

Santiago Ramón y Cajal is considered by many to be the father of modern neuroscience

All of which paved the way for the modern era of neuroscience, beginning with the work of the Spanish anatomist Santiago Ramón y Cajal (pictured) at the dawn of the 20th century. His spectacular observations identified the neuron as the building block of the brain. He found them to have a diversity of forms that is not found in the cells of other organs. Most surprisingly, he noted that insect neurons matched and sometimes exceeded the complexity of human brain cells. This suggested that our abilities depend on the way neurons are connected, not on any special features of the cells themselves.

Cajal's "connectionist" view opened the door to a new way of thinking about information processing in the brain, and it still dominates today.

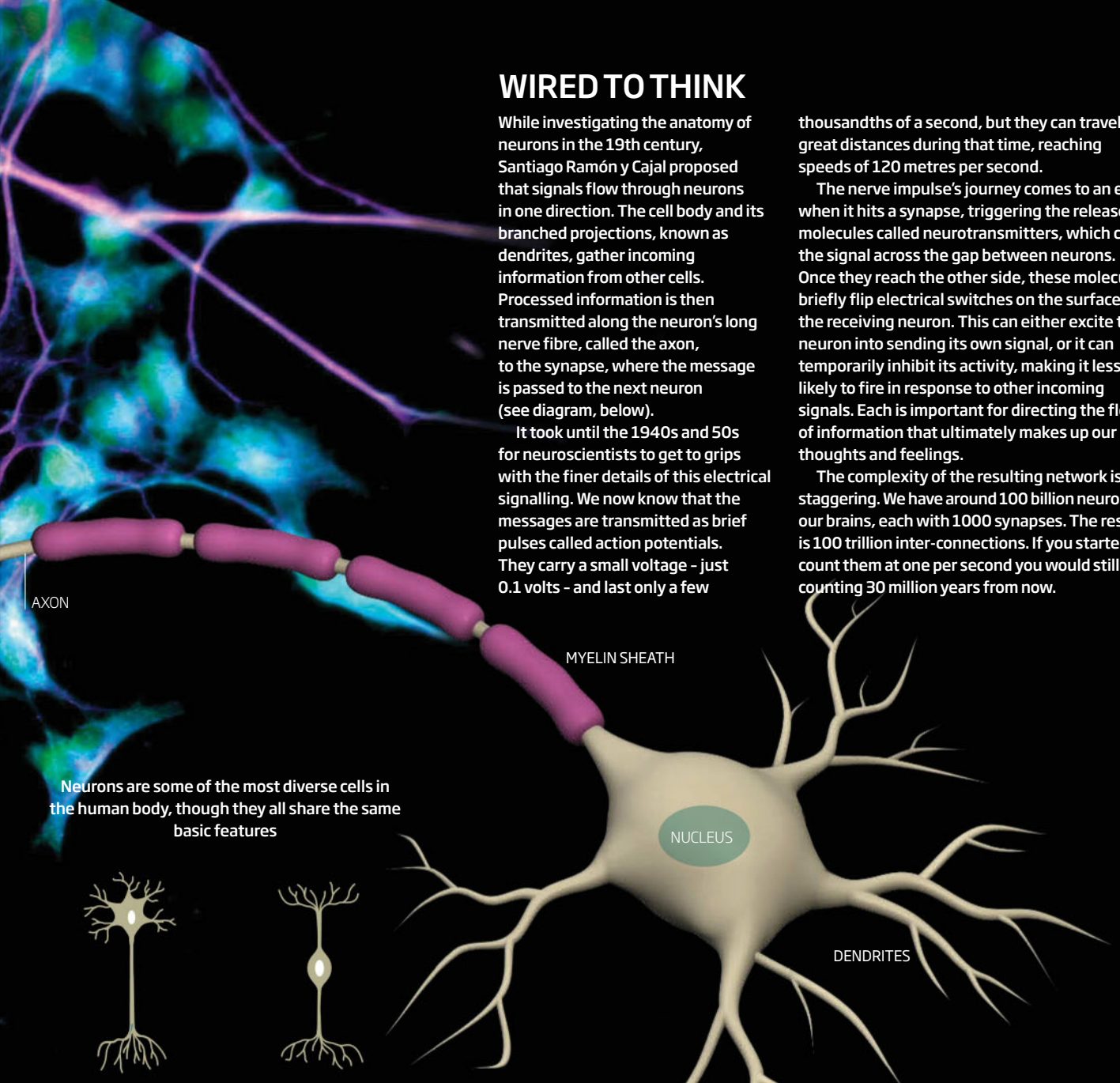
WIRED TO THINK

While investigating the anatomy of neurons in the 19th century, Santiago Ramón y Cajal proposed that signals flow through neurons in one direction. The cell body and its branched projections, known as dendrites, gather incoming information from other cells. Processed information is then transmitted along the neuron's long nerve fibre, called the axon, to the synapse, where the message is passed to the next neuron (see diagram, below).

It took until the 1940s and 50s for neuroscientists to get to grips with the finer details of this electrical signalling. We now know that the messages are transmitted as brief pulses called action potentials. They carry a small voltage – just 0.1 volts – and last only a few thousandths of a second, but they can travel great distances during that time, reaching speeds of 120 metres per second.

The nerve impulse's journey comes to an end when it hits a synapse, triggering the release of molecules called neurotransmitters, which carry the signal across the gap between neurons. Once they reach the other side, these molecules briefly flip electrical switches on the surface of the receiving neuron. This can either excite the neuron into sending its own signal, or it can temporarily inhibit its activity, making it less likely to fire in response to other incoming signals. Each is important for directing the flow of information that ultimately makes up our thoughts and feelings.

The complexity of the resulting network is staggering. We have around 100 billion neurons in our brains, each with 1000 synapses. The result is 100 trillion inter-connections. If you started to count them at one per second you would still be counting 30 million years from now.

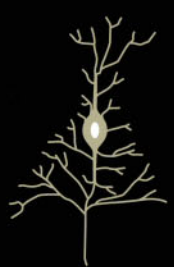


Neurons are some of the most diverse cells in the human body, though they all share the same basic features

MOTOR
Send signals to parts of the body, such as muscle, to direct movement

INTER
Provide a connective bridge between other neurons

SENSORY
Transmit signals to the brain from the rest of the body



PYRAMIDAL
Involved in many areas of cognition, such as object recognition within the visual cortex

THE PLASTIC BRAIN

Unlike the electronic components of a computer, our networks of neurons are flexible thanks to a special class of neurotransmitter. These "neuromodulators" act a bit like a volume control, altering the amount of other neurotransmitters released at the synapse and the degree to which neurons respond to incoming signals. Some of these changes help to fine-tune brain activity in response to immediate events, while others rewire the brain in the long term, which is thought to explain how memories are stored.

Many neuromodulators act on just a few neurons, but some can penetrate through large swathes of brain tissue creating sweeping changes. Nitric oxide, for example, is so small (the 10th smallest molecule in the known universe, in fact) that it can easily spread away from the neuron at its source. It alters receptive neurons by changing the amount of neurotransmitter released with each nerve impulse, kicking off the changes that are necessary for memory formation in the hippocampus.

Through the actions of a multitude of chemical transmitters and modulators, the brain is constantly changing, allowing us to adapt to the world around us.

MAPPING THE MIND

Our billions of neurons, joined by trillions of neural connections, build the most intricate organ of the human body. Attempts to understand its architecture began with reports of people with brain damage. Localised damage results in highly specific impairments of particular skills – such as language or numeracy – suggesting that our brain is modular, with different locations responsible for different mental functions.

Advanced imaging techniques developed in the late 20th century gave a more nuanced approach by allowing researchers to peer into healthy brains as volunteers carried out different cognitive tasks. The result is a detailed map of where different skills arise in the brain – an important step on the road to understanding our complex mental lives.

FOREBRAIN

Many of our uniquely human capabilities arise in the forebrain, which expanded rapidly during the evolution of our mammalian ancestors. It includes the thalamus, a relay station that directs sensory information to the cerebral cortex for higher processing; the hypothalamus, which releases hormones into the bloodstream for distribution to the rest of the body; the amygdala, which deals with emotion; and the hippocampus, which plays a major role in the formation of spatial memories.

Among the most recently evolved parts are the basal ganglia, which regulate the speed and smoothness of intentional movements initiated by the cerebral cortex. Connections in this region are modulated by the neurotransmitter dopamine, provided by the midbrain's substantia nigra. A deficiency in this source is associated with many of the symptoms of Parkinson's disease, such as slowness of movement, tremor and impaired balance. Although drugs that boost levels of the neurotransmitter in the basal ganglia can help,a cure for Parkinson's is still out of reach.

Finally, there is the cerebral cortex – the enveloping hemispheres thought to make us human. Here plans are made, words are formed and ideas generated. Home of our creative intelligence, imagination and consciousness, this is where the mind is formed.

Structurally, the cortex is a single sheet of tissue made up of six crinkled layers folded inside the skull; if it were spread flat it would stretch over 1.6 square metres. Information enters and leaves the cortex through about a million neurons, but it has more than 10 billion internal connections, meaning the cortex spends most of its time talking to itself.

Each of the cortical hemispheres have four principal lobes (see upper diagram, right). The frontal lobes house the neural circuits for thinking and planning, and are also thought to be responsible for our individual personalities. The occipital and temporal lobes are mainly concerned with the processing of visual and auditory information, respectively. Finally, the parietal lobes are involved in attention and the integration of sensory information.

The body is "mapped" onto the cortex many times, including one map representing the senses and another controlling our movements. These maps tend to preserve the basic structure of the body, so that neurons processing feelings from your feet will be closer to those dealing with sensations from your legs than those crunching data from your nose, for example. But the proportions are distorted, with more brain tissue devoted to the hands and lips than the torso or legs. Redrawing the body to represent these maps results in grotesque figures like Penfield's homunculus (below left).

The communications bridge between the two cerebral hemispheres is a tract of about a million axons, called the corpus callosum. Cutting this bridge, a procedure sometimes performed to alleviate epileptic seizures, can split the unitary manifestation of "self". It is as if the body is controlled by two independently thinking brains. One smoker who had the surgery reported that when he reached for a cigarette with his right hand, his left hand would snatch it and throw it away!

As we have seen, different tasks are carried out by different cortical regions. Yet all you have to do is open your eyes to see that these tasks are combined smoothly: depth, shape, colour and motion all merge into a 3D image of the scene. Objects are recognised with no awareness of the fragmented nature of the brain's efforts. Precisely how this is achieved remains a puzzle. It's called the "problem of binding" and is one of the many questions left to be answered by tomorrow's neuroscientists.

NATURAL HISTORY MUSEUM, LONDON; BACKGROUND: GILLES PERESS/MAGNUM

"Cutting the bridge between the brain's two hemispheres can split the 'self'. It is as if the body is controlled by two independently thinking brains"

MIDBRAIN

The midbrain plays a role in many of our physical actions. One of its central structures is the substantia nigra, so called because it is a rich source of the neurotransmitter dopamine, which turns black in post-mortem tissue. Since dopamine is essential for the control of movement, the substantia nigra is said to "oil the wheels of motion". Dopamine is also the "reward" neurotransmitter and is necessary for many forms of learning, compulsive behaviour and addiction.

Other regions of the midbrain are concerned with hearing, visual information processing, the control of eye movements and the regulation of mood.

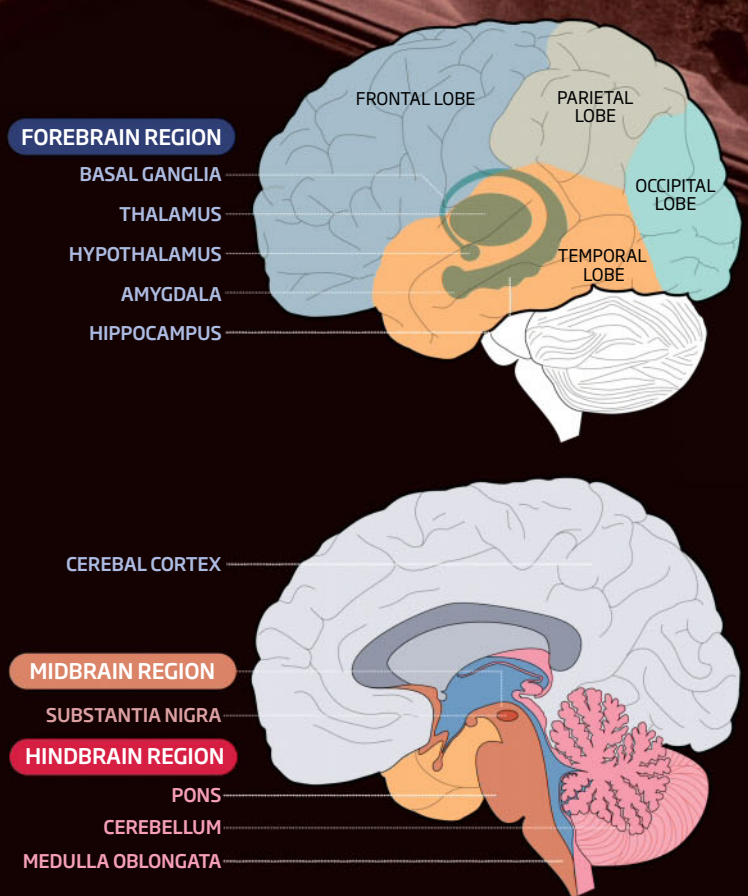
HINDBRAIN

As its name suggests, the hindbrain is located at the base of the skull, just above the neck. Comparisons of different organisms suggest it was the first brain structure to have evolved, with its precursor emerging in the earliest vertebrates. In humans it is made up of three structures: the medulla oblongata, pons and cerebellum.

The medulla oblongata is responsible for many of the automatic behaviours that keep us alive, such as breathing, regulating our heart beat and swallowing. Significantly, its axons cross from one side of the brain to the other as they descend to the spinal cord, which explains why each side of the brain controls the opposite side of the body.

A little further up is the pons, which also controls vital functions such as breathing, heart rate, blood pressure and sleep. It also plays an important role in the control of facial expressions and in receiving information about the movements and orientation of the body in space.

The most prominent part of the hindbrain is the cerebellum, which has a very distinctive rippled surface with deep fissures. It is richly supplied with sensory information about the position and movements of the body and can encode and memorise the information needed to carry out complex fine-motor skills and movements.



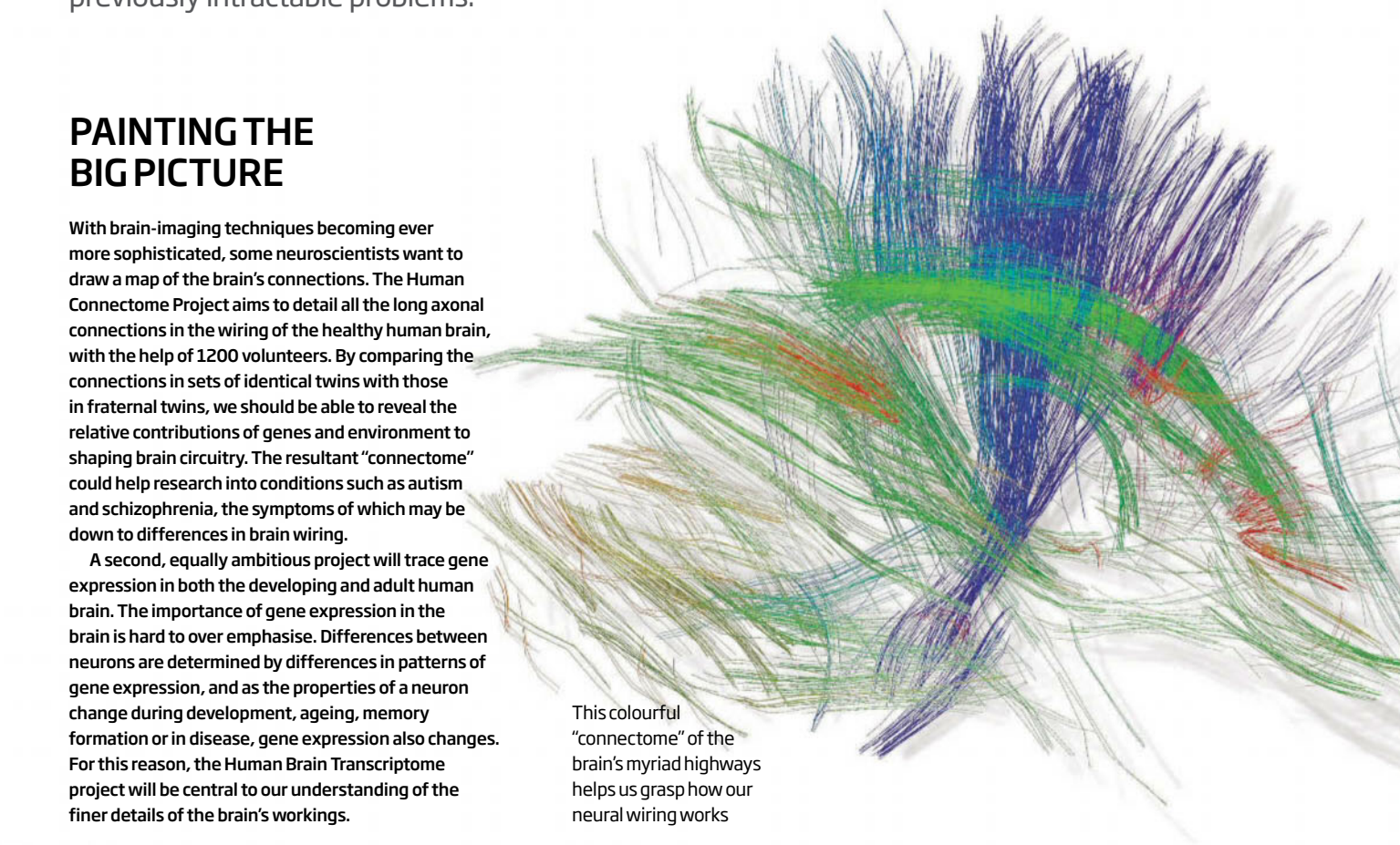
THE QUEST FOR THE HUMAN MIND

Almost 2500 years since Hippocrates first saw the brain as the centre of our thoughts, we can now interrogate its inner world with some of the most advanced technology known to humanity. The ultimate aim is to work out how the brain generates our conscious mind. Such understanding is still a long way off, but we are beginning to get to grips with many previously intractable problems.

PAINTING THE BIG PICTURE

With brain-imaging techniques becoming ever more sophisticated, some neuroscientists want to draw a map of the brain's connections. The Human Connectome Project aims to detail all the long axonal connections in the wiring of the healthy human brain, with the help of 1200 volunteers. By comparing the connections in sets of identical twins with those in fraternal twins, we should be able to reveal the relative contributions of genes and environment to shaping brain circuitry. The resultant "connectome" could help research into conditions such as autism and schizophrenia, the symptoms of which may be down to differences in brain wiring.

A second, equally ambitious project will trace gene expression in both the developing and adult human brain. The importance of gene expression in the brain is hard to over emphasise. Differences between neurons are determined by differences in patterns of gene expression, and as the properties of a neuron change during development, ageing, memory formation or in disease, gene expression also changes. For this reason, the Human Brain Transcriptome project will be central to our understanding of the finer details of the brain's workings.



This colourful "connectome" of the brain's myriad highways helps us grasp how our neural wiring works

"The SPAUN brain simulation performs well in memory tasks – it even passed a human IQ test"

DECODING THE SENSES

How do neurons encode an idea of something so that we can immediately recognise a familiar face, our house or a favourite book? Most neuroscientists believe that the brain stores our concept of an object over many neurons, with all these cells having to work together for you to recognise something. According to this theory, the activity of any one neuron is not representative of a particular object – it could respond to similar features in other objects. Instead, it is the behaviour of the group that determines what meaning comes to mind.

But some neuroscientists claim that we may encode concepts using smaller, more selective, networks of neurons. According to this view, a neuron may sometimes specialise in a single idea. For instance, in one study volunteers were shown pictures of movie stars and famous buildings while the researchers recorded the resulting activity from a selection of single neurons. The results were surprising, showing, for example, that one of the neurons studied responded to many different pictures of the actor Jennifer Aniston.

In some cases it wasn't just pictures that triggered the neuron's activity; some also responded to a word representing the object or person. It's almost as if the neuron being studied somehow encoded the essence of the person or object, which may explain why we can recognise things from different perspectives or in unfamiliar surroundings.

Does this picture set your Jennifer Aniston neuron on fire?

CLOCKWISE FROM TOP LEFT: LABORATORY OF NEURO IMAGING AT UCLA AND MARTINOS CENTER FOR BIOMEDICAL IMAGING AT MGH CONSORTIUM OF THE HUMAN CONNECTOME PROJECT; STARMAX/XPOSURE; RAVI DEEPRES/GALLERY STOCK

REFLECTING ON EACH OTHER

Some neuroscientists believe that the discovery of "mirror neurons" will transform our understanding of the human mind and brain just as DNA transformed evolutionary biology. They could potentially help to de-mystify the most human of our qualities, such as empathy.

So what are mirror neurons? The defining characteristic is that they fire both when we perform an action such as reaching for a coffee cup, and when we see someone else doing the same. This suggests that they embody an understanding of the meaning or intentions of the actions of others, and through a similar mechanism allow us to grasp their emotions too.

It has also been suggested that they lie behind language. According to one theory, human language originated with physical gestures – and mirror neurons were instrumental in helping us to translate the meaning of these gestures. Although the idea is controversial, evidence is accumulating. For example, functional MRI studies show that a mirror neuron system lies close to a language centre called Broca's area.

A CONSCIOUS COMPUTER?

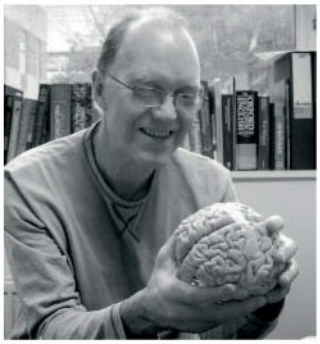
"Consciousness is a fascinating but elusive phenomenon," wrote the late Stuart Sutherland of the University of Sussex in Brighton, UK. "It is impossible to specify what it is, what it does, or why it evolved. Nothing worth reading has been written on it."

The problem arises because although consciousness must come from a physical structure, no one has been able to work out how. A potential breakthrough may lie in attempts to create robots with artificial brains capable of conscious thought and understanding.

One approach is to build an accurate, large-scale model or simulation of the neuronal networks of the human cortex in the hope that this will capture the signatures of human cognition. The Semantic Pointer Architecture Unified Network project is a promising example. To date, the SPAUN model consists of 2.5 million artificial neurons and recent reports suggest it can perform tasks of the kind that contribute to human cognition. For example, it can recognise visual images, perform well in a variety of memory tasks, and has even passed an IQ test.

By using more advanced models of this kind, it may be possible to test out the prerequisites of consciousness in a way that would not be feasible with a human or animal brain.

WELLCOME TRUST CENTRE FOR NEUROIMAGING



Michael O'Shea

Michael O'Shea is a professor of neuroscience in the school of life sciences and co-director of the Centre for Computational Neuroscience and Robotics at the University of Sussex in the UK. He is the author of *The Brain: A very short introduction* (Oxford University Press, 2005)

FUTURE HORIZONS

Is the mind beyond human understanding?

The answer to that question may be a humbling one. Consider, for example, Nobel prize-winning physicist Erwin Schrödinger's view on the knowledge-limitation problem, elegantly expressed in his book *What is Life?*. Of our attempts to understand the molecular interactions on which life depends, he writes:

"That is a marvel – than which only one is greater; one that, if intimately connected with it, yet lies on a different plane. I mean the fact that we, whose total being is entirely based on a marvellous interplay of this very kind, yet possess the power of acquiring considerable knowledge about it. I think it is possible that this knowledge may advance to little short of a complete understanding – of the first marvel. The second may well be beyond human understanding."

Are we really beyond the brain's capacity for understanding when attempting to fathom the conscious mind? Perhaps we should not be so despairing, for if we are indeed groping in the dark because current physics is incomplete, then we can be hopeful that new physical laws will shed light on that deepest of mysteries: the physical mechanisms and functions of consciousness.

RECOMMENDED READING

Principles of Neural Science (5th edition) by Eric Kandel and others (McGraw-Hill Medical, 2012)

The Brain: A very short introduction by Michael O'Shea (Oxford University Press, 2005)

Shadows of the Mind: A search for the missing science of consciousness by Roger Penrose (Oxford University Press, 1994)

The Astonishing Hypothesis: The scientific search for the soul by Francis Crick (Scribner Book Company, 1994)

WEBSITES

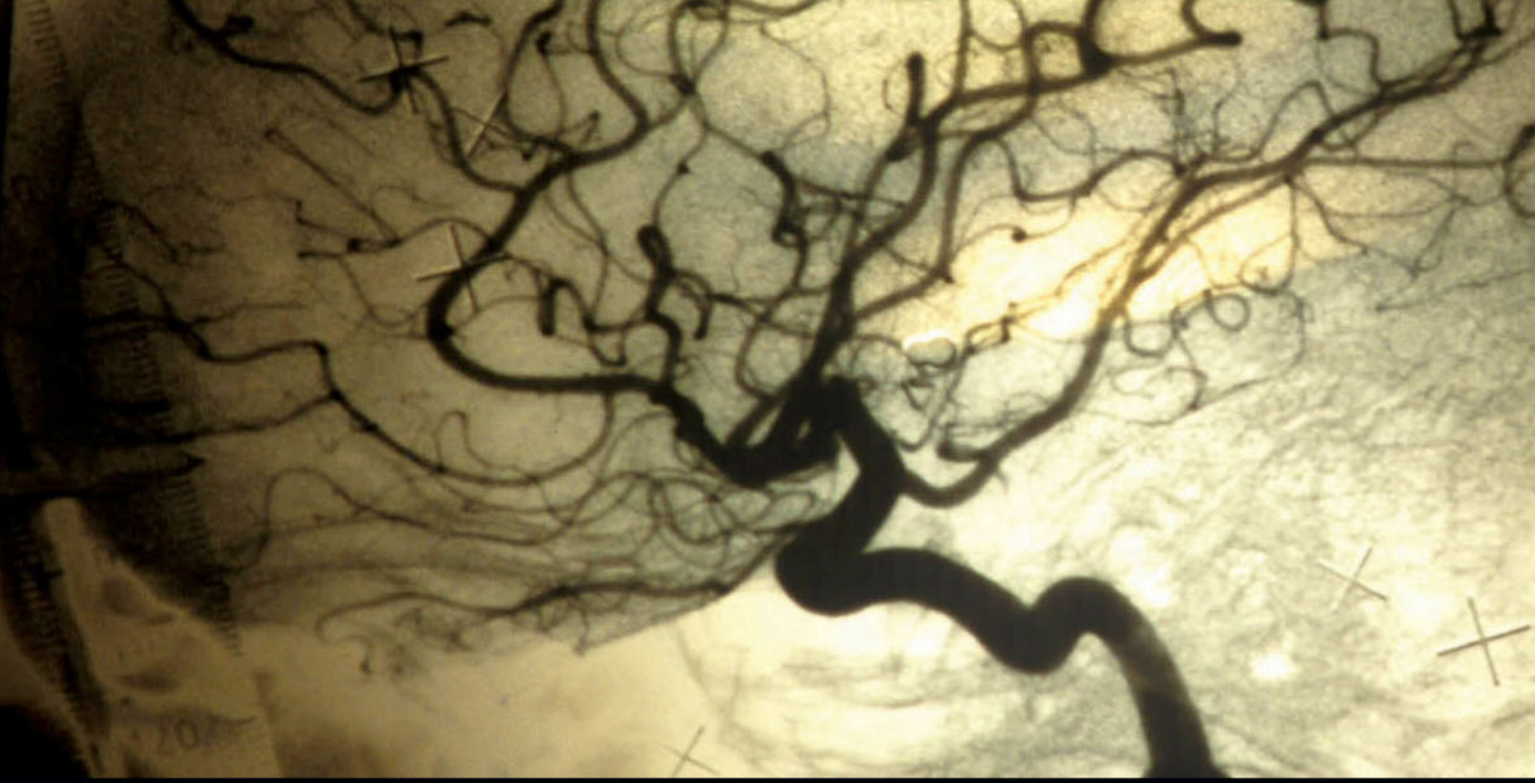
Sackler Centre for Consciousness Science: sussex.ac.uk/sackler

The Society for Neuroscience – Brain Facts: bit.ly/W3CKIQ

Large-scale human brain projects: humanconnectomeproject.org and humanbrainproject.eu/files/HBP_flagship.pdf

Cover image

Wellcome Trust Centre for Neuroimaging



INTELLIGENCE

LINDA GOTTFREDSON

INSTANT EXPERT

WHAT IS INTELLIGENCE?

Intelligence matters to us. In surveys people rank it second only to good health. Women worldwide believe smarter men make better husband material. Entrepreneurs hawk brain-boosting games, foods, supplements and training programmes. And the media quickly broadcast any scientific study claiming to discover how we can make ourselves, or our children, smarter. Yet our keen private interest in intelligence is matched by a reluctance to acknowledge publicly that some people have more of it than others. Democratic people value social equality above all, so they mistrust anything that might generate or justify inequality – but intelligence is no more equally distributed in human populations than height is. This tension has led to rancorous controversy over intelligence and intelligence testing but it has also benefited the science by pushing it exceedingly hard. A century of clashes and stunning discoveries has upended assumptions and revealed some fascinating paradoxes. Intelligence is definitely not what most of us had imagined.

QUANTIFYING INTELLIGENCE

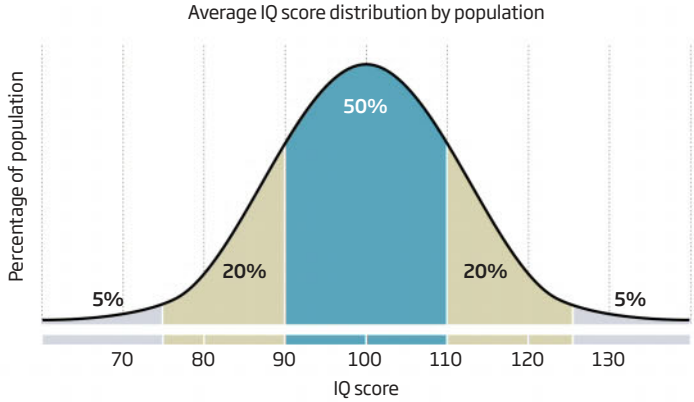
The first intelligence quotient (IQ) test was born of a desire to help the most vulnerable. In 1904 the French Ministry of Education commissioned psychologist Alfred Binet to find a practical way to identify children who would fail elementary school without special help. Binet assembled 30 short, objective questions on tasks such as naming an everyday object and identifying the heavier of two items. A child's performance on these, he believed, would indicate whether their learning was "retarded" relative to their peers. His invention worked and its success spawned massive intelligence-testing programmes on both sides of the Atlantic. Organisations turned to IQ tests to screen large pools of applicants: military recruits for trainability, college applicants for academic potential and job applicants for employability and promotability. The tests were eagerly adopted at first as a way to select talent from all social levels, but today their use can be considered contentious, partly because they do not find equal amounts of intelligence everywhere.

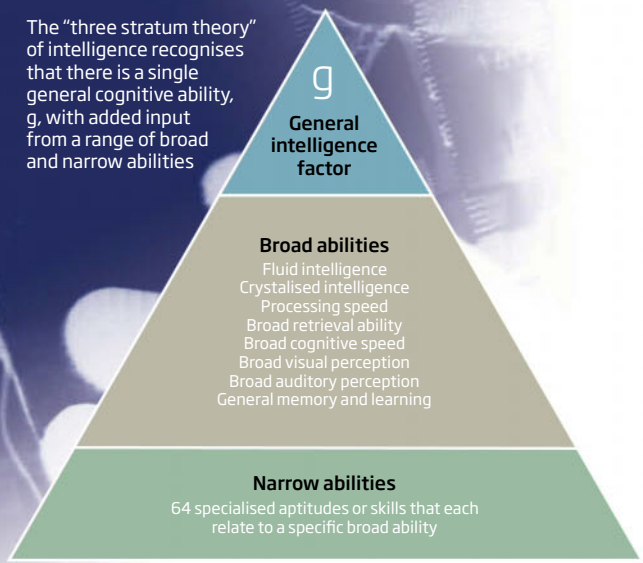
Nevertheless, intelligence testing continues because it has practical value. Many colleges, employers and the armed services still use paper-and-pencil or computer-based intelligence tests to screen large groups of applicants. The gold standard, however, is the orally administered, one-on-one IQ test, which requires little or no reading and writing. These include the Stanford-Binet and Wechsler tests, which take between 30 and 90 minutes and combine scores from areas such as comprehension, vocabulary and reasoning to give an overall IQ. These batteries are used to diagnose, treat or counsel children and adults who need personal or academic assistance. Ability testing is governed by detailed ethical standards and professionally administered tests must meet strict criteria including lack of cultural bias and periodic updating. In fact, IQ tests are the most technically sophisticated of all psychological tests and undergo the most extensive quality checks before publication.

Alfred Binet invented the IQ test to identify those schoolchildren most in need of help

US NATIONAL LIBRARY OF MEDICINE/SPL

Average IQ score distribution by population





The "three stratum theory" of intelligence recognises that there is a single general cognitive ability, g, with added input from a range of broad and narrow abilities

JOEMCNALLY/GETTY

WHAT DO IQ TESTS MEASURE?

A century ago, British psychologist Charles Spearman observed that individuals who do well on one mental test tend to do well on all of them, no matter how different the tests' aims, format or content. So, for example, your performance on a test of verbal ability predicts your score on one of mathematical aptitude, and vice versa. Spearman reasoned that all tests must therefore tap into some deeper, general ability and he invented a statistical method called factor analysis to extract this common factor from the web of positive correlations among tests. This showed that tests mostly measure the very same thing, which he labelled the general factor of intelligence or "g factor". In essence, g equates to an individual's ability to deal with cognitive complexity.

Spearman's discovery lay neglected in the US until the 1970s, when psychologist Arthur Jensen began systematically testing competing ideas about g. Might g be a mere artefact of factor analysis? No, it lines up with diverse features of the brain, from relative size to processing speed. Might g be a cultural artefact, just reflecting the way people think in western societies? No, in all human groups – and in other species too – most cognitive variation comes from variation in g.

Jensen's analyses transformed the study of intelligence, but while the existence of g is now generally accepted, it is still difficult to pin down. Like gravity, we cannot observe it directly, so must understand it from its effects. At the behavioural level, g operates as an indivisible force – a proficiency at mentally manipulating information, which undergirds learning, reasoning, and spotting and solving problems in any domain. At the physiological level, differences in g probably reflect differences in the brain's overall efficiency or integrity. The genetic roots of g are even more dispersed, probably emerging from the joint actions of hundreds if not thousands of genes, themselves responding to different environments.

Higher g is a useful tool, but not a virtue. It is especially handy when life tasks are complex, as they often are in school and work. It is also broadly protective of health and well-being, being associated with lower rates of health-damaging behaviour, chronic illness, post-traumatic stress disorder, Alzheimer's and premature death.

Higher g helps an individual get ahead socioeconomically but it has little connection with emotional well-being or happiness. Neither does it correlate with conscientiousness, which is a big factor in whether someone actually fulfils their intellectual potential.

DIFFERENT TYPES OF INTELLIGENCE

Consider the engineer's superior spatial intelligence and the lawyer's command of words and you have to wonder whether there are different types of intelligence. This question was debated ferociously during the early decades of the 20th century. Charles Spearman, on one side, defended the omnipotence of his general factor of intelligence, g. On the other, Louis Thurstone argued for seven "primary abilities", including verbal comprehension (in which females excel) and spatial visualisation (in which males excel). Thurstone eventually conceded that all his primary abilities were suffused with the same g factor, while Spearman came to accept that there are multiple subsidiary abilities in addition to g on which individuals differ.

This one-plus-many resolution was not widely accepted until 1993, however. It was then that American psychologist John B. Carroll published his "three stratum theory" based on a monumental reanalysis of all factor analysis studies of intelligence (see diagram, above left). At the top is a single universal ability, g. Below this indivisible g are eight broad abilities, all composed mostly of g but each also containing a different "additive" that boosts performance in a broad domain such as visual perception or processing speed. These in turn contribute to dozens of narrower abilities, each a complex composite of g, plus additives from the second level, together with life experiences and specialised aptitudes such as spatial scanning.

This structure makes sense of the many differences in ability between individuals without contradicting the dominance of g. For example, an excellent engineer might have exceptional visuospatial perception together with training to develop specialist abilities, but above all a high standing on the g factor. The one-plus-many idea also exposes the implausibility of multiple-intelligence theories eagerly adopted by educators in the 1980s, which claimed that by tailoring lessons to suit the individual's specific strength – visual, tactile or whatever – all children can be highly intelligent in some way.

WHAT MAKES SOMEONE SMART?

Intelligence tests are calibrated so that, at each age, the IQ average score is 100 and 90 per cent of individuals score between IQ 75 and 125. The typical IQ difference between strangers is 17 points and it is 12 between full siblings. Everybody accepts that intelligence varies. But what makes some people smarter than others? How do nature and nurture interact to create that variation as we develop? Are differences in g set at birth, or can we increase someone's intelligence by nurturing them in the right environment?

NATURE AND NURTURE

Each of us is the embodiment of our genes and the environment working together from conception to death. To understand how these two forces interact to generate differences in intelligence, behavioural geneticists compare twins, adoptees and other family members. The most compelling research comes from identical twins adopted into different homes – individuals with identical genes but different environments – and non-kin adopted into the same home – unrelated individuals sharing the same environment. These and other studies show that IQ similarity most closely lines up with genetic similarity.

More intriguingly, the studies also reveal that the heritability of intelligence – the percentage of its variation in a particular population that can be attributed to its variation in genes – steadily increases with age. Heritability is less than 30 per cent before children start school, rising to 80 per cent among western adults. In fact, by adolescence, separated identical twins answer IQ tests almost as if they were the same person and adoptees in the same household as if they were strangers. The surprising conclusion is that most family environments are equally effective for nurturing intelligence – the IQ of an adult will be the same almost regardless of where he or she grew up, unless the environment is particularly inhumane.

Why does the shared environment's power to modify IQ variation wane and genetic influences increase as children gain independence? Studies on the nature of nurture offer a clue. All children enter the world as active shapers of their own environment. Parents and teachers experience this first-hand as their charges frustrate attempts to be shaped in particular ways. And increasing independence gives young people ever more opportunities to choose the cognitive complexity of the environments they seek out. The genetically brighter an individual, the more cognitively demanding the tasks and situations they tend to choose, and the more opportunities they have to reinforce their cognitive abilities.

Given that an individual's ability to exploit a given environment is influenced by their genetic endowment, and given that "better" family environments tend not to produce overall increases in IQ, it is not surprising that attempts to raise low IQs by enriching poor school or home environments tend to disappoint. Narrow abilities can be trained up but g apparently cannot. This makes sense if g is an overall property of the brain. That does not mean intensive early educational interventions lack positive effects: among other things they may reduce rates of teenage pregnancy, delinquency and school dropout. Besides, even if we cannot boost low intelligence into the average range, we do know how to help all children learn more than they currently do and achieve more with the intelligence they have.

"Intriguingly, the heritability of intelligence is less than 30 per cent before children start school, rising to 80 per cent among adults"

Identical twins are a natural laboratory in which to study how intelligence develops

LEFT GETTY; BACKGROUND DOUG CORRANCE/GETTY

OLDER AND WISER

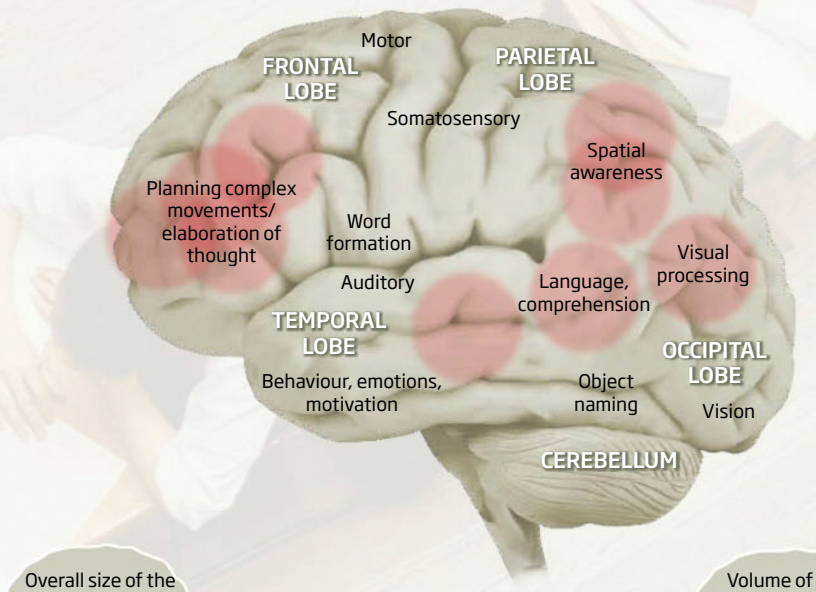
The brain is a physical organ and no less subject than any other to ageing, illness and injury. The normal developmental trajectory is that aptitude at learning and reasoning – mental horsepower – increases quickly in youth, peaks in early adulthood, and then declines slowly thereafter and drops precipitously before death. The good news is that some important abilities resist the downturn.

Some IQ researchers distinguish between tests of fluid intelligence (gF) and crystallised intelligence (gC). The first assess on-the-spot learning, reasoning and problem solving; the second assess the crystallised fruits of our previous intellectual endeavours, such as vocabulary in one's native language and broad cultural knowledge. During youth, gF and gC rise in tandem, but they follow different trajectories thereafter. All gF abilities decline together, perhaps because the brain's processing speed slows down with age. However, most people's gC abilities remain near their personal peak into old age because they reside in the neural connections that gF has laid down over a lifetime of learning and practice. Of course, age-related memory loss will affect an individual's ability to recall, but exactly how this affects intelligence is not yet known.

This has practical implications. On the positive side, robust levels of gC buffer the effects of declining gF. Older workers are generally less able to solve novel problems, but they can often compensate by calling upon their larger stores of experience, knowledge and hard-won wisdom. But gC can also disguise declines in gF, with potentially hazardous results. For example, health problems in later life can present new cognitive challenges, such as complex treatments and medication regimes, which individuals with ample gC may appear to understand when actually they cannot cope.

There are ways of slowing or reversing losses in cognitive function. The most effective discovered so far is physical exercise, which protects the brain by protecting the body's cardiovascular health. Mental exercise, often called brain training, is widely promoted, but it boosts only the particular skill that is practised – its narrow impact mirroring that of educational interventions at other ages. Various drugs are being investigated for their value in staving off normal cognitive decline, but for now preventive maintenance is still the best bet – avoid smoking, drinking to excess, head injuries and the like.

Intelligence is distributed across many areas of the brain and people with the highest IQ tend to have increased volume in a network of regions (shaded) including key language areas



Overall size of the brain, relative to the body, correlates with IQ

Volume of the cortex, the brain's grey matter, correlates with IQ

Intelligence requires integration of sensory and other information

Learning and experience can increase the size of specific brain areas

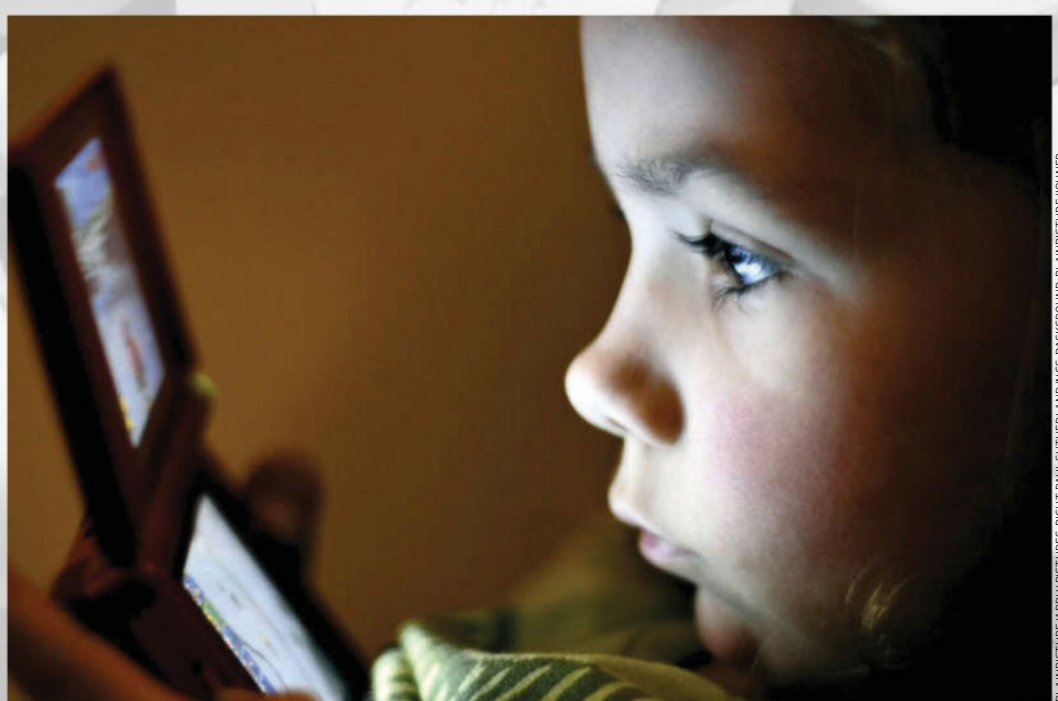
High IQ is associated with faster mental processing speed

Volume of tissue linking the brain's hemispheres correlates with IQ

BOOSTING BRAINPOWER

Who wouldn't like to be more intelligent? If someone invented a safe and effective smart drug that could boost g by 20 points it would surely sell faster than Viagra. Unfortunately, everything we have learned about intelligence indicates that this is highly unlikely. If increasing intelligence is not an option, can you do more with what you have, by finding effective ways to work smarter, perhaps?

REALISING YOUR ASSETS



PLAINPICTURE/APPLY PICTURES; RIGHT: PAUL SUTHERLAND/NGS; BACKGROUND: PLAINPICTURE/JOHNER

IQ tests are designed to measure an individual's maximum cognitive ability but in everyday life we rarely perform at our best. Too often we arrive at work sleep-deprived, stressed, distracted, hungry, sick, addled by medicine or hung-over – all of which reduce cognitive acuity. This is compounded by the fact that many employers fail to recognise that mental performance varies over a day or week. Organisations squander their members' cognitive assets when they pace tasks poorly or flout normal sleep cycles, such as when schools start too early for the typical student, or when shift-workers have to put up with constantly changing schedules.

What's more, to fully realise their abilities, individuals of different intelligence levels often require different kinds of support. Educational and military psychologists have shown that people of below-average intelligence learn best when given concrete, step-by-step, hands-on instruction and lots of practice, whereas individuals of above-average intelligence learn best when allowed to structure their own learning. One-size-fits-all instruction stunts the learning of both types of individuals. Schools can get far more out of pupils by educating them to their personal potential and employers can boost the achievements of their staff with well-targeted assistance such as mentoring, supervision and training.

Brainpower also needs protecting and nurturing. Chronic illness, alcohol abuse and head injuries cause cumulative cognitive damage, accelerating the effects of ageing and increasing the risk of dementia. With vaccinations and care, most such assaults are preventable. We can also reduce exposure to human-made hazards that damage the brain, such as pesticides, lead, radiation and exposure to drugs in the womb. The best way to get the most from our native intelligence right into old age is to maintain good health of both body and mind. Healthy body, healthy mind is a cliché because it's true.

Brain training games can only improve particular skills but not overall intelligence

"As modern life becomes ever more complex, technological upgrades can feel like brain downgrades"

SIMPLIFY YOUR WORLD

Modern life is becoming ever more complex. When parents have to turn to their children to operate the latest electronic gadget, technological upgrades can feel like brain downgrades. The rising complexity of daily life can be a source of humour, embarrassment and inconvenience but, given that the ability to deal with cognitive complexity is the essence of intelligence, this complexity can also be detrimental to personal well-being. One largely overlooked way we can achieve more with the intelligence we have is to recognise this and try to reduce needless complexity in everyday life.

The potentially harmful effects of cognitive overload are particularly clear in the field of healthcare. High rates of non-adherence to treatments are the bane of medical providers, and these increase when treatment plans are morecomplex and patients less intelligent. Given the complexity of self-care regimes, it is hardly surprising that some people make dangerous errors or fail to comply.

Effective management of diabetes, for example, requires a person to keep blood sugar levels within a healthy range, which means coordinating diet, exercise and medication throughout the day, which in turn requires planning for contingencies, recognising when blood sugar is veering too high or low, knowing how to regain control and conceptualising the imperceptible but cumulative damage caused by failing to maintain control. There is no set recipe for people with diabetes to follow – their bodies and circumstances differ. Moreover, they get little training, virtually no supervision and no days off. Effectively managing one's diabetes is a cognitively complex job and poor performance has serious consequences, including emergency room visits, lost limbs or eyesight, and even death. The lower the diabetic person's IQ, the greater the risks.

Attempts to improve health outcomes in situations like this often focus on changing the behaviour of patients, but an equally effective approach might be to lower unnecessary cognitive hurdles to successful prevention, treatment and self-management of illnesses. Many doctors are unaware that even a seemingly simple prescription medicine label or appointment slip may be incomprehensible to some patients. There is wide scope to simplify the cognitive demands on patients and to provide assistance with essential tasks that are inherently complex. And patients who are very susceptible to cognitive overload can benefit from triage, with healthcare providers identifying the behaviours most critical for success and then providing training, monitoring and feedback to ensure they are mastered.

In healthcare and beyond, managing cognitive overload is a great missed opportunity, a chance to reduce the risks of illness, accidents and premature death by reshaping everyday environments to meet people's individual cognitive needs.

COGNITIVE ENHANCEMENT

Brain implants, transplants and downloads may be far in the future, but other forms of cognitive enhancement have a long history. For centuries people have used brain-boosting drugs. Caffeine and nicotine, for example, both increase alertness for short periods. Today there are more choices than ever. One recent survey of US universities found that as many as 25 per cent of students routinely take Ritalin or Adderall to boost memory and concentration – both drugs are actually designed to treat attention-deficit hyperactivity disorder. Another favourite is modafinil, licensed to treat narcolepsy and various sleep disorders, but which can also reduce fatigue and maintain alertness in healthy individuals burning the midnight oil. There are dozens more drugs in the pipeline with the potential for cognitive enhancement – some act on the same nicotinic receptors as cigarettes; others are being developed for the express purpose of augmenting memory.

Even if they are effective, however, such drugs do not increase intelligence, they only enhance certain aspects of cognition such as memory or alertness. And there may be unknown risks associated with them, particularly those that have been developed for other purposes and have had few trials on healthy people. However much we would like to boost our brainpower, many of us are not prepared to take these risks. That might help explain the rise in recent years of what are called "superfoods" as a natural solution to cognitive enhancement. Unfortunately, while eating blueberries, salmon, avocados, and dark chocolate is obviously safer, it may not be as effective as many people hope. If such "brain foods" work at all, it is probably primarily by promoting general health when consumed as part of a wholesome, balanced diet.

In our desire to be cleverer we are constantly on the look-out for new cognitive enhancers. They range from the sublime, such as learning to play a musical instrument, to the impractical, such as transcranial direct current stimulation, which involves placing electrodes on the scalp to zap the brain with a tiny electrical current. Each claims to improve one or more specific abilities such as concentration, visual perception or memory, but the jury is still out on whether these improvements have real-world value.

Perhaps the most universally accessible brain toner is one of the most ancient – meditation. Growing evidence suggests that training in mindfulness meditation improves not just psychological well-being but also produces measurable improvements in a range of cognitive areas, including attention and memory, probably by reducing susceptibility to stress and distraction.

Superfoods may make you healthier but they won't increase your IQ

Linda S. Gottfredson

Linda S. Gottfredson is a professor of education at the University of Delaware in Newark. She focuses on the social implications of intelligence, including how cultural institutions are shaped by the wide variation in human cognitive capability that is characteristic of all groups. She is also interested in the evolution of human intelligence and especially the idea that it may have been driven by a need to overcome novel hazards associated with innovation.

ARE WE GETTING SMARTER?

Over the past century, each successive generation has answered more IQ test items correctly than the last, the rise being equivalent to around 3 IQ points per decade in developed nations. This is dubbed the "Flynn effect" after the political scientist James Flynn, who most thoroughly documented it. Are humans getting smarter, and if so, why?

One possible explanation is that today's world supports or demands higher levels of intelligence. Flynn himself suggests that intelligence has risen in part because we view life more analytically, through "scientific spectacles". However, the idea that cultural environments have potent and widespread effects on how smart we are does not square with what we know about the high heritability of intelligence. Environmental variation contributes relatively little to the IQ differences in a birth cohort as its members mature over the decades. How, then, could it create such big IQ differences across successive birth cohorts living in the same era?

Another theory puts rising IQ down to physiological changes. In the past century, human height has been increasing in tandem with IQ throughout the developed world. Better public health measures have reduced the need for our immune systems to consume resources to combat infectious disease, leaving us able to spend more on growth – and larger, smarter brains may be just one consequence. Not only that, as more people travelled and married outside their local group, populations may have benefited genetically from hybrid vigour. Inbreeding is known to lower intelligence, and outbreeding can raise it.

It is also possible that the Flynn effect does not in fact reflect a rise in general intelligence, or g. After all, can the average IQ of adults at the end of the second world war really have been 20 points less than today? That would put them in the bottom 10 per cent of intelligence by today's standards, making them legally ineligible to serve in the US military on grounds of poor trainability. It defies belief.

Instead of an overall increase in g, perhaps just certain biologically rooted cognitive abilities are increasing. An IQ test comprises a series of subtests, and it turns out that scores in some of these have increased a lot – including our ability to identify similarities between common objects – whereas others have not increased at all – such as scores in the vocabulary and arithmetic subtests. That would imply changes in specific brain regions rather than the whole brain.

The inter-generational rise in IQ test scores is a brain-twister for researchers trying to figure out what it means. Nevertheless, it does not undermine the use of IQ tests within generations. Today's IQ tests are not intended to give an absolute measure of intelligence akin to grams and kilograms, but only to rate an individual's intellectual capacity relative to others born in the same year – no matter what the cohort, the mean score is always set at 100. As for the variation in g that IQ tests measure, it seems as wide and as consequential as ever.

RECOMMENDED READING

Intelligence: A Very Short Introduction by Ian Deary (Oxford University Press, 2001)

The Genetic and Environmental Origins of Learning Abilities and Disabilities in the Early School Years by Yulia Kovas and others (Blackwell, 2007)

The g Factor: The Science of Mental Ability by Arthur Jensen (Praeger, 1998)

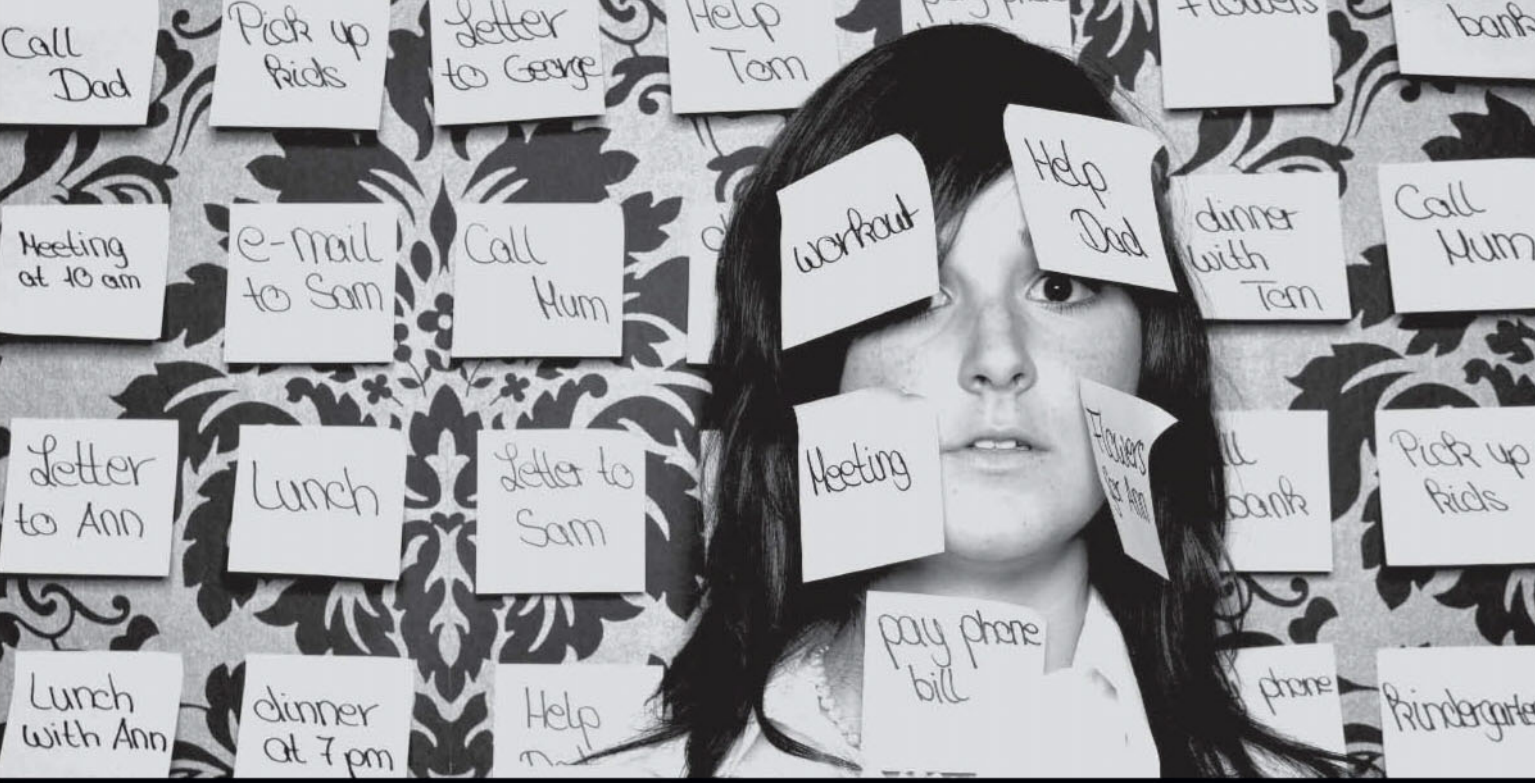
Correcting Fallacies about Educational and Psychological Testing edited by Richard Phelps (American Psychological Association, 2009)

Intelligence, vol 24(1) (special issue called Intelligence and Social Policy)

Intelligence, vol 37(2) (special issue called Intelligence and the Brain)

Journal of Personality and Social Psychology, vol 86, p 96 (the Special Section on Cognitive Abilities)

Cover image
Joe McNally/Getty



MEMORY

JONATHAN K. FOSTER

INSTANT EXPERT

WHAT IS MEMORY?

Remembering the past is an integral part of human existence. Without a good memory, you would not be able to drive to work, hold a meaningful conversation with your children, read a book or prepare a meal.

Memory has fascinated humans since ancient times; Plato famously compared our memory to a wax tablet that is blank at birth and slowly takes on the impression of the events from our life. Only in the past hundred years, though, have psychologists developed systematic objective techniques that have enabled us to study our recollections of the past with scientific accuracy and reproducibility. These range from laboratory tests of our ability to remember verbal and visual materials to more recent brain-imaging approaches.

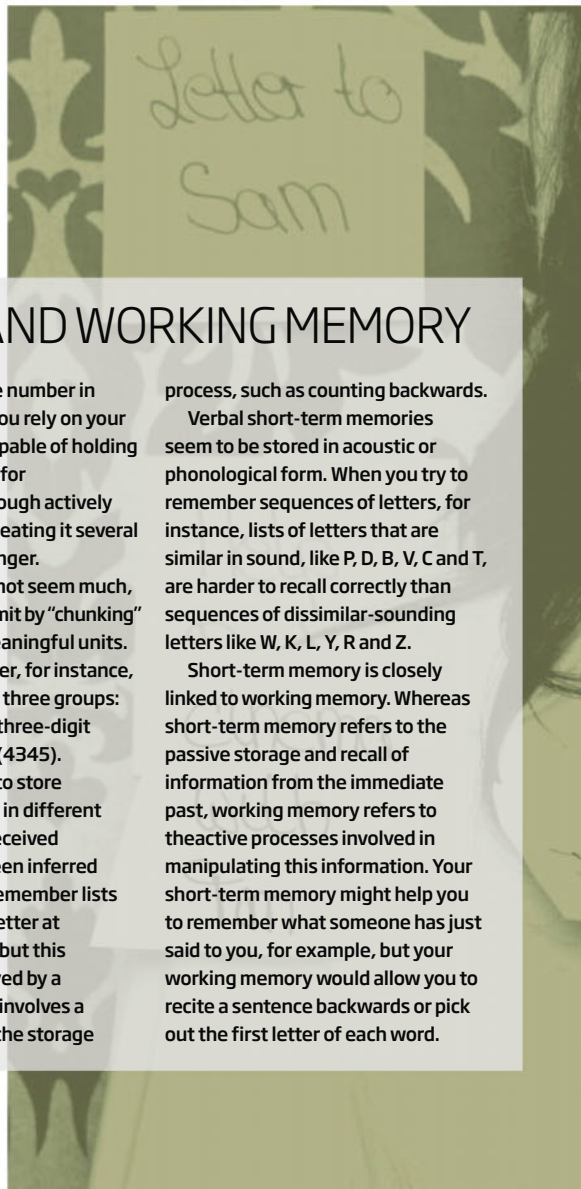
It has become clear through these studies that, unlike Plato's wax tablet, human memory comprises many different components. If you consider how long a memory lasts, for example, there appear to be at least three subtypes of storage: sensory, short term and long term. Memories can also be distinguished by the type of information that is stored.

SENSORY MEMORY

During every moment of an organism's life, its eyes, ears and other sensory organs are taking in information and relaying it to the nervous system for processing. Our sensory memory store retains this information for a few moments. So twirling a sparkler, for example, allows us to write letters and make circles in the air thanks to the fleeting impression of its path.

Johann Segner, an 18th-century German scientist, was one of the first people to explore this phenomenon. He reportedly attached a glowing coal to a cartwheel, which he rotated at increasing speeds until an unbroken circle of light could be perceived. His observations were followed by the systematic investigations of American psychologist George Sperling 100 years later. By studying people's ability to recall an array of letters flashed briefly on a screen, he found that our fleeting visual impressions – dubbed "iconic memory" – last for just a few hundred milliseconds. Studies of "echoic" sound memories followed soon afterwards, showing that we retain an impression of what we hear for several seconds. Of note, echoic memories may be impaired in children who are late talkers.

Sensory memories are thought to be stored as transient patterns of electrical activity in the sensory and perceptual regions of the brain. When this activity dissipates, sensory memory fades. While they last, though, sensory memories provide a detailed representation of the entire sensory experience, from which relevant pieces of information can be extracted into short-term memory and processed further via working memory.



SHORT-TERM AND WORKING MEMORY

When you hold a restaurant's phone number in your mind as you dial the number, you rely on your short-term memory. This store is capable of holding roughly seven items of information for approximately 15 to 20 seconds, though actively "rehearsing" the information by repeating it several times can help you to retain it for longer.

Seven items of information may not seem much, but it is possible to get around this limit by "chunking" larger pieces of information into meaningful units. To recall a 10-digit telephone number, for instance, a person could chunk the digits into three groups: the area code (such as 021), then a three-digit chunk (639) and a four-digit chunk (4345).

Your short-term memory seems to store verbal and visuospatial information in different subsystems. The verbal store has received most attention. Its existence has been inferred from studies asking volunteers to remember lists of words: people tend to be much better at recalling the last few items in a list, but this effect disappears if the test is delayed by a fewseconds, especially if the delay involves a verbal activity that interferes with the storage process, such as counting backwards.

Verbal short-term memories seem to be stored in acoustic or phonological form. When you try to remember sequences of letters, for instance, lists of letters that are similar in sound, like P, D, B, V, C and T, are harder to recall correctly than sequences of dissimilar-sounding letters like W, K, L, Y, R and Z.

Short-term memory is closely linked to working memory. Whereas short-term memory refers to the passive storage and recall of information from the immediate past, working memory refers to theactive processes involved in manipulating this information. Your short-term memory might help you to remember what someone has just said to you, for example, but your working memory would allow you to recite a sentence backwards or pick out the first letter of each word.

LONG-TERM MEMORY

Important information is transferred to the brain's long-term storage facility, where it can remain for years or even decades. Your date of birth, phone number, car registration number and your mother's maiden name are all held here.

Unlike short-term memory, we seem to store long-term memories by their meaning. If you try to recall significant information after a delay, for instance, you probably won't be able to reproduce the exact wording but you should be able to recall its meaning or gist. This can lead to errors, however.

Long-term memories can take different forms. Semantic memories, for example, concern your knowledge of facts, such as Paris being the capital of France, though you may not remember the exact circumstances in which you acquired this information. Episodic memories concern events from your life that occurred at a specific time and place, such as the day you passed your driving test.

Long-term memories can also be categorised by the way they influence your behaviour. Consciously recalled events or pieces of information are termed explicit memories, whereas implicit memory refers to experiences that influence your behaviour, feelings or thoughts without you actively recollecting the events or facts. For instance, if you pass an Italian restaurant on the way to work in the morning, you might think about going out for an Italian meal that evening, without being aware that you had been influenced by your morning journey.

From early influential work by Canadian psychologist Donald Hebb in the 1940s through to the more recent Nobel-prizewinning work of American neuropsychiatrist Eric Kandel, we know that long-term memories are maintained by stable and permanent changes in neural connections. And through brain imaging techniques we now have the capacity to study these processes non-invasively. The diencephalon and the hippocampal regions seem to be essential for consolidating information from short-term into long-term memories (see diagram, above left). Their exact roles in retrieving older memories from across our lifespan is still contentious; the hippocampus might be part of the "library" system that stores our long-term memories, or instead just the "printing press" which creates these memories.

"Twirling a sparkler allows us to write letters and make circles in the air thanks to our sensory memory"

MIKE HEWITT/GETTY; BACKGROUND: PLAINPICTURE/ANDRE SCHUSTER

THE MAKING OF MEMORIES

Unlike a DVD or computer hard drive, our brains do not faithfully replay our previous experiences as they happened. Certain pieces of information and events are retained but others disappear or become distorted, and sometimes we even seem to remember things that never happened. Over the past few decades, researchers have developed a good understanding of the factors that influence the contents of our memories.

ENCODING AND PROCESSING

Several important factors influence how well we remember, such as the way we process information at the time of encoding. In the 1970s, Fergus Craik and Robert Lockhart at the University of Toronto, Canada, described different levels of processing, from the "superficial", dealing only with the physical properties of the material to be remembered, through to "deeper" processes such as the meaning of the material. Deeper encoding produced better memories. Suppose you were asked to study a list of words, for instance. If you were asked to provide a definition of each word as you studied it, you would be much more likely to remember the list than if you were asked to count the number of vowels in each word. This finding illustrates that we are active, rather than passive, agents in the remembering process.

Context can also exert important effects on memory. For example, in some famous studies from the 1970s, divers were asked to remember lists of words when they were either on the shore or beneath the water's surface. They were then asked to recall the information at a later time – again, either on land or underwater. Their recall turned out to be substantially better when the learning and test sessions occurred in the same environment; for example, the divers remembered the information they had learned under the sea more clearly if they were tested underwater than if they were tested on the beach.

Your physiological or psychological state can exert a similar influence on your memory. This is potentially significant for students studying for exams. If you revise while you are calm, but then feel nervous or threatened in the actual examination, you might not recall information so well compared with someone whose mood is more even across study and test. Equally, you may be more likely to remember a drunken night's escapades when you next drink alcohol than when you are sober.

MATTHEW ODFIELD/SPL; RIGHT: GPL/EVERETT/REX FEATURES

Memories formed under the sea are best recalled underwater, showing that context plays an important role in recollection

IMAGINED MEMORIES

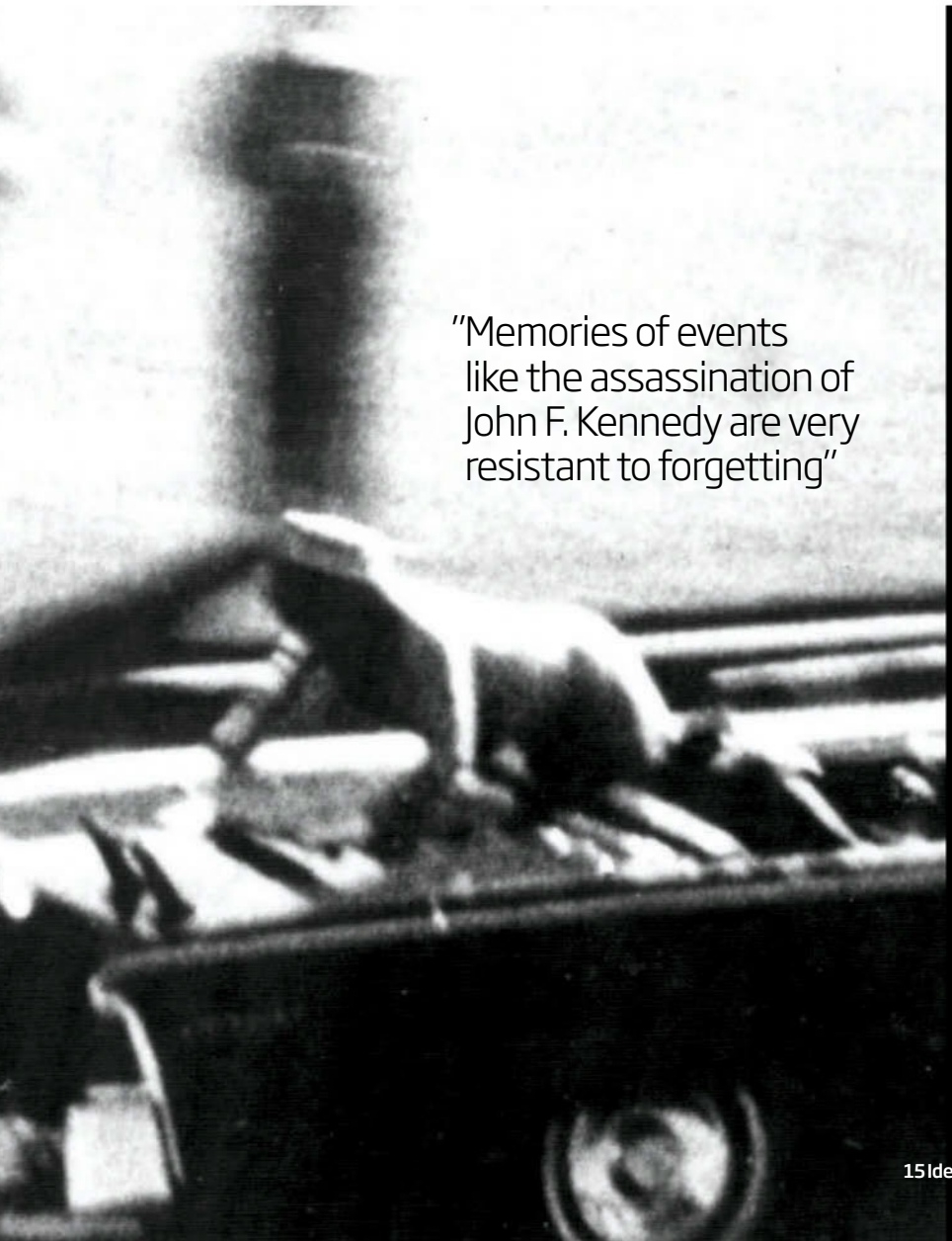
It is very easy to lead someone's memory astray. For example, if I witness a traffic accident and I am later asked whether the car stopped before or after the tree, I am much more likely to "insert" a tree into my memory of the scene, even if no tree was actually present. This occurrence reflects the fact that when we retrieve a memory, we also re-encode it and during that process it is possible to implant errors.

Elizabeth Loftus at the University of California, Irvine, and colleagues have shown that this "misinformation effect" can have huge implications for the court room, with experiments repeatedly demonstrating that eyewitness testimonies can be adversely influenced by misleading questioning. Fortunately, these findings also suggest ways for police, lawyers and judges to frame the questions that they ask in a way that makes reliable answers more likely.

Related to the misinformation effect are "recovered" and false memories. A team led by Henry Roediger and Kathleen McDermott at Washington University in St Louis, Missouri, has built an extensive body of research showing that false memories can be induced relatively easily. People can be encouraged to "remember" an item that is linked in its meaning to a series of previously presented items but which itself was not presented. Researchers have also shown that misleading information can create memories of personal events that the individual strongly believes to have happened in their past but which never took place. In one famous experiment, Loftus persuaded subjects that they had seen Bugs Bunny at Disneyland, despite the fact that Bugs is a Warner Brothers character. Such findings may represent a serious concern for legal cases in which adults undergoing therapy believe that they have recovered genuine memories of abuse in childhood.

SUPERSTOCK/GETTY

Attempts to recover repressed memories on the therapist's couch may lead patients to "recall" imagined events



"Memories of events like the assassination of John F. Kennedy are very resistant to forgetting"

FLASHBULB MEMORIES AND THE REMINISCENCE BUMP

Certain events seem to stay with us as particularly vivid and detailed memories for years afterwards. This is especially true if the events are unusual, arousing or associated with strong emotions.

The assassination of John F. Kennedy in 1963, the death of Princess Diana in 1997, and the destruction of 9/11, for instance, are all very memorable for people who were alive when these events occurred. Memory for such events appears to be very resistant to forgetting – many people are able to remember where they were and who they were with when they heard the news, even decades later. This effect has been termed flashbulb memory.

Another common phenomenon, known as the reminiscence bump, refers to the wealth of memories that we form and store between adolescence and early adulthood. When we are older, we are more likely to remember events from this period than any other stage of life, before or after. It could be that the reminiscence bump is due to the particular emotional significance of events that occur during that period. These events include meeting one's partner, getting married or becoming a parent, and events that are life-defining in other ways, such as starting work, graduating from university or backpacking around the world.

Flashbulb memories and the reminiscence bump are the subject of much debate. For instance, some have questioned whether the rich details that we seem to recall in a flashbulb memory may in fact have merely been inferred from our general knowledge of the event – we may recall details of the fatal car accident in which Princess Diana died due to the fact that it has been replayed in the media many times since 1997.

MEMORY LOSS

All too often, our memory can fail us. We all forget names and important facts from time to time, but in the most serious forms of amnesia people may have no concept of their recent past whatsoever. Understanding memory failings can help researchers to unravel how we form and retain memories.

WHY DO WE FORGET?

Any effective memory device needs to do three things well: encode information and represent it in a storable form, retain that information faithfully and enable it to be accessed at a later time. A failure in any of these components leads us to forget.

Distraction or reduced attention can cause a memory failure at the encoding stage, while a problem in storage – following brain injury, for example – can cause encoded information to be deleted.

JAMES PORTER/WORKBOOK/GETTY; BACKGROUND: PLAINPICTURE/CLAUDIA BELOUR

Memories can also become less distinctive if the storage of other memories interferes with them, perhaps because they are stored in overlapping neural assemblies.

Often, memory problems occur when we try to retrieve information, leading to the feeling that a fact is "on the tip of my tongue". For example, it can be intensely frustrating to forget someone's name at a party, only to remember it a few hours later. This problem might arise because our search algorithms aren't perfect and an appropriate "cue", needed to enable the right information to be accessed, is not available.

Despite its drawbacks, forgetting can also be useful and adaptive. Other things being equal, we tend to remember things that are important for our functioning and survival; for example, information that is potentially rewarding or threatening.

Names often escape us, perhaps because the brain can't distinguish the signal from neural noise

AMNESIA

The "amnesic syndrome" (also known as classical amnesia) is one of the purest forms of memory impairment. It is typically caused by a brain injury either to the hippocampus or a nearby region called the diencephalon. These regions are critically involved in consolidating memories into long-term storage. In the case of patient S.J., whom my colleagues and I have studied, the profound amnesia was due to an apparently very focal lesion of the hippocampus caused by an infection. In the case of the patient known as N.A., the damage was caused by an accident in which a fencing foil penetrated his brain via a nostril, damaging the diencephalon.

People with classical amnesia lose their ability to recall events before the brain injury (known as retrograde amnesia), and also cannot lay down new memories after the injury (termed anterograde amnesia). The amnesic syndrome does not affect all forms of memory, however. Knowledge of speech, language and other elements of semantic memory are usually preserved, as is short-term memory. People with the condition can also remember specific skills, such as how to drive, and sometimes master new abilities, even though they typically can't remember the events that led them to learn the skill!

By contrast, the ability to retain new information over any significant period of time is profoundly affected in people with classical amnesia; they find it almost impossible to learn the name of the person they just met, for instance.

The study of memory impairment has considerably informed our knowledge about how memory operates in the fully functional state. As psychologist Kenneth Craik stated: "In any well-made machine one is ignorant of the working of most of the parts... it is only a fault which draws attention to the existence of a mechanism at all."

CAN I IMPROVE MY MEMORY?

At the moment, none of us can reliably improve the biological hardware involved in memory, though it is comparatively easy to damage it via drug and alcohol abuse or injury. So-called "smart drugs" are claimed to improve the functioning of our memory circuits, but although some treatments have been shown to help people with impaired memory due to brain damage or illness, their effects seem to be limited in healthy people.

We can ensure that we make the best possible use of our memory by living and eating healthily, and by using a range of mnemonics, some of which have been known for thousands of years. Most mnemonics are based on the principles of reduction or elaboration. As the name suggests, a reduction mnemonic reduces the information to be remembered (through an acronym like the words Roy G Biv, which helps children to remember the colours of the rainbow). An elaboration code increases the information to be retained, perhaps through a catchy phrase like "Richard of York gave battle in vain", which again is used to remember a rainbow's colours. More detailed systems include the peg word system, which uses the power of visual imagery by assigning a memorable rhyming word to a number :"one is bun", "two is shoe", "three is tree". A list can then be remembered by linking each item in the sequence to each peg word, through the memorable image.

There are many other ways of increasing your chances of recall, such as by elaborating or rehearsing information, organising it in a new way, or attempting to explain what you are studying to someone else. Timing study periods so that you attempt to remember a piece of information after steadily increasing intervals can also help to lay the foundations for effective long-term recall.

"We can make the best use of our memory by living and eating healthily and by using mnemonics"

Jonathan K. Foster

Jonathan K. Foster is a clinical neuropsychologist and professor of clinical neuroscience. He is affiliated to Curtin University, the Neurosciences Unit of the Health Department of Western Australia and the University of Western Australia. He worked previously in the UK and North America. He maintains an active clinical practice consulting patients with memory loss and other types of cognitive impairment.

FORESTALLING DECLINE

There is decline in many types of memory with ageing, but not in semantic memory or general knowledge which typically increases across the years and is well preserved with age. Age-related memory decline is most pronounced in the condition known as dementia.

As many as 1 in 20 people will develop some form of dementia by 65 years of age – a figure that rises to 1 in 5 in the over-80s. The condition is extremely distressing, because affected individuals deteriorate irreversibly until they essentially become a shell of their former selves. The level of care they need can affect many other family members too. With the so-called "demographic time bomb" of an ageing population in many countries, it seems this major societal issue will only become more important over time.

The various factors that might predict the onset of age-related memory loss and dementia are the focus of much research. For example, studies have identified genetic factors that seem to influence memory loss in later life. Specifically, possession of the epsilon 4 variant of the *APOE* gene appears to represent an important risk factor for later-onset dementia. The identification of this and other risk factors (such as lower levels of cognitive stimulation, exercise and a higher body mass index) may help us to prioritise more vulnerable individuals for screening, intervention and possible future treatment.

Indeed, in some of our own research, my collaborators and I have shown that moderate physical activity may help to counter age-related cognitive decline. What's more, it now seems that conditions such as diabetes may be an important risk factor for Alzheimer's disease – the most common form of dementia. These discoveries may lead to ways to limit memory decline as we age.

RECOMMENDED READING

Memory: A very short introduction by Jonathan K. Foster (Oxford University Press, 2008)

Memory: Systems, Process or Function by Jonathan K. Foster & Marko Jelicic (Oxford University Press, 1999)

Memory by Alan Baddeley, Michael W. Eysenck & Michael C. Anderson (Psychology Press, 2009)

Cover image:
Plainpicture/Andre Schuster

SLEEP

DERK-JAN DIJK
RAPHAËLLE WINSKY-SOMMERER

INSTANT
EXPERT

THE ORIGINS AND PURPOSE OF SLEEP

Birds, fish, reptiles and other mammals have at least one thing in common with us: they sleep. Sleep is a central part of our lives and is clearly crucial. Although sleep has fascinated philosophers, writers and scientists for centuries, it wasn't until the early 1950s that scientific research began in earnest. Since then, sleep science has revealed much about the structure and patterns of sleep. Even so, its origins and functions remain largely mysterious

WHAT IS SLEEP?

Strictly speaking, the term "sleep" only applies to animals with complex nervous systems. Nevertheless it is possible to identify sleep-like states in invertebrates that allow us to define sleep more broadly. These include cycles of rest and activity, a stereotypical body position, lack of responsiveness and compensatory rest after sleep deprivation. Insects in particular have a state very similar to sleep, as do scorpions and some crustaceans.

Even microorganisms, which lack a nervous system, have daily cycles of activity and inactivity driven by internal body clocks known as circadian clocks. The origins of sleep might therefore date back to the dawn of life 4 billion years ago, when microorganisms changed their behaviour in response to night and day.

Some researchers consider sleep part of a continuum of inactive states found throughout the animal kingdom. Once we understand exactly which aspects of an organism benefit from these states, we may be able to provide a meaningful answer to the question of whether simple organisms sleep.

M. WATSON/ARDEA.COM; BELOW: HANK MORGAN/SPL

All animals with complex nervous systems sleep. Lower animals and even microorganisms display sleep-like rest states

"Insects in particular have a state very similar to sleep, as do scorpions and some crustaceans"

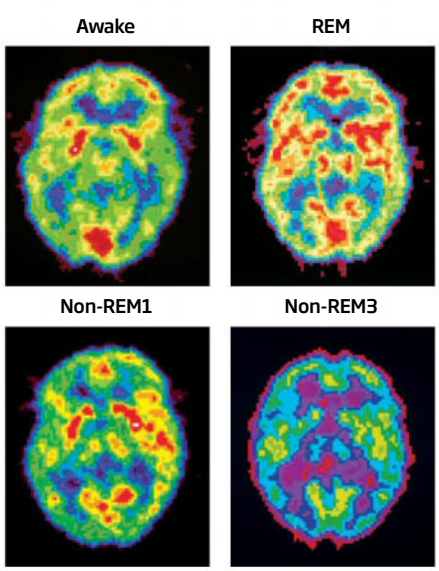
REASONS FOR REST

There are many explanations for sleep, ranging from keeping us out of harm's way to saving energy, regulating emotions, processing information and consolidating memory. Each has strengths – and weaknesses too. Rather than seek a single, universal function of sleep we might do better to study its influence at each level of biological organisation.

At the level of the whole organism a primary function of sleep may be the regulation of autonomic nervous activity such as heart rate; sleep disorders are often associated with dysfunction of the autonomic nervous system, such as an abnormal heartbeat. At the level of the brain it may support memory consolidation by reducing the amount of information travelling through the central nervous system. However, memory consolidation occurs when we are awake too.

At the level of nerve cells, sleep alters firing rates of neurons and also changes the temporal distribution and synchronisation of firing across networks of cells, which may alter their connectivity. The regulation of nerve-cell connectivity, called synaptic homeostasis, can help prevent the nervous system from becoming overloaded. Support for this idea has come from recent studies of fruit flies.

One neglected role of sleep in humans is social isolation. As social animals, we may need sleep to consolidate the rules and insights of our complex social lives.



PET scans show differences in brain activity between wakefulness and various sleep states. Activity is red, inactivity blue (see "Slumber cycles", above right)

SLUMBER CYCLES

During sleep, complex changes occur in the brain. These can be observed with an electroencephalogram (EEG), which measures the brain's electrical activity and associated brainwaves.

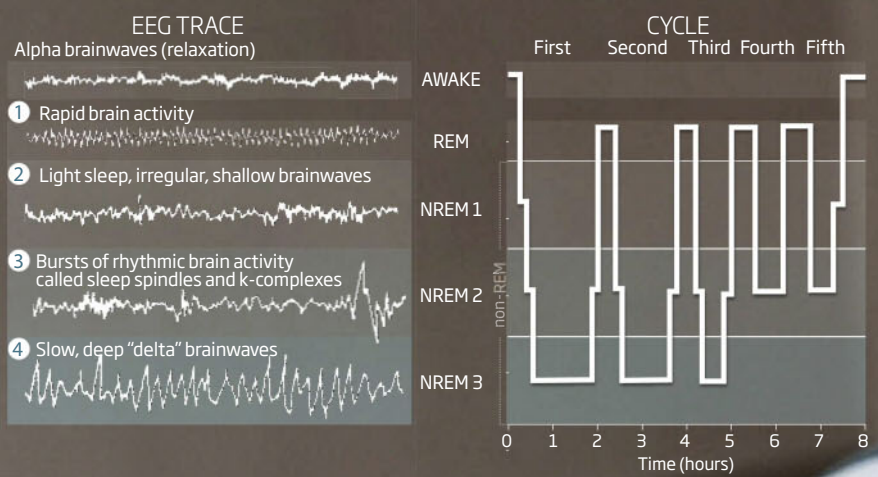
After lying awake for 10 minutes or so we enter non-rapid eye movement sleep or NREM sleep. NREM sleep is divided into three stages, NREM1, NREM2 and NREM3, based on subtle differences in EEG patterns. Each stage is considered progressively "deeper".

After cycling through the NREM stages we enter rapid-eye-movement or REM sleep. The EEG during REM sleep is similar to wakefulness or drowsiness. It is during this stage that many of our dreams occur.

Each cycle lasts for about 1.5 hours and a night's sleep usually consists of five or six cycles.

In addition to changes in brain activity, sleep is also characterised by a reduction in heart rate of about 10 beats per minute, a 1 to 1.5 °C fall in core body temperature as well as a reduction in movement and sensation.

Sleep scientists break sleep into four distinct stages
A typical night's sleep involves several cycles, which includes both REM and non-REM (NREM) sleep



HALF AWAKE

Sleep feels like an on-or-off condition, but brains can be awake and asleep at the same time. This phenomenon is well known in dolphins and seals, animals that can sleep "uni-hemispherically": one half of their brain is asleep while the other half shows electrical activity characteristic of wakefulness.

A study in rats found that after prolonged wakefulness, some neurons go offline and display sleep-like activity. This mosaic brain state is accompanied by occasional lapses in attention. Sleep researchers are investigating if human and other animal sleep is a "global" state or whether the process of sleep can, to some extent, be regulated locally. There is mounting evidence for the latter. For example, the most active brain regions during wakefulness subsequently undergo deeper sleep for longer.

This localised view of sleep could lead to a better understanding of cases when wakefulness intrudes into sleep, such as in sleep-talking, sleepwalking and episodes of insomnia in which people report being awake all night even though recording brainwaves (see "Slumber cycles", above) from a single location suggests they have been asleep.

It also promises to explain how sleep can intrude into wakefulness, such as during lapses of attention when we are sleep-deprived. These "micro sleeps" can be particularly dangerous when driving and various ways to detect them have been developed, for instance by monitoring how a car moves relative to white lines on roads or analysing the movements of the eyes for signs of sleepiness .

PASCAL AMAR/TENDANCE FLOUET

SLEEP PATTERNS

New insights into our sleep requirements – from the amount we need to when, where and how we do it – have been supplied by recent research. The bottom line is that the quality of sleep varies from person to person and also changes throughout life

HOW MUCH IS ENOUGH?

A newborn baby may sleep as much as 18 hours a day, while a middle-aged executive may manage on as few as 5 hours. But how much is healthy and how much do we need? The short answer is that there is no answer: your needs depend on your age and gender, and it varies between individuals.

Young animals generally sleep more than adults and humans are no exception. The structure and intensity of sleep is different too. In young people there is a preponderance of REM sleep, and non-REM sleep is very deep, probably to aid brain maturation. Deep sleep and REM sleep both contribute to plasticity of neuronal networks, which could help with the acquisition of new skills. A recent study found babies sleep more during growth spurts.

During adolescence, sleep becomes shallower and shifts to later hours, reflecting extensive brain rewiring. The frontal lobe – responsible for executive functions such as planning and inhibiting inappropriate behaviour – shows a marked fall in synapse density as the result of neuronal pruning. Teenagers are not just being lazy when they don't want to get out of bed. Their adolescent biology may also prefer an adjustment of school hours.

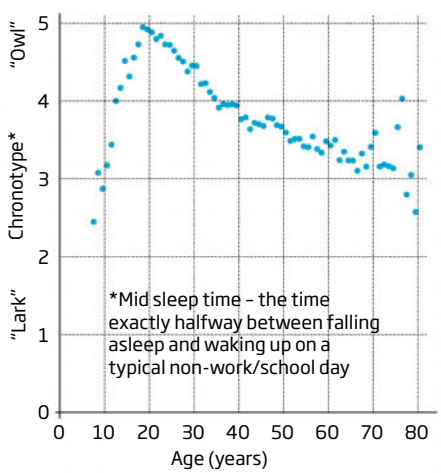
There is good evidence that young people don't get enough sleep. When they live on an 8-hour sleep schedule they remain sleepy, and much more so than older people on the same schedule. If young adults are forced to stay in bed in darkness for 16 hours a day they initially sleep for as long as 12 hours. However, after several days they level off to just under 9 hours, showing that they were paying off a sleep debt.

As we age, we sleep less and spend less time in both deep sleep and REM sleep. At the same time, learning new skills proves to be more difficult. One fascinating question is whether, by preventing age-related changes in sleep, we can halt the age-related decline in mental dexterity.

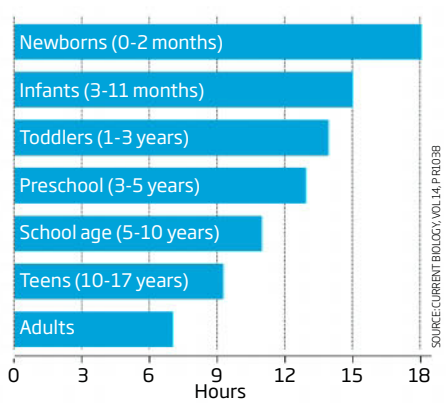
If older volunteers are forced to stay in bed in darkness, they tend to sleep for around 7.5 hours. Studies of those who live longest find they report sleeping between 6 and 7.5 hours a night.

Gender differences also occur in sleep and circadian rhythms. Such differences have been seen in cats, rodents, fruit flies and humans. Women enjoy a greater quantity of deep sleep and sleep longer. There are gender differences in the circadian clock too: the period of the clock is 6 minutes shorter in women than in men and this may explain why on average women go to bed earlier, wake up earlier and are more likely to rate themselves as "morning types" than men.

Teenagers become increasingly "owlish" as they approach the age of 20, then gradually go back the other way



How much sleep do we need?



"It is possible to improve sleep by developing better evening lighting and being more aware of the effects of artificial light"

CATHERINE LARRE/MILLENNIUM

WHAT CONTROLS SLEEP?

The human body and many of its functions are synchronised with a 24-hour cycle via numerous biological clocks. For instance, brain activity while we are awake is largely dependent on the time of day. The circadian timing system is made up of a core set of genes, the protein products of which act in interconnected feedback loops to precisely regulate their own production over a 24-hour cycle. Although the "central" biological clock, located in the brain, largely dictates wakefulness and sleep, other factors come into play.

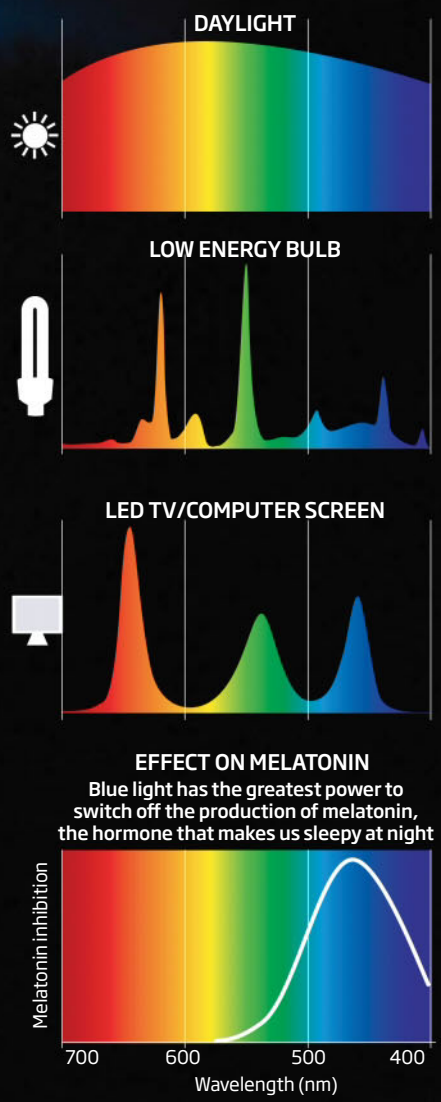
A key factor is light (see diagram, below right). We all know that it is difficult to fall asleep in a brightly lit room. This is because light stimulates specialised cells in the retina, called intrinsically photosensitive retinal ganglion cells, which synchronise the biological clock to the day-night cycle and stimulate other brain areas that are involved in alertness. Even ordinary room light, or the light from a computer screen, can influence the clock and suppress secretion of the sleep-inducing hormone melatonin.

A recent study carried out at the Surrey Sleep Research Centre in Guildford in the UK, found that reducing the intensity of evening light, and/or using a light that contains less blue and more yellow, minimises the disruptive effect on sleep. This suggests that it is possible to improve sleep by developing better evening lighting and being more aware of the disruptive effects of artificial light.

However, even in bright sunlight it can be difficult to stay awake if you have been active for a very long time. This "sleep pressure" is largely created by a neuromodulator called adenosine. Caffeine, the most widely used stimulant in the world, blocks the receptors where adenosine acts in the brain.

Sleeplessness blues

Artificial light and computer screens can play havoc with circadian clocks because they contain wavelengths of light that suppress the secretion of the sleep-promoting hormone melatonin



SOURCE: JOURNAL OF NEUROSCIENCE, VOL 21, P 6405 / JOURNAL OF APPLIED PHYSIOLOGY, VOL 110, P 1432

BAD DAY, GOOD NIGHT

We often sleep well on holiday thanks to an absence of alarm clocks and work-related stress. Lab studies have indeed shown that worry means shallower sleep and more waking up. We also know that age has a profound effect on sleep quality. Older people are more susceptible to the sleep-disrupting effects of stress, caffeine and alcohol. For such people, a few drinks in the evening may severely disrupt sleep in the second half of the night.

Unsurprisingly, the best time to sleep is at night. If we attempt to sleep during the day without shifting our biological clock, we tend to slumber for between 1 and 3 hours less than at night.

Daytime sleep is also different in quality. During the night, REM sleep increases as sleep progresses but during the day it often decreases. Sleep spindles, brainwaves that are characteristic of non-REM sleep and are implicated in memory consolidation, are more abundant at night. Other differences can be seen in the rest of the body. Sleep at night results in a lower body temperature and higher concentrations of the sleep hormone melatonin, whereas the reverse occurs during daytime sleep. These changes can be harmful to health.

Night shift workers often have disrupted daytime sleep and in the long run tend to have a higher risk of cardiovascular diseases and diabetes.

THE FUTURE OF SLEEP

Society has changed enormously over the past century, not least in the way we sleep. Light bulbs, televisions, computers and shift work have all altered traditional sleeping habits. This has huge implications not only for sleep deprivation, but also for our health. New discoveries about sleep and sleeplessness may help us cope with future change

SLEEPLESS SOCIETY?

It is often said that we sleep less than we used to. Half a century ago people supposedly slept for more than 8 hours a night while we now sleep for 7 on average.

The evidence for this is not particularly convincing, simply because 50 years ago scientists weren't too interested in how much people were sleeping. However, there is no doubt that some of us are sleep-deprived. Fortunately we also understand the effects of sleeplessness better than ever.

Sleeplessness is bad for you. A fifth of road accidents are linked to fatigue. Sleep deprivation also affects health directly: the past decade has seen several epidemiological studies linking sleeplessness to ill health and higher mortality rates. For example, there is a link between a decrease in sleep and a rise in obesity and diabetes, which might be explained by the way sleep deprivation disrupts appetite mechanisms and stimulates hunger.

A 17-year study of more than 10,000 British civil servants found that those who had cut their sleep from 7 to fewer than 5 hours were 1.78 times more likely to die from all causes. For example, risk of death from cardiovascular disease rose by 2.25 times.

It is not just sleep deprivation that is unhealthy. Sleeping too long is also associated with an increased risk of dying compared with people who sleep for the optimum period, though the reasons for this are unknown (see diagram, right).

A better understanding of insomnia is also urgently needed, since it damages the performance of businesses to the tune of around $60 billion per year in the US alone.

There are more subtle effects of sleeplessness. One study suggests that sleep deprivation increases the activity of the amygdala, a region of the brain involved in the regulation of anxiety and mood. In addition, some form of sleep disruption is found in almost all psychiatric disorders.

Sleepy people are hungry people, which may explain the link between sleep deprivation, obesity and diabetes

"The light-dark cycle is the natural timekeeper and light remains the alarm clock par excellence"

NEW WAYS TO SNOOZE

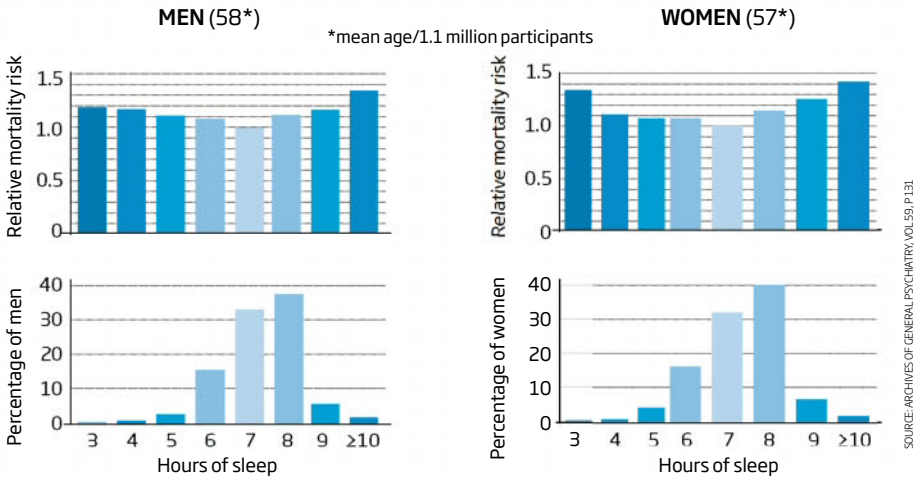
Sleeping pills (or hypnotics) are designed to mimic the mechanisms of natural sleep, which involve many molecules that transmit information in the brain and body, such as neurotransmitters and hormones. Most hypnotics target specific neurotransmitters and the receptor sites in the brain on which they act, especially GABA, the brain's principal inhibitory signalling molecule. Some hypnotics target hormones such as melatonin.

Most of these drugs are effective, the newer ones especially so. But they all alter sleep structure and brainwaves to some extent and most sleep medications have side effects such as grogginess the next morning or memory loss for events that happened while awake during the night.

Based on new understanding of the circuits involved in the sleep-wake cycle, there are drugs in development for the treatment of insomnia. They target several proteins in the brain on which messenger chemicals act, such as the receptors for serotonin, orexin, histamine or melatonin. Some of these new drugs appear to change only the timing or duration of sleep, without changes in the generation of brainwaves. If these findings are supported by bigger studies, these drugs promise to be valuable advances on current sleeping pills.

THE PERFECT SEVEN

Sleeping less than 7 hours a night raises your chances of dying sooner - but so does sleeping for too long



SOURCE: ARCHIVES OF GENERAL PSYCHIATRY, VOL 59, P 131

PHOTONONSTOP/SUPERSTOCK; ABOVE LEFT: TOM GRILL/CORBIS

NEW WAYS TO RISE

In the modern world many of us need an alarm clock to rouse us. Left to our own devices, however, we eventually wake up. The signal to do so is generated in a region of the brain called the suprachiasmatic nucleus (SCN) in the hypothalamus. This plays a central role in sleep-wake timing by regulating the rhythms of hormones such as melatonin, as well as neurotransmitters and neuromodulators such as orexins.

Because the internal wake-up call is gentle, it is easy to oversleep. Furthermore, the call often arrives a bit late because most of us have a body clock that naturally runs for longer than 24 hours. Exposure to light from phones, TVs and computers in the evening may also shift the clock to a later hour, making matters worse.

In the absence of an alarm clock, we most often wake up during REM sleep, which some sleep scientists have called the "gate to wakefulness". Large parts of the brain are active during REM sleep and switching to fully alert wakefulness may therefore be easier. But during deep sleep many brain areas are deactivated. If woken from this state by an alarm clock (or at the end of a long afternoon nap) we feel groggy.

There are new kinds of alarm clock that attempt to ensure a more pleasant awakening. One company offers a way to monitor sleep states to ensure the alarm happens during REM sleep, and cellphones and wrist-based monitors claim to detect differences of movement between REM sleep and non-REM sleep, again with the aim of ensuring a more gentle crossing into consciousness. Another approach is to use light: after all, the light-dark cycle is the natural timekeeper and light remains the alarm clock par excellence.

COURTESY OF THE UNIVERSITY OF SURREY

Derk-Jan Dijk

Derk-Jan Dijk is a professor of sleep and physiology at the University of Surrey in Guildford, UK, and director of the Surrey Sleep Research Centre.

Raphaëlle Winsky-Sommerer

Raphaëlle Winsky-Sommerer is reader in sleep and circadian rhythms at the University of Surrey in Guildford, UK.

MOVING TO A 24/7 SOCIETY

Sleep has always been an important part of human society and will inevitably remain so, not least because of its pleasures and benefits. Even so, we may request more control over how and when we sleep.

Artificial light and caffeine have already handed us that control to some extent by freeing us from the dictates of sunrise and sunset. A next step towards a culture of on-demand rest and work may be more powerful sleep and alertness compounds. These could give us effective wakefulness and quality sleep at any time of day. In the ideal scenario this would enable a truly 24/7 global society without jet lag and the negative health consequences of shift work.

There are still many issues to deal with before we arrive at this ideal state of slumber. For one thing, our understanding of sleep-wake regulation and pharmacological tools to manipulate it remains incomplete.

We may dream of a future of satisfying sleep and productive wakefulness but even light bulbs and caffeine are likely to cause us problems. For example, artificial evening light and coffee reduce sleepiness, which drives us to stay up later even though the next working day starts at the same time. This leads to a cycle of sleep deprivation and more coffee or energy drinks to combat it.

It is only relatively recently that epidemiological studies of sleep duration and health have confirmed that we need adequate sleep for good health.

The growing scientific interest in sleep, its circadian organisation and recognition of its importance to health will transform how we approach it in the future.

RECOMMENDED READING

Sleep in Animals: A state of adaptive inactivity by Jerry Siegel, in *Principles and Practice of Sleep Medicine* edited by Meir Krieger, Thomas Roth and William Dement (Saunders, 2010)

"Sleep as a fundamental property of neuronal assemblies" by James Krueger, David Rector, Sandip Roy, Hans Van Dongen, Gregory Belenky and Jaak Panksepp (*Nature Reviews Neuroscience*, vol 9, p 910)

"Is sleep essential?" by Chiara Cirelli and Giulio Tononi (*PloS Biology*, vol 6, p e216)

Cover image:
Flore-Aël Surun/Tendance Floue

CHAPTER FOUR

GLOBAL HAZARDS

EARTHQUAKES

SUSAN HOUGH

INSTANT EXPERT

HISTORY & MEASUREMENT

What is an earthquake? We have always been aware of the planet's rumblings, but it has taken us centuries to grasp the true causes. And as for sizing them up, seismologists only settled on a reliable scale of measurement in the 1930s

WHAT AND WHERE

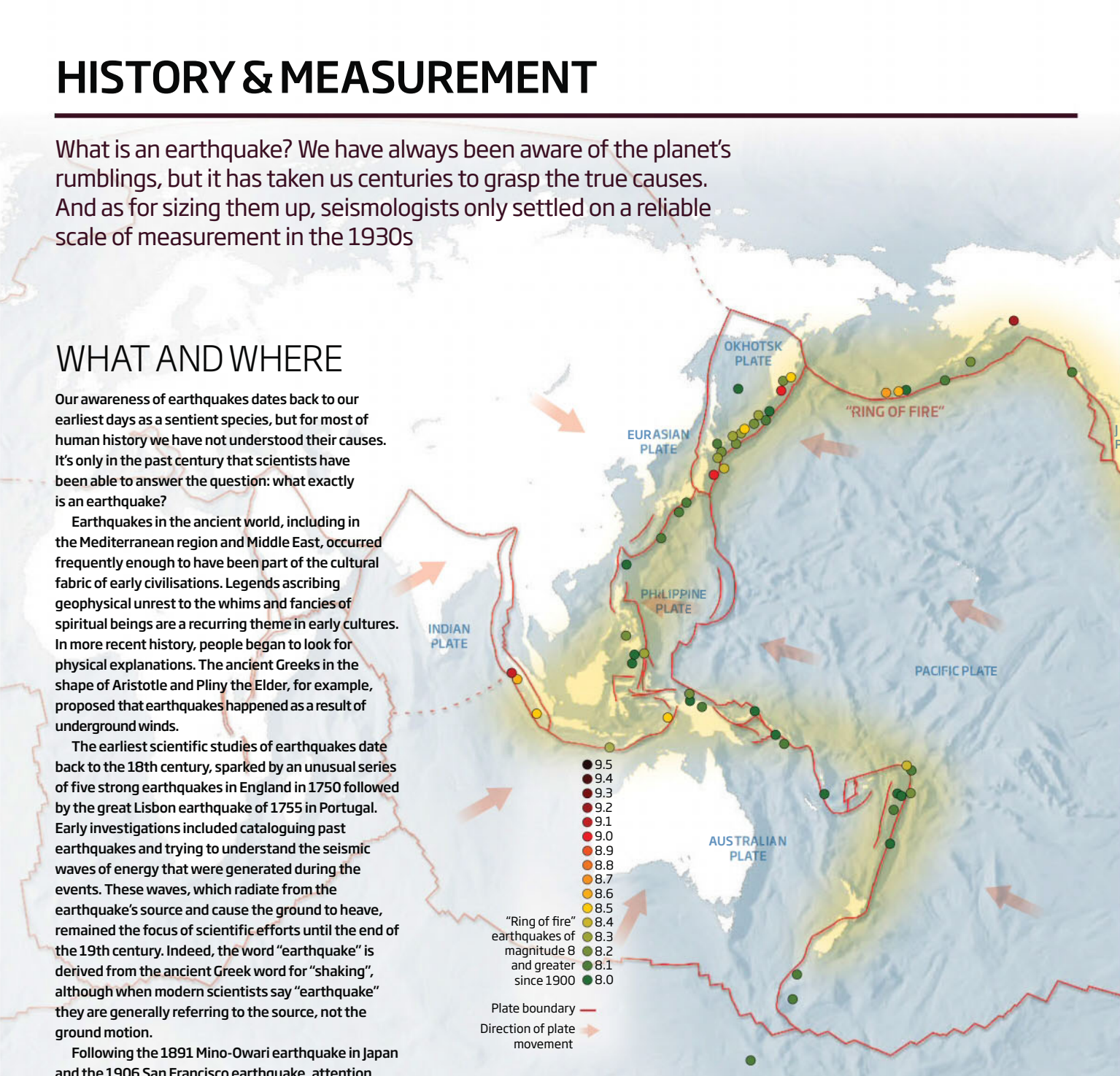
Our awareness of earthquakes dates back to our earliest days as a sentient species, but for most of human history we have not understood their causes. It's only in the past century that scientists have been able to answer the question: what exactly is an earthquake?

Earthquakes in the ancient world, including in the Mediterranean region and Middle East, occurred frequently enough to have been part of the cultural fabric of early civilisations. Legends ascribing geophysical unrest to the whims and fancies of spiritual beings are a recurring theme in early cultures. In more recent history, people began to look for physical explanations. The ancient Greeks in the shape of Aristotle and Pliny the Elder, for example, proposed that earthquakes happened as a result of underground winds.

The earliest scientific studies of earthquakes date back to the 18th century, sparked by an unusual series of five strong earthquakes in England in 1750 followed by the great Lisbon earthquake of 1755 in Portugal. Early investigations included cataloguing past earthquakes and trying to understand the seismic waves of energy that were generated during the events. These waves, which radiate from the earthquake's source and cause the ground to heave, remained the focus of scientific efforts until the end of the 19th century. Indeed, the word "earthquake" is derived from the ancient Greek word for "shaking", although when modern scientists say "earthquake" they are generally referring to the source, not the ground motion.

Following the 1891 Mino-Owari earthquake in Japan and the 1906 San Francisco earthquake, attention shifted to the mechanisms that give rise to these events. Using data from triangulation surveys – an early forerunner to GPS – conducted before and after the 1906 earthquake, geophysicist Harry Fielding Reid developed one of the basic tenets of earthquake science, the theory of "elastic rebound". This describes how earthquakes occur due to the abrupt release of stored stress along a fault line (see diagram, right).

Another half-century elapsed before the plate tectonics revolution of the mid-20th century provided an explanation for the more fundamental question: what drives earthquakes? We now know that most earthquakes are caused by the build-up of stress along the planet's active plate boundaries, where tectonic plates converge or slide past each other.



Other earthquake causes have also been identified, such as post-glacial rebound, when the crust returns to its non-depressed state over timescales of tens of thousands of years following the retreat of large ice sheets. Such processes, however, make up only a tiny percentage of the overall energy released by earthquakes due to plate tectonics.

Thus has modern science established the basic framework to understand where, how and why earthquakes happen. The devil continues to lurk in the details.

JIM MCHUGH/SYGMA/CORBIS; BEARIMAGES/UNIVERSITY OF CALIFORNIA

SIZING UP

How do we measure earthquakes? By the early 20th century, geologists knew that some earthquakes create visible rips across Earth's surface, which gives some indication of their force. But since most fault ruptures are entirely underground, we need other methods to size up and compare earthquakes.

The earliest scales were called intensity scales, which typically assign Roman numerals to the severity of shaking at a given location. Intensity scales remain in use today: well-calibrated intensity values determined from accounts of earthquake effects help us study historical earthquakes and their effects within densely populated areas, for example. Following an earthquake in Virginia in 2011, over 140,000 people reported their accounts to the US Geological Survey's "Did You Feel It?" website.

To size up an earthquake directly, one needs to record and dissect the waves it generates. Today, this is done with seismometers employing digital recording, but it wasn't always so. The first compact instrument capable of faithfully recording small local earthquakes was called a Wood-Anderson seismometer. When the ground shook, a mass suspended on a tense wire would rotate, directing a light onto photosensitive film. The image "drawn" by the light reflected the severity of the seismic waves passing through.

In the early 1930s, Charles Francis Richter used these seismometers to develop the first magnitude scale – borrowing the word "magnitude" from astronomy. Richter's scale uses a logarithm to produce magnitude values that are easily tractable: each one unit increase in magnitude corresponds to a 30-fold increase in energy release. A magnitude-7 earthquake thus releases almost 1000 times more energy than a magnitude 5 earthquake.

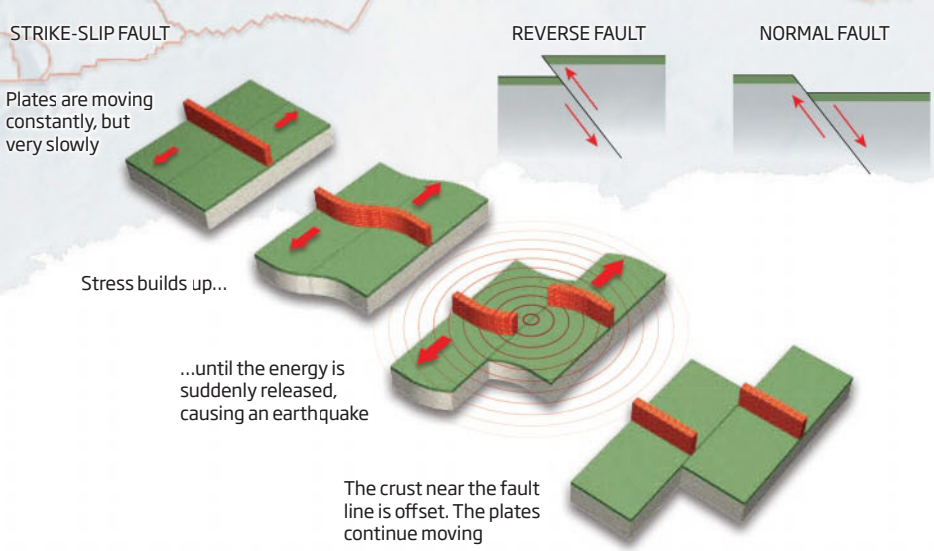
Magnitude values are relative: no physical units are attached. Richter tuned the scale so that magnitude 0 (M0) was the smallest earthquake that he estimated could be recorded by a surface seismometer under ordinary conditions. Earthquakes with negative magnitudes are possible but thus unlikely to be recorded. The scale is open-ended, but Richter might have had an upper limit of M10 in mind: he also tuned the scale so that the largest recorded earthquakes in California and Nevada were around M7, and surmised that the 1906 San Francisco earthquake was probably around M8. (The largest earthquake recorded since then was in Chile in 1960, with an estimated magnitude of 9.5.)

Relationships have been developed since to relate the energy released by earthquakes to magnitude. In the 1960s, Keiitti Aki introduced a fundamentally different quantity: the "seismic moment". This provides a full characterisation of the overall size of an earthquake and is the measure generally used in scientific analyses.

The so-called moment-magnitude scale was introduced to convert the seismic moment to an equivalent Richter magnitude. This figure is the one usually reported in the media. Strictly speaking this reported value is not "on the Richter scale", because it is calculated differently to Richter's formulation. Still, following Richter's approach, moment-magnitude values have no physical units, and are useful for comparing earthquakes.

Charles Richter (left) borrowed the term "magnitude" from astronomy.

The "elastic rebound" theory describes how earthquakes occur at faults due to the movement of plates



GROUND MOTION

Understanding the shaking caused by earthquakes is crucial if we areto prepare for these events – but the impact of an earthquake on people and cities depends on more than magnitude alone. Earth's crust can amplify or dampen the severity of shaking

SHAKE, RATTLE AND ROLL

Seismic waves cause perceptible ground motion if they are strong enough. For seismic hazard assessment, the study of ground motion is where the rubber meets the road. If we understand the shaking, we can design structures and infrastructures to withstand it.

The severity of earthquake shaking is fundamentally controlled by three factors: earthquake magnitude, the attenuation of energy as waves move through the crust and the modification of shaking due to the local geological structure.

Bigger earthquakes generally create stronger shaking, but not all earthquakes of a given magnitude are created equal. Shaking can depend significantly on factors such as the depth of the earthquake, the orientation of a fault, whether or not the fault break reaches the surface and whether the earthquake rupture is relatively faster or slower than average.

Attenuation of seismic waves varies considerably in different regions. In a place like California or Turkey, where the crust is highly fractured and relatively hot, waves dissipate – or attenuate – quickly. Following the 1906 San Francisco earthquake, pioneering geologist G. K. Gilbert observed: "At a distance of twenty miles [from the fault] only an occasional chimney was overturned… and not all sleepers were wakened." In regions that are far from active plate boundaries, such as peninsular India or the central and eastern US, waves travel far more efficiently. The three principal mainshocks of the 1811-1812 New Madrid earthquake sequence in the central US damaged chimneys and woke most sleepers in Louisville, Kentucky, some 400 kilometres away. In 2011, the magnitude-5.8 Virginia earthquake was felt in Wisconsin and Minnesota, over 1500 km away.

Local geological structures such as soft sediment layers can amplify wave amplitudes. For example, the M8 earthquake along the west coast of Mexico in 1985 generated a ringing resonance in the lake-bed sediments that underlie Mexico City. And in Port-au-Prince, some of the most dramatic damage in the 2010 Haiti earthquake was associated with amplification by small-scale topographic features such as hills and ridges.

Characterisation of the full range and nature of site response remains a prime target for ground motion studies, in part because of the potential to map out the variability of hazard throughout an urban region, called "microzonation". This offers the opportunity to identify those parts of urban areas that are relatively more and less hazardous, which can guide land-use planning and appropriate building codes. Rubber, meet road.

MUSTAFA OZER/AFP/GETTY; BACKGROUND IMAGE: SIPA PRESS/REX FEATURES

TKYODO/REUTERS

TSUNAMI!

The tsunami that hit Japan in 2011 caused more damage and deaths than the shaking

Undersea earthquakes can generate a potentially lethal cascade: a fault break can cause movement of the seafloor, which displaces the water above to form a tsunami wave.

Tsunamis can also be generated when earthquakes trigger undersea slumping of sediments, although these waves are generally more modest in size.

Tsunami waves spread out through the ocean in all directions, travelling in the open ocean about as fast as a jet plane. They have a very long wavelength and low amplitude at sea, but grow to enormous heights as the wave energy piles up against the shore.

"Earthquakes far from major plate boundaries can often be felt over 1000 kilometres away"

STRONGEST LINKS

Earthquakes are often related to one another – one can lead to another – but there are common misconceptions about what drives them and the ways that they are linked.

It is an enduring misperception that a large earthquake is associated with a sudden lurching of an entire tectonic plate. If one corner of the Pacific plate moves, shouldn't it be the case that other parts of the plate will follow suit? The idea might be intuitive, but it is wrong. The Earth's tectonic plates are always moving, typically about as fast as human fingernails grow. What actually happens is that adjacent plates lock up, causing warping of the crust and storing energy, but only over a narrow zone along the boundary. So when an earthquake happens, this kink is catching up with the rest of the plate.

Earthquake statistics do tell us, however, that the risk of aftershocks can be substantial: on average, the largest aftershock will be about one magnitude unit smaller than the mainshock. Aftershocks cluster around the fault break, but can also occur on close neighbouring faults. As the citizens of Christchurch, New Zealand, learned in 2011, a typical largest aftershock (M6.1) had far worse consequences than the significantly bigger mainshock (M7), because the aftershock occurred closer to a population centre.

In addition to aftershock hazard, there is always a chance that a big earthquake can beget another big earthquake nearby, typically within tens of kilometres, on a timescale of minutes to decades. For example, the 23 April 1992 M6.1 Joshua Tree earthquake in southern California was followed by the 28 June 1992 M7.3 Landers earthquake, approximately 35 kilometres to the north. Such triggering is understood as a consequence of the stress changes caused by the movements of the rocks. Basically, motion on one fault will mechanically nudge adjacent faults, which can push them over the edge, so to speak, following delays ranging from seconds to years.

An additional mechanism is now recognised as giving rise to triggering: the stress changes associated with seismic waves. Remote triggering occurs commonly – but not exclusively – in active volcanic and geothermal areas, where underground magmatic fluid systems can be disrupted by passing seismic waves.

Overwhelmingly, remotely triggered earthquakes are expected to be small. Here again, recent advances in earthquake science as well as centuries of experience tell us that earthquakes do not occur in great apocalyptic cascades. However, in recent decades scientists have learned that faults and earthquakes communicate with one another in far more diverse and interesting ways than the classic foreshock-mainshock-aftershock taxonomy suggests.

PREDICTION

Many avenues for earthquake forecasting have been explored, from prior changes in animal behaviour to electromagnetic signals. Yet predicting exactly when an earthquake will happen remains impossible today. Still, there is a great deal we do know about Earth's shaking in the future

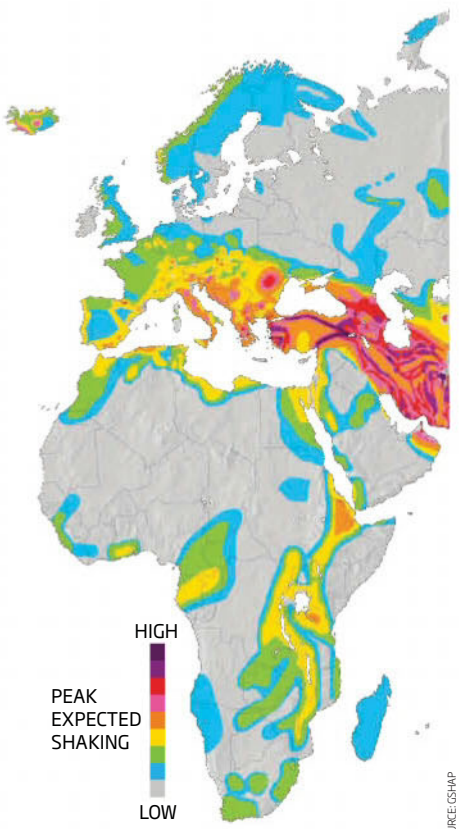
FORECASTING: WHAT WE KNOW

When seismologists are asked whether earthquakes can be predicted, they tend to be quick to answer no. Sometimes even we geologists can forget that, in the ways that matter, earthquakes are too predictable. We know where in the world they are likely to happen. For most of these zones, we have quite good estimates of the expected long-term rates of earthquakes (see map, right). And while we often cannot say that the next Big One will strike in a human lifetime, we can say it is very likely to occur within the lifetime of a building.

We know the largest earthquakes occur along subduction zones, where a tectonic plate dives beneath another into Earth's mantle, with rupture lengths of more than 1000 kilometres and an average slip along a fault of tens of metres. But any active plate boundary is fair game for a big earthquake, at any time. For example, two years before the 2010 earthquake in Haiti, geophysicist Eric Calais and his colleagues published results of GPS data from the region, noting that "the Enriquillo fault is capable of a M7.2 earthquake if the entire elastic strain accumulated since the last major earthquake was released in a single event". While this exact scenario did not play out in 2010, it wasn't far off. We can say for sure that people living on plate boundaries will always face risk.

Future large earthquakes are expected in California. Research by James Lienkaemper and his colleagues estimates that sufficient strain is stored on the Hayward fault in the east San Francisco Bay area to produce a M7 earthquake. An earthquake this size is expected, on average, every 150 years. The last one was in 1868. Local anxieties inevitably mount knowing such information, but earthquakes occur by irregular clockwork: if the average repeat time is 150 years, it could vary between 80 to 220 years. So we are left with the same vexing uncertainty: an "overdue" earthquake might not occur for another 50 years, or it could happen tomorrow. On a geological timescale there is not much difference between sooner versus later. On a human timescale, however, sooner versus later seems like all the difference in the world.

Earth scientists have made great strides in forecasting the expected average rates of damaging earthquakes. The far more challenging problem remains finding the political will and resources to prepare for the inevitable.



SOURCE: GSHAP

Geologists use hazard maps to illustrate earthquake risk in a region. This one shows the peak shaking that policymakers should prepare for in the next 50 years

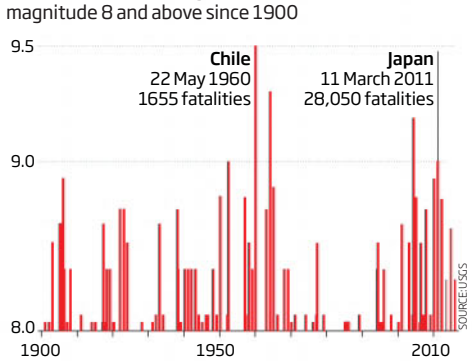
COLORADO STATE UNIVERSITY/NATIONAL SCIENCE FOUNDATION

MEGAQUAKE MYTHS

Since the M9.1 Sumatra-Andaman earthquake struck on Boxing Day in 2004, another five earthquakes with magnitudes of 8.5 or greater have occurred on the planet, including the Tohoku, Japan, earthquake in 2011 (see diagram, below). This apparent spate has led some to wonder if earthquake frequency is increasing. Careful statistical analysis reveals that it is not.

The recent rate of very large earthquakes is unusual, but not a statistically significant increase relative to expected variability. And the overall energy release by earthquakes in the past eight

Earthquakes measuring magnitude 8 and above since 1900



WHY SO DIFFICULT?

"Shake tables" test how buildings will act in an earthquake

In the 1970s and 1980s, leading scientists were quoted in the media expressing optimism that reliable short-term prediction of earthquakes was around the corner. This was fuelled by promising results from the Soviet Union, and the apparently successful prediction of the 1975 earthquake in Haicheng, China. Since then, this optimism has given way to varying degrees of pessimism. Why are earthquakes so hard to predict?

Any number of possible precursors to earthquakes have been explored: small earthquake patterns, electromagnetic signals and radon or hydrogeochemical changes. Many of these seemed promising, but none have stood up to rigorous examination.

Consider this example. In March 2009, Italian laboratory technician Giampaolo Giuliani made a public prediction that a large earthquake would occur in the Abruzzo region of central Italy. His evidence? An observed radon anomaly. The prediction was denounced by local seismologists. The M6.3 L'Aquila earthquake struck the area on 6 April, killing 308 people.

This gets to the issue of reliable precursors. It is possible that radon was released because of the series of small earthquakes, or foreshocks, that preceded the main earthquake. It is also possible it was coincidence. Scientists explored radon as a precursor in the 1970s and quickly discovered how unreliable it is. Once in a while radon fluctuations might be associated with an impending earthquake, but usually they are not. Meanwhile big earthquakes hit regions where radon anomalies did not take place. The same story has played out with many other proposed precursors.

That's not to say that seismologists have neglected to investigate precursors – on the contrary they are examining them with increasingly sophisticated methods and data. However, a common bugaboo of prediction research is the difficulty of truly prospective testing. To develop a prediction method based on a particular precursor, researchers compare past earthquakes with available recorded data. One might, for example, identify an apparent pattern of small earthquakes that preceded the last 10 large earthquakes in a given region. Such retrospective analyses are plagued by subtle data selection biases. That is, given the known time of a big earthquake, one can often look back and pick out apparently significant signals or patterns.

This effect is illustrated by the enduring myth that animals can sense impending earthquakes. It is possible that animals respond to weak initial shaking that humans miss, but any pet owner knows that animals behave unusually all the time – and it's soon forgotten. People only ascribe significance with hindsight.

At present most seismologists are pessimistic that prediction will ever be possible. But the jury is still out. One of the big unanswered questions in seismology is: what happens in the earth to set an earthquake in motion? It is possible that some sort of slow nucleation process is involved, and therefore possible that earthquake precursors exist.

For this as well as all earthquake prediction research, the challenge is to move beyond the retrospective and the anecdotal, into the realm of statistically rigorous science.

California schoolchildren perform earthquake practice drills

JUSTIN SULLIVAN/GETTY

years is still below the combined energy release of the two largest recorded earthquakes: the 1960 Chilean quake and Alaska's quake on Good Friday 1964.

Anthropogenic climate change could conceivably influence earthquake rates in some areas: the post-glacial rebound associated with the retreat of glaciers provides a source of stress that can drive earthquakes. Such earthquakes could have a significant local impact, but their overall energy release will continue to be dwarfed by that of earthquakes caused by plate tectonics.

While there is no reason to believe that megaquakes are on the rise, there is little doubt that more and worse megadisasters due to earthquakes lie ahead in our future – they are the inevitable consequence of explosive population growth and concomitant construction of vulnerable dwellings in the developing world.

Susan Hough

Susan Hough is a seismologist at the US Geological Survey in Pasadena, California, and a Fellow of the American Geophysical Union. She led the Earthquake Disaster Assistance Team effort to deploy seismometers in Haiti following the January 2010 earthquake.

WHITHER EARTHQUAKE SCIENCE

In the 1970s, during the heyday of earthquake prediction research, Charles Richter remained an ardent and vocal sceptic, a stance that drove a wedge between him and more optimistic colleagues. Overwhelmingly, the lessons of subsequent decades have vindicated Richter's views. Yet asked in 1979 if he thought earthquake prediction would ever be possible, he replied: "Nothing is less predictable than the development of an active scientific field."

Indeed, the 25 years since Richter's death have witnessed developments he could not have imagined, including the recent recognition that many subduction zones generate a kind of seismic chatter, dubbed non-volcanic tremor, and that patches along subduction zones can slip slowly without releasing seismic waves. Non-volcanic tremor, which is thought to occur along the deep extension of faults into layers that are too hot to remain fully brittle, has also been identified along a few faults outside subduction zones. Could the processes at play in the deeper layers be the key to understanding the occurrence of large earthquakes?

Other intriguing but controversial ideas have been proposed, including the theory that electromagnetic precursors are generated before faults rupture. Scientific discourse on such research is couched within polarised debates about proposed prediction methods. Some scientists now wonder if the pessimism about the feasibility of reliable earthquake prediction has led the field to shy away from investigations that could help us understand earthquake processes.

Some fairly basic questions still beg for answers: why does an earthquake start at a particular time and place? Why does an earthquake stop? As a big earthquake starts, does it "know" it will be a big earthquake? Or is it merely a small one that gets out of hand?

As seismologists work to develop a more complete understanding of earthquakes, and to refine hazard assessments, one sobering lesson has emerged: expect the unexpected. While hazard maps characterise the expected long-term rates of earthquakes in many regions, an "overdue" earthquake might not strike for another 100 years.

Moreover, even in well studied areas, the historical record is too short to understand fully the variability of the earthquake cycle associated with a given plate boundary. Geological investigations of prehistoric earthquakes can start to extend our knowledge to more geologically meaningful timescales, but such results are limited and typically characterised by high uncertainties. Our understanding of both the variability of earthquake repeat times and the largest possible earthquake in a given area is limited at best. Our expectations for the largest possible earthquakes are often too strongly shaped by the events in the historical record only. We should know better.

FURTHER READING

Predicting the Unpredictable: The tumultuous science of earthquake prediction by Susan Hough, Princeton University Press, 2009

Earthshaking Science: What we know (and don't know) about earthquakes by Susan Hough, Princeton University Press, 2002

Introduction to Seismology by Peter Shearer, Cambridge University Press, 1999

Earthquakes 5th edition by Bruce Bolt, W. H. Freeman, 2003

WEBSITES

US Geological Survey Earthquake Hazards Program
earthquake.usgs.gov

European-Mediterranean Seismological Centre
emsc-csem.org

Global Seismic Hazard Assessment Program
www.seismo.ethz.ch/static/GSHAP

The Great ShakeOut US earthquake drills
shakeout.org

Cover image: Osman Orsal/Reuters

EXTREME WEATHER

JEFF MASTERS

INSTANT
EXPERT

DRIVING FORCE

The thin layer of gas making up the Earth's amazing atmosphere is prone to moods of spectacular beauty – and stunning violence. To understand what makes the weather go wild, we must start by looking at the intricate workings of the forces that set the atmosphere in motion and thus drive our planet's weather

THE SUN: DRIVER OF THE WEATHER

Earth's atmosphere is heated unequally at the poles and equator. This occurs because of simple geometry. We live on a sphere orbiting the sun, and sunlight falls from directly overhead on the equator, but at a sharply slanted angle near the north and south poles. The polar regions thus receive less sunlight for a given area than the equator. This difference is the fundamental driver of weather on the planet. Heat naturally moves from hotter to colder areas, so the atmosphere and oceans transport heat from the equator to the poles. A planet without temperature differences would be a planet without weather, where the wind never blows. But on Earth the wind always blows, and sometimes it blows very hard.

TEMPERATURE, PRESSURE AND WINDS

Air is warmed in three main ways: radiation, conduction and convection. The sun radiates photons that are absorbed by air molecules, making the molecules move faster – that is, get warmer. It has the same effect on the ground, whose molecules then conduct heat energy to the thin layer of air in contact with it.

As the molecules in this parcel of warm air zing around more rapidly, its volume increases. Since this makes the air parcel less dense than the surrounding air, it becomes more buoyant and thus rises. Cooler, heavier air flows into the space it has vacated, where it in turn becomes heated and rises, continuing the cycle. This vertical movement of heat is called convection, and the rising parcels of air are known as thermals.

In this way, temperature differences cause variations in density and pressure that drive winds both vertically and horizontally as the air flows to try to equalise the pressure.

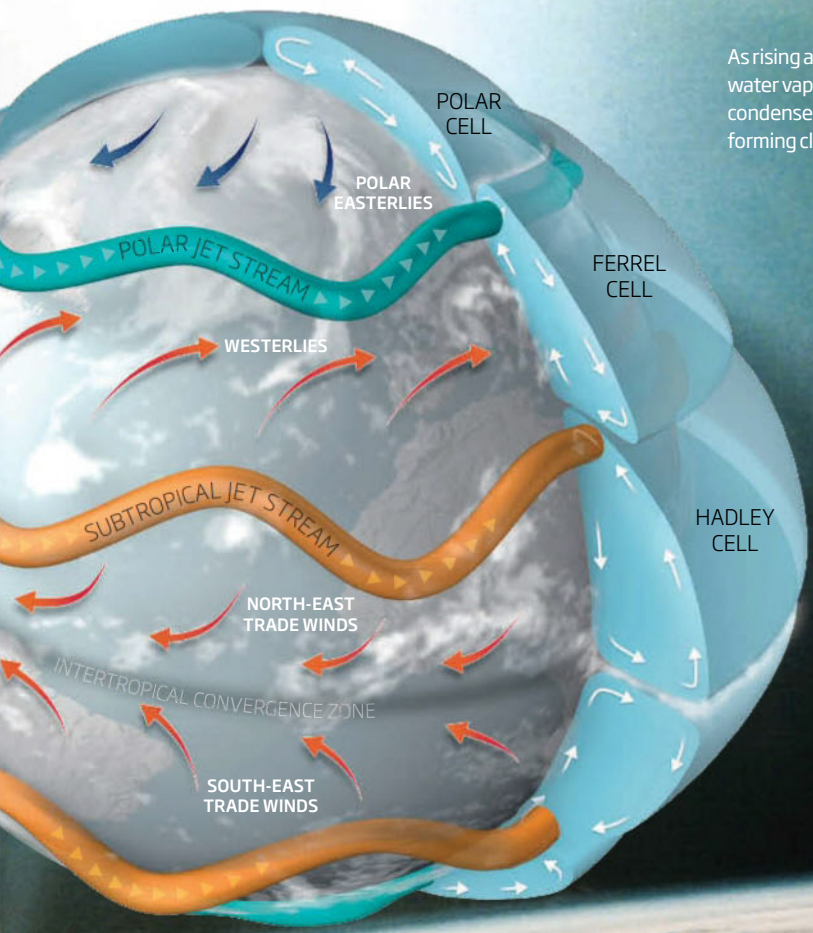
"If Earth spun faster, there would be more of these wind belts, as can be seen on Jupiter"

XINHUA/ANATOLIA NEWS AGENCY/EYEVINE

Rising currents called thermals form as the sun heats air near the ground

The speed at which the Earth rotates shapes its wind patterns

As rising air cools, water vapour condenses out, forming clouds



WIND BELTS AND THE CORIOLIS EFFECT

If Earth did not rotate, global wind patterns would be very simple. Hot air would rise at the equator, then spread out horizontally towards the poles once it reached the top of the atmosphere. At the poles it would cool, sinking as it became more dense, then flow along the surface back to the equator. Surface winds would thus flow only from north to south in the northern hemisphere, and south to north in the southern hemisphere.

On a rotating sphere, the surface – and the air above it – moves fastest at the equator and not at all at the poles. Thus, Earth's rotation deflects winds to the right in the northern hemisphere, and left in the southern. This deflection is called the Coriolis effect.

The rotation of the Earth creates a Coriolis effect strong enough to produce three interlocking bands of surface winds in each hemisphere: the equatorial trade winds, the mid-latitude westerlies and the polar easterlies (see diagram above). If Earth spun faster, there would be more of these wind belts. Jupiter's very fast rotation rate – its days last just 10 hours – gives that planet many more bands of winds than Earth. At high altitudes, fast west-to-east bands of wind called jet streams develop above the slower-moving surface winds. While this general pattern of wind belts predominates, because we do not live on a uniform sphere but on one with oceans, mountains, forests and deserts, actual wind patterns are far more complicated and variable.

INSTABILITY AND PRECIPITATION

Instability occurs when air is less dense than its surroundings. It is greatest when there is cold, dense air aloft, and warm, moist air at low levels, since moist air is less dense than dry air. In an unstable atmosphere, air that starts moving upwards will keep on moving upwards.

As air moves upwards, it expands further. Expansion causes cooling, since as the molecules of air move farther apart, the amount of kinetic energy in a given volume falls. At some point it becomes too cool for water vapour in the rising air to remain in the gas phase. When air reaches this point, called the dew-point temperature, water vapour begins condensing out of the air, forming clouds and precipitation – rain, hail or snow. Thus, two ingredients are needed to generate precipitation: sufficient water vapour in the air, and a mechanism to lift air so that it cools to its dew-point temperature.

The three main ways air gets lifted to cause cooling and condensation are:

1) solar heating of the ground, causing thermals to develop

2) air masses of different densities meeting and creating "fronts" that push air upwards, and

3) air being forced upwards by mountains in its way.

While air temperatures fall with altitude in the lower atmosphere, or troposphere, at an altitude of around 11 kilometres the air suddenly begins to warm again. This "temperature inversion" marks the bottom of the stratosphere. It is caused by the heating of air as ozone absorbs ultraviolet light. No clouds form in the stratosphere, since air from the troposphere cannot rise above the inversion. This puts a lid on instability. If there was no inversion, we would get more extreme weather.

BACKGROUND: NASA; TOP: GUNTER CIVICA

CONCENTRATED FURY

Tornadoes reveal the atmosphere at its most violent. Spawned by thunderstorms, they produce the fastest winds of any natural phenomenon. It is fortunate for us that most tornadoes are small, and that the seriously destructive ones are extremely rare

THUNDERSTORMS

If the sun's heat is strong enough, the upward-moving thermals it creates form puffy-topped cumulus clouds. In some cases, the tops of these cauliflower-shaped clouds may reach the top of the troposphere. In the upper parts of these clouds, freezing temperatures create ice and snow, and collisions between the frozen particles separate electric charge. When the charge builds up to a critical level, a lightning bolt strikes, reuniting the positive and negative charges.

The cause of the subsequent thunderclap is still being debated. One recent theory is that it is driven by energy released after N_2 and O_2 molecules have been split apart – an air explosion.

Whatever the reason, the cumulus cloud is now a cumulonimbus cloud – a thunderstorm. As well as providing life-sustaining rains for most of the planet, thunderstorms also bring a variety of hazards. The world's heaviest rainfall events are invariably caused by thunderstorms; the heavy rainfall of tropical cyclones is due to the thunderstorms embedded within them. Severe thunderstorms can generate destructive straight-line winds with speeds up to 240 kilometres per hour and generate hailstones a large as a grapefruit. Finally, thunderstorms spawn nature's most violent windstorm – the tornado.

Hailstones can do serious damage to people, plants and planes even when they are far smaller than this

RIGHT AND BACKGROUND: JIM REED/CORBIS; ABOVE: JIM REED/SPL

MID-LATITUDE CYCLONES

The unequal heating of the equator and poles often leads to storms thousands of kilometres across, which transport heat towards the poles. These huge storms, called mid-latitude cyclones, are the familiar low-pressure systems and winter blizzards that give the mid-latitudes much of their precipitation.

Mid-latitude cyclones form where sharp temperature contrasts exist along a front separating cold, dry polar air from warm, moist tropical air. These great storms are primarily powered by the release of potential energy as cold, dense air moving down and towards the equator displaces warmer, less dense air moving polewards and upwards.

An additional energy source is latent heat. A lot of energy is needed to turn liquid water into vapour, and this energy is released when the vapour condenses. When air is lifted and cooled in a storm, and water vapour condenses, the release of latent heat warms the surrounding atmosphere. That makes this air rise higher, which releases yet more latent heat, powering the storm. The storm acts as a heat engine, converting the heat energy into kinetic energy – wind.

HECTOR MATA/AFP/GETTY

TORNADOES

During a tornado in Bridge Creek, Oklahoma, on 3 May 1999, Doppler radar revealed a wind speed of 486 kilometres per hour about 30 metres above the ground – the fastest ever recorded. Winds of this strength cause total destruction, sweeping strong timber-frame houses off their foundations and badly damaging steel-reinforced concrete structures.

In the past decade there have been 15 top-end tornadoes earning an EF-5 designation on the Enhanced Fujita scale (winds exceeding 322 kilometres per hour). Unlike hurricanes, tornadoes are quite small, ranging from 75 metres across to about 3 kilometres. They descend from cumulonimbus (thunderstorm) clouds, which can be over land or water. Those that form or move over water are called waterspouts and tend to be much weaker than tornadoes that develop over land.

A very particular set of conditions is needed for tornadoes to form. Most important is the presence of instability and wind shear. A low-altitude flow of warm, moist air from an ocean area combined with a flow of cold, dry polar air high up creates the conditions for maximum instability, which means that parcels of air heated near the surface rise rapidly, creating powerful updrafts.

If a strong jet stream is present, with high winds near the top of the troposphere, there will be vertical wind speed shear. If the winds also change from southerly near the surface to westerly aloft, there is vertical wind direction shear. These two types of shear make the updraft rotate, creating a rotating thunderstorm, or supercell. Supercells spawn the vast majority of strong (EF-2 and EF-3) and violent (EF-4 and EF-5) tornadoes.

A third ingredient that is usually needed to generate supercell thunderstorms is the "cap". This is a region in the middle layers of the atmosphere where dry, stable air has intruded. It prevents air rising very high until later in the day, when solar heating eventually generates enough instability for one thermal to burst through the cap. The result is a single, large supercell instead of a number of smaller, spread-out thunderstorms.

These conditions are most common in the Midwestern US. The Gulf of Mexico provides a source of warm, moist air at low levels, and when this low-density air slides underneath high-density cold, dry air flowing southwards from Canada, an explosively unstable atmosphere often results. Add to this mix a mid-level intrusion of dry, stable air from the desert regions to the west and a powerful jet stream aloft creating plenty of wind shear, and dozens or even hundreds of tornadoes can result. During the 25-28 April 2011 tornado outbreak, 355 tornadoes – including four top-end EF-5s – ripped though 21 US states and Canada, killing 324 people.

While the vast majority of the world's tornadoes occur in the US, they do affect other nations too. Bangladesh averages three tornadoes per year, and many of these are strong and violent. The world's deadliest tornado was an EF-5 that hit Bangladesh on 26 April 1989, killing more than 1300 people.

The tornado Super Outbreak of 1974 (right) caused extensive damage across the US and Canada

HECTOR MATA/AFP/GETTY; FRED STEWART/AP

HEAT ENGINES

Warm ocean waters provide the power that drives the world's most fearsome storms. Hurricanes can be more than 2000 kilometres wide, generate storm surges of over 10 metres and deposit more than a metre of rain in a day

TROPICAL CYCLONES

The word "cyclone" can be used to describe any rotating storm system. This includes hurricanes, tornadoes and the ordinary low-pressure systems that develop in the mid-latitudes.

Tropical cyclones, though, form only over warm ocean waters of at least 26 °C, and unlike storm systems over land, derive their energy exclusively from latent heat. Hurricanes, typhoons, tropical storms and tropical depressions are all examples of tropical cyclones.

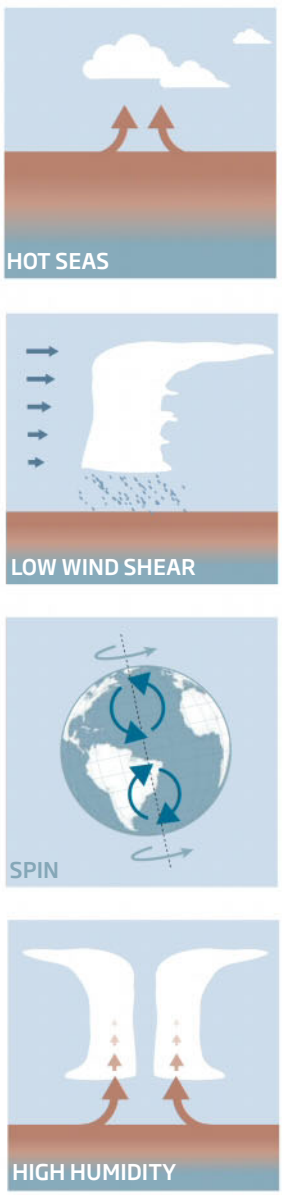
Like tornadoes, tropical cyclones need a particular set of ingredients in order to form, starting with warm ocean water. In addition, vertical wind shear must be very low: in other words, the difference in wind speed between the surface and the top of the troposphere must be less than around 10 metres per second. Any faster and the shear tilts and stretches the core of a developing tropical cyclone, carrying away its heat and moisture.

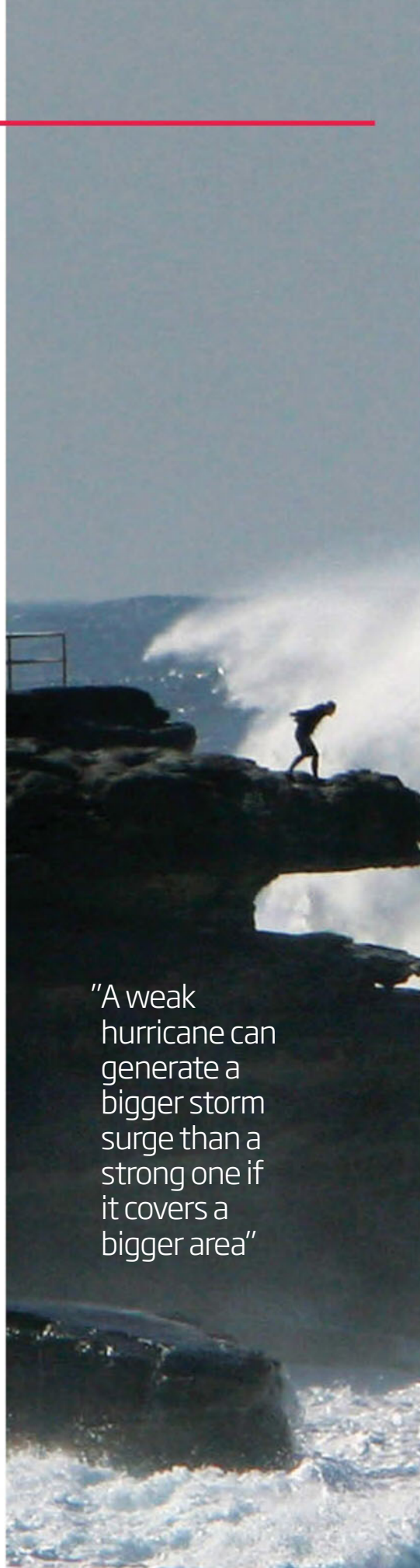
Strong upper-level winds associated with the jet stream or upper-level low-pressure systems are the most common source of wind shear. The more equator-ward position of the subtropical jet stream in winter and spring is the main reason why hurricanes and typhoons rarely form in the Caribbean Sea or western Pacific in those seasons, even though ocean temperatures are warm enough year-round to support such storms. Tropical cyclones also need high moisture through a deep layer of the atmosphere. Dry air from Africa or North America often disrupts a hurricane in the process of forming.

Finally, a tropical cyclone needs something to get it spinning. In the Atlantic this is usually provided by disorganised areas of low pressure called African easterly waves, which emerge from the coast of Africa and move westwards towards the Caribbean.

Hurricanes get more spin from the effect of Earth's rotation. Since the amount of vertical spin of the atmosphere due to the Coriolis force is zero at the equator and maximum at the poles, tropical cyclones generally cannot form within about 5 degrees of latitude from the equator. They then tend to expand as they move polewards, due to the increasing amount of vertical spin.

RECIPE FOR A HURRICANE





"A weak hurricane can generate a bigger storm surge than a strong one if it covers a bigger area"

ESPEN RASMUSSEN/PANOS; FARRIGHT/SANJIT DAS

Hurricanes bring damaging winds, storm surges and heavy rain

TYPHOON OR HURRICANE?

A tropical cyclone starts life as a tropical depression – an organised, spinning storm system with wind speeds of less than 63 kilometres per hour. When the winds grow faster, the system is given a name and classified as a tropical storm. When the winds reach 119 km/h, a ring of intense thunderstorms called the "eyewall" forms around the storm's centre. Within the eyewall is the "eye" of the storm, a clear, calm region of sinking air.

Once the sustained winds exceed 119 km/h, the storm is classified as a hurricane if it is in the Atlantic or eastern Pacific, or as a typhoon in the western Pacific. In the Indian Ocean or in the southern hemisphere it is simply called a cyclone or tropical cyclone. There is no meteorological difference between these differently named storms.

Very warm water extending to a depth of 50 metres or more can help fuel rapid intensification of a tropical cyclone to "major" hurricane status, with winds of 178 km/h or more – the most fearsome and destructive type of storm on the planet.

Traditionally, hurricanes are ranked from 1 to 5 on the Saffir-Simpson scale, based on the maximum sustained wind speed. However, this scale can be misleading. A weak storm that covers a huge area of the sea can generate a larger storm surge than a smaller but more intense hurricane with a higher Saffir-Simpson rating. To give a better idea of the storm surge potential, the experimental Integrated Kinetic Energy scale has been developed. It is a measure of both wind speed and the area over which high winds extend.

RIGHT: NOAA; ISS

The strongest winds in a hurricane occur in the eyewall around the centre

MONSOON DEPRESSIONS

Monsoons operate via the same principle as the familiar summer afternoon sea breeze, but on a grand scale. In summer, the land gets hotter than the sea: that's because on land, the sun's heat is concentrated close to the surface, while at sea, wind and turbulence mix warm water at the surface with cooler water lower down. Also, the molecular properties of water mean it takes more energy to raise its temperature than it does to heat the soil and rock that make up dry land.

As a result, a low-pressure region of rising air develops over land areas. Moisture-laden ocean winds blow towards this region and are drawn upwards when they reach land. The rising air expands and cools, releasing its moisture as some of the heaviest rains on Earth – the monsoon.

Each summer, monsoons affect all continents except Antarctica and are responsible for rains that sustain the lives of billions of people. In India, home to 1.25 billion people, the monsoon provides 80 per cent of the annual rainfall. Monsoons have their dark side too: hundreds of people in India and surrounding nations die every year in floods and landslides triggered by the heavy rains.

The most deadly flooding events usually come from monsoon depressions, also known as monsoon lows. A monsoon depression is similar to, but larger than, a tropical depression. Both are spinning storms hundreds of kilometres in diameter with sustained winds of 50 to 55 kilometres per hour, nearly calm winds at their centre and very heavy rains. Each summer, some seven monsoon depressions form over the Bay of Bengal and track westwards across India. In 2010, two major monsoon depressions crossed India into Pakistan in July and August, bringing heavy rains and causing the most costly floods in Pakistan's history ($10 billion).

NEWSPIX/REX

WEATHER UNDERGROUND

Jeff Masters

Jeff Masters co-founded the Weather Underground online weather information service in 1995 while working on his PhD. He flew with the US National Oceanic and Atmospheric Administration's hurricane hunters from 1986 to 1990.

HURRICANE **HUNTING**

"Another updraft, much stronger, grabs the aircraft. I regret forgetting to fasten my shoulder harness as I struggle to keep from bashing into the computer console. Seconds later, a huge downdraft blasts us, hurling the loosened gear against walls and floor. Gerry and Lowell are barely in control of the aircraft. Hugo is a category 5 hurricane, and we are in the eyewall at 1500 feet! One strong downdraft could send us plunging into the ocean. We have to make it to the eye, where we can climb to a safer altitude.

"We're almost there! Then, disaster. Thick, dark clouds suddenly envelop the aircraft. A titanic fist of wind smashes us. I am thrown into the computer console, bounce off, and for one terrifying instant find myself looking down at a precipitous angle at Sean across the aisle from me."

I served for four years as a flight meteorologist on the National Oceanic and Atmospheric Administration's P-3 hurricane hunter aircraft. During this mission in hurricane Hugo in 1989, we hit the most extreme turbulence any hurricane hunter aircraft has ever survived, with forces of 5.7 *g*. The wings are only rated to stay on at 6 *g*. The pilot lost control of the aircraft and one of our engines burst into flame during the encounter.

Six missions between 1945 and 1974 were lost with all hands. The aircraft used in those days were poorly equipped and wind speed had to be estimated by looking at the degree to which the sea surface had been churned to foam – which meant flying beneath the lowest clouds at an altitude of just 200 metres.

The reason for flying into hurricanes is to measure the strongest winds, which occur in the eyewall – the ring of violent thunderstorms that surrounds the calm eye. Knowing the exact strength of the eyewall winds is crucial for issuing proper warnings.

Satellites cannot measure winds in the eyewall directly, as they have a limited ability to see through clouds and rain. That an airplane could safely penetrate the eyewall and survive was first demonstrated on 27 July 1943, by Colonel Joseph Duckworth. This dangerous task will likely be taken over by UAVs, with no crew on board, in 20 or so years from now. However, only crewed aircraft can carry the heavy Doppler radar instruments needed to fully probe the structure of a hurricane, so these aircraft will continue to fly into the less dangerous parts of hurricanes for a long time to come.

RECOMMENDED READING

The AMS Weather Book: The Ultimate guide to America's weather by Jack Williams (University of Chicago Press)

Meteorology Today by Donald Ahrens (Brooks Cole)

Extreme Weather by Christopher Burt (W. W. Norton)

Divine Wind by Kerry Emanuel (Oxford University Press)

Tornado Alley by Howard Bluestein (Oxford University Press)

The Rough Guide to Climate Change by Robert Henson (Rough Guides)

Cover image:
Jim Reed/Science Photo Library

MASS EXTINCTIONS

MICHAEL J. BENTON

INSTANT EXPERT

DEATH ON A MASSIVE SCALE

Every now and again, life on Earth faces a crisis. At least five times in the past 540 million years half or more of all species have been wiped out in a short space of time. These mass extinctions are important punctuation marks in the history of life, as once-dominant groups are swept away and replaced with new ones. What triggers this wholesale regime change? How does life recover? And are we in the middle of a mass extinction of our own making?

Woolly rhinos and mammoths died out in an extinction event 11,000 years ago

BOTH: ITV/REX FEATURES; BACKGROUND: PASCAL BLONDE/GETTY

WHAT IS A MASS EXTINCTION?

Extinction is a normal part of evolution. Species come and go continually – around 99.9 per cent of all those that have ever existed are now extinct. The cause is usually local. For example, a lake might dry up, an island might sink beneath the waves or an invasive species might outcompete another. This normal loss of species through time is known as the background rate of extinction. It is estimated to be around 1 extinction per million species per year, though it varies widely from group to group.

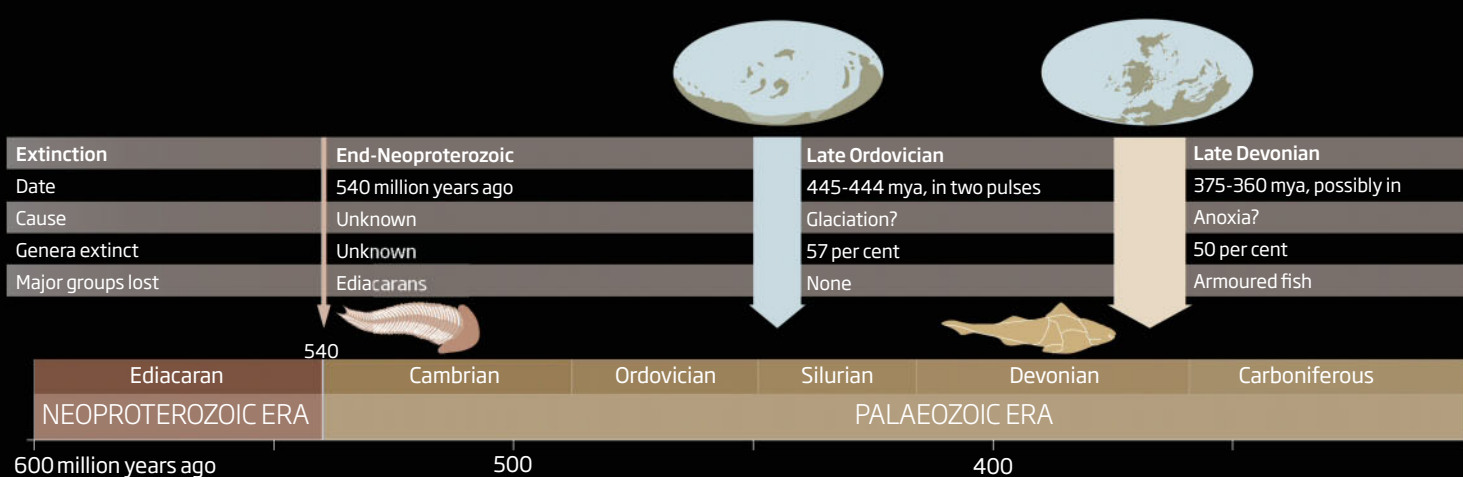
The vast majority of species meet their end in this way. Most dinosaurs did not die out in the asteroid strike – after 165 million years of evolution, hundreds or thousands of species had already been and gone.

Sometimes many species disappear together in a short time. At the end of the ice ages 11,000 years ago, for example, mammoths, woolly rhinos, cave bears and other large mammals adapted to cold conditions died out across Europe and North America. There have been many such "extinction events" through the history of life.

Occasionally extinction events are global in scale, with many species of all ecological types – plants and animals, marine and terrestrial – dying out in a relatively short time all over the world. This is a mass extinction.

There is no exact definition of a mass extinction. The loss of 40 to 50 per cent of species is about the norm, but this is only the upper end of a spectrum of extinction events. There is no set timescale either: some extinctions happen relatively quickly while others take several million years. It depends on the cause.

"Until quite recently, geologists were conditioned against seeing evidence of major crises of any kind"



THE BIG FIVE (OR IS IT SIX, OR SEVEN?)

We now recognise that there have been several mass extinctions over the past 600 million years – the period over which macroscopic life has existed in relative abundance. The first of these was about 540 million years ago, at the end of the Neoproterozoic era (see geological timescale, below), when the enigmatic Ediacaran animals disappeared. Some palaeontologists also identify the late Cambrian as another time of mass extinction.

Three further mass extinctions punctuate the Palaeozoic era. The late Ordovician, between 445 and 444 million years ago, saw substantial losses among the dominant animals of the time: trilobites, brachiopods, corals and graptolites. The late Devonian mass extinction, beginning around 375 million years ago, was another long and drawn out affair. Armoured fish known as placoderms and ostracoderms disappeared, and corals, trilobites and brachiopods suffered some heavy losses. The Palaeozoic ended with the enormous end-Permian mass extinction (see page 98).

Another 50 million years or so passed before the next mass extinction, at the end of the Triassic. Fish, molluscs, brachiopods and other marine groups saw substantial losses, while extinctions on land opened the way for the dinosaurs. They dominated for 165 million years before being wiped out in the most recent extinction, what is today called the Cretaceous-Palaeogene (KPg) event (see page 99).

WHEN THE PENNY DROPPED

Given how important mass extinctions are to understanding the history of life, it may seem surprising that no one was much interested in the idea until the 1970s. Of course, the great Victorian palaeontologists such as Richard Owen and Thomas Huxley were aware that dinosaurs and other ancient creatures were extinct, but they did not see any role for sudden, dramatic events.

Following Charles Darwin, palaeontologists argued that extinction was a normal process: species originated at some point by splitting from existing species, and at some point they died out.

This mindset can be traced back to Charles Lyell, who in the 1830s argued that the foundation of sane geology was uniformitarianism. This holds that "the present is the key to the past": all geological phenomena can be explained by processes we see today, extrapolated over enormous periods of time.

In fact, until quite recently, geologists were conditioned against seeing any evidence of major crises. Woe betide anyone who believed in past impacts and explosions, the marks of an unscientific catastrophist! Until the 1950s geologists even denied that the Earth had been hit by meteorites, arguing, for example, that Meteor Crater in Arizona was a volcanic collapse feature.

This all began to change in the 1960s, a time of ferment and revolution for geologists when ideas of an immobile Earth were rejected in favour of the dynamic reality of plate tectonics.

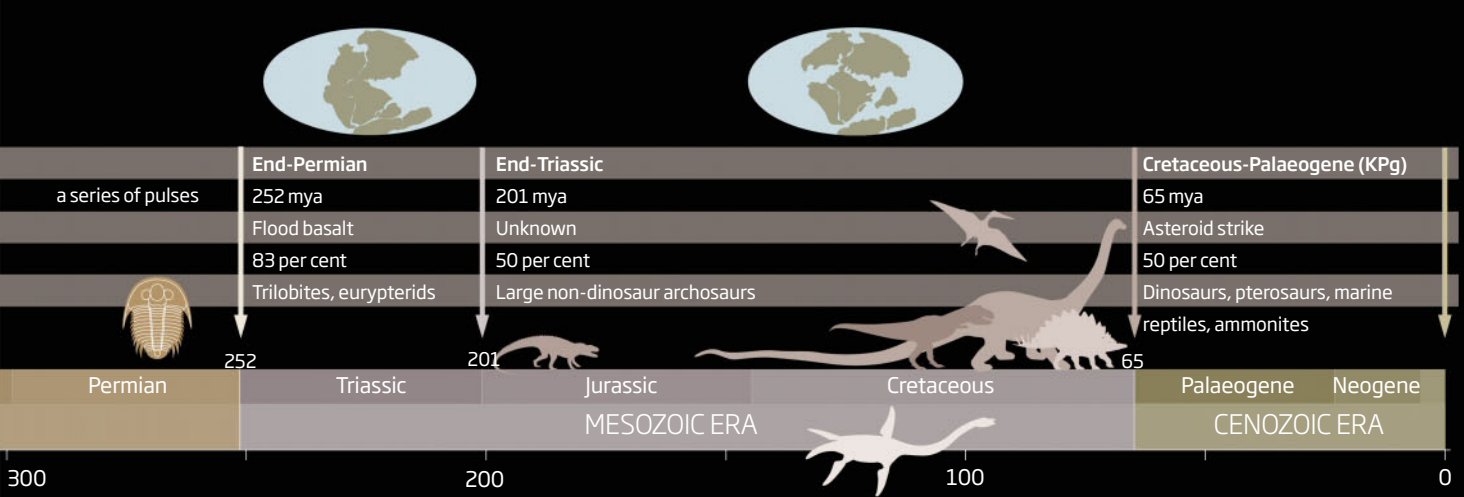
Arizona's Meteor Crater, the birthplace of impact geology

CHARLES AND JOSETTE LENARS/CORBIS

That decade also saw the birth of impact geology. Gene Shoemaker of the California Institute of Technology in Pasadena identified rare minerals, such as coesite and stishovite, in the floor of Meteor Crater, and argued that these were evidence of an impact. At the time such minerals were unknown in nature and had only been created in the lab using enormous temperatures and pressures.

Shoemaker also investigated a large circular depression called Nördlinger Ries in Bavaria, Germany. There he found coesite and stishovite, along with suevite, a type of rock composed of partially melted material. The depression is now considered to be an impact crater some 16 million years old.

Around the same time, palaeontologist Norman Newell of Columbia University in New York began building the case that the fossil record contained evidence of large-scale extinctions. With his work the concept of mass extinctions began to gain currency. Even so, when Luis Alvarez at the University of California, Berkeley, and his colleagues proposed in 1980 that the dinosaurs had been killed off by an asteroid impact the world was still not ready to believe it. Opposition to the idea was substantial, and it took another decade to convince the world that this massive catastrophe really happened.



THE TERRIBLE TWO

By their very nature extinction events are a big deal, but two really stand out, one for its sheer scale and the other for its sudden, spectacular and shocking cause

WHEN LIFE NEARLY DIED

The skull of *Dinogorgon*, which died 252 million years ago along with most other animals and plants

One mass extinction truly dwarfs all the others. Whereas earlier and later events each seem to have extinguished about 50 per cent of species, the end-Permian extinction was associated with a loss of 80 to 90 per cent of species in the sea and on land. Several major groups disappeared, including trilobites and giant sea scorpions called eurypterids.

The vast scale of the extinction is shown by the fact that two major structural ecosystems disappeared – reefs and forests. Nothing like that has happened in any of the other mass extinctions.

Reefs first appeared in the Cambrian, and by the Permian had become a major ecosystem hosting substantial biodiversity, as they do today. With the loss of the dominant reef-builders, the rugose and tabulate corals, the Earth was cleared entirely of reefs. It took 15 million years for new groups of coral to evolve and build reefs once more.

Forests likewise virtually disappeared. There is a famous "coal gap" in the early and middle Triassic when no forests anywhere became sufficiently established to produce coal deposits. Key groups of forest insects, soil churners and vertebrates disappeared too.

Such a huge devastation of life might seem to imply a colossal impact. Evidence for this, however, is weak to non-existent. The most-favoured explanation is volcanic eruptions: 252 million years ago, massive volcanoes erupted in Siberia and they continued to belch forth viscous basalt lava and massive clouds of gases for 500,000 years. These were not conventional cone-shaped volcanoes but great rifts in the Earth's crust. The rock from the eruptions now forms a vast formation known as the Siberian Traps.

Sulphur dioxide caused flash freezing for a short time by blocking the sun, but this gas dissipated rapidly. More long-lasting was the greenhouse gas carbon dioxide, which caused global warming and ocean stagnation. Repeat eruptions kept pumping carbon dioxide into the atmosphere, perhaps overwhelming the normal feedback in which plants mop up the excess through photosynthesis. The warming probably also released frozen masses of methane, an even more potent greenhouse gas, from the deep oceans.

The earliest Triassic rocks contain evidence of repeated cycles of ocean stagnation: their black colour and rich supply of pyrite indicate oxygen-poor conditions. These dark, sulphurous rocks contain very few fossils, in contrast to the abundant and diverse fossils in the limestones just below the extinction level. On land, the volcanic gases mixed with water to produce acid rain. Trees died and were swept away together with the soils they anchored, denuding the landscape. Land animals perished as their food supplies and habitats disappeared.

The slaughter of life in the sea and on land left a devastated Earth. Pulses of flash warming continued for 5 million years, delaying the recovery of life. Some "disaster taxa" such as *Lystrosaurus*, a pig-sized herbivore, gained a foothold here and there, but it took 10 to 15 million years for complex ecosystems to become re-established.

THE DEMISE OF THE DINOSAURS

The extinction of the dinosaurs 65 million years ago, at the Cretaceous-Palaeogene (KPg) boundary, is the most recent of the major mass extinctions and the one most amenable to study. Rocks from before, during and after the event are more abundant, detailed and datable than those for older events. So its cause was just waiting to be resolved.

Up to the 1970s the best evidence suggested that the dinosaurs – along with pterosaurs, mosasaurs, plesiosaurs, pliosaurs, ammonites and many other groups – declined slowly over some 10 million years as a result of cooling climates.

Then came the bombshell. In 1980 Luis Alvarez, who had already won a Nobel prize in physics, his geologist son Walter and other colleagues published an astounding paper in *Science*. The team had set out to use the element iridium as a geological timekeeper, but ended up with remarkably different findings.

Iridium is very rare on Earth's surface, and the minute quantities that are present arrived on meteorites. These hit the Earth at a low but steady rate, so iridium can be used to mark the passage of time: the concentration of iridium in a sedimentary rock indicates how long the rock took to form.

The method worked well when the team applied it to thick sections of sedimentary rock on either side of the KPg boundary at Gubbio in Italy. But at the boundary itself they found a sharp spike in iridium, 10 times the normal amount. If they had stuck to their original hypothesis, they would have concluded that the rocks were laid down by unusually slow sedimentation over a vast time span. But they rejected that in favour of the idea that the spike indicated a sudden influx of iridium from a very large meteorite or asteroid. This, they argued, was what had caused the mass extinction.

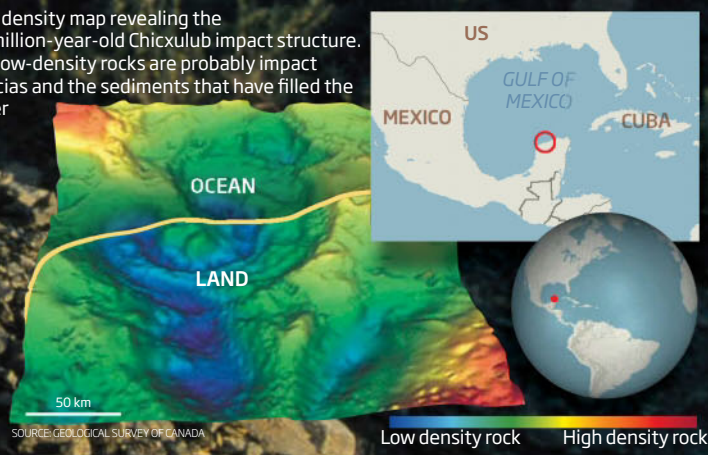
The team reasoned that such an impact would have sent up a vast cloud of dust that encircled the globe, blacking out the sun, preventing photosynthesis and so causing massive loss of life. They calculated that a crater some 100 to 150 kilometres in diameter was required, implying an asteroid 10 kilometres across.

The paper caused an outcry, mainly because it drew such a remarkable conclusion from modest evidence – but such is the stuff of the most daring scientific advances. As the 1980s progressed, geologists found more and more evidence for an impact, including iridium spikes in dozens of locations around the world, the high pressure minerals coesite and stishovite, "shocked" quartz grains, glassy spherules of melted rock and the sudden extinction of many groups of plankton worldwide. Around the Caribbean they also found ancient tsunami debris, and in 1991 the crater itself was identified at Chicxulub on Mexico's Yucatán peninsula (see map, below). As predicted, it was 130 kilometres across.

There are still some serious loose ends to tie up, not least the role played by massive volcanic eruptions on the Deccan plateau of India around the time of the extinction. A handful of geogists dispute whether the impact coincides with the extinction. Even so, the consensus now is that the Alvarez team was right.

ROGER RESSMEYER/CORBIS

Luis (left) and Walter Alvarez in 1985 with a sample of the rock that led to their impact theory



A 3D density map revealing the 66-million-year-old Chicxulub impact structure. The low-density rocks are probably impact breccias and the sediments that have filled the crater

SOURCE: GEOLOGICAL SURVEY OF CANADA

PATTERNS OF EXTINCTION AND RECOVERY

Like unhappy families, all mass extinctions are unhappy in their own way. But their aftermaths are surprisingly similar. It takes millions of years, but life eventually bounces back

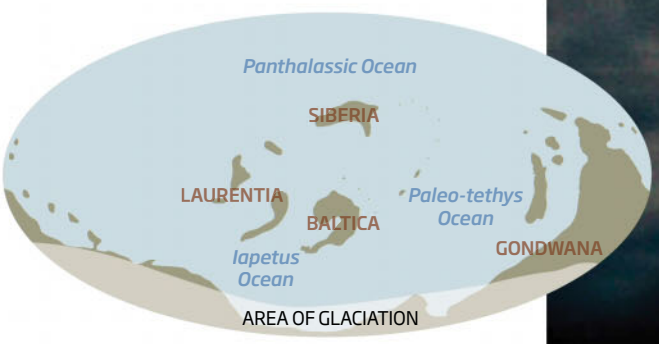
IMPACTS. VOLCANOES. WHAT ELSE?

The causes of two of the largest mass extinctions are now reasonably well understood (see pages 98 and 99). But what of the others? In some cases it is difficult to say. The fossil record clearly shows a huge loss of life but not what caused it. Over the years, a number of possibilities have been put forward, but the cause of two of the big five – the end-Neoproterozoic and end-Triassic – remains uncertain.

CONTINENTAL MOVEMENTS. During the Permian and Triassic, all continents were fused into a supercontinent, Pangaea. At one time, the end-Permian mass extinction was linked to this, based on the suggestion that fusion of continents removes intercontinental seas, each with its own unique fauna, and allows land animals and plants to mix. It now seems, however, that such movements are too slow to lead to massive species loss.

ICE AGES. The late Ordovician mass extinction has been explained as a consequence of a massive ice age, particularly the growth of a huge southern ice cap (see map, below). As the ice spread, species migrated towards the equator and warm-adapted species may have disappeared. Sea levels fell dramatically, reducing many inland seas and causing widespread extinction.

ANOXIA. The late Devonian extinction has been linked to a lack of oxygen in the ocean, possibly caused by sudden temperature changes or massive increases in the supply of sediment from the land caused by the rise of terrestrial plants.



IS THERE A COMMON PATTERN?

In the 1980s, as the Alvarez hypothesis gained ground, it seemed reasonable to assume that all mass extinctions were caused by impacts. Though there have been numerous "discoveries" of craters and other impact signatures coinciding with the other mass extinctions, none has stood up to scrutiny. It now seems that the KPg event was unique – the only mass extinction caused by an impact. In fact, we now think that each mass extinction had its own unique cause.

Another idea that was fashionable in the 1980s was that mass extinctions are periodic. Some palaeontologists claimed to have found patterns in the fossil record showing a mass extinction every 26 million years, and they explained this by suggesting that a "death star", dubbed Nemesis, periodically swings into our solar system and perturbs the meteorite cloud. But Nemesis has never been found and evidence for this pattern is now widely doubted.

Common features have emerged, however. For example, it does seem that some species are more vulnerable to extinction than others. Large body size makes animals especially susceptible as it is associated with high food requirements, large feeding range and small population size. Species with specialised diets or limited distribution are also likely to suffer. In contrast, the survivors tend to have large population sizes, live in many habitats in many parts of the world, and have a varied diet.

This is not to say that mass extinctions are highly selective. David Raup at the University of Chicago famously characterised the death of species during mass extinctions as the result of "bad luck rather than bad genes", meaning that the normal rules of natural selection break down. Their success – or lack of it – in normal times has little bearing on their chances of survival when the meteorite hits or the volcano erupts. This holds lessons for current and future extinctions (see page 102). For example, if humans destroy habitats wholesale then all species are vulnerable, whatever their size, diet or habitat.

LIFE REBOUNDS

"In South America, giant birds and crocodilians vied to replace dinosaurs as the top carnivores"

Mass extinctions are devastating, and yet life eventually returns to normal. The rate of recovery depends on many factors, but the most important is the scale of the extinction.

After most mass extinctions life recovers within a few million years, though the end-Permian event was different. It was twice as large as most of the others, and so it is no surprise that the recovery time was greatest.

Recovery also depends on which plants and animals survive. If the mass extinction hit all groups more or less equally, as most seem to, then there is a good chance that one or two species from each major group will survive. These act as an ecological framework, occupying most of the broad niches, and so the basic ecosystem structure survives. New species evolve to fill the gaps and the recovered ecosystem may be quite comparable to the one that existed before the disaster.

A more selective event, on the other hand, might leave broad sectors of ecospace vacant. A variety of the survivors then jockey for position, evolving to fill the vacant niches.

After the KPg event it was by no means a foregone conclusion that mammals would take over. Indeed, in North America and Europe, giant flightless birds became the dominant carnivores, some of them famously preying upon ancestral (admittedly terrier-sized) horses. In South America, giant birds and crocodilians vied with each other to become the top carnivores, and mammals only replaced them some 30 million years later.

Mass extinctions, then, have a creative side. Marginal groups sometimes get a chance to expand and become dominant. Most famously, mammals benefited from the demise of the dinosaurs. In fact, mammals first evolved in the late Triassic, at the same time as the dinosaurs, but they remained small and probably nocturnal because dinosaurs occupied all the key niches.

The end-Permian mass extinction was even more creative, with a yawning post-extinction eco-space providing opportunities for the survivors. In the sea, molluscs (bivalves and gastropods) took over roles previously occupied by brachiopods. Scleractinian corals rebuilt the reefs, and new kinds of light-scaled fish moved into roles previously occupied by more primitive ones. On land, the key beneficiaries of the extinction might have been the dinosaurs, whose earliest ancestors emerged within 5 million years of the crisis.

Parts of the post-dinosaur world were briefly ruled by giant birds like *Gastornis* (above), before mammals took over

MAIN: ANDREI SPIRACHE/GETTY

BOTTOM: D.CHAPMAN/WOODFALL WILDIMAGES/NHPA; BELOW: DEAGOSTINI/GETTY

BRISTOL UNIVERSITY

Michael J. Benton

Michael J. Benton is professor of vertebrate palaeontology at the University of Bristol in the UK. His research focuses on the end-Permian mass extinction.

APOCALYPSE NOW?

It is often said that we are living through the sixth mass extinction, this one induced by human activity. The point is well made: the present biodiversity crisis appears to be comparable in scale to many of the biotic crises of the past.

There can be no doubt that many species have gone extinct on our watch. We know, for example, that the dodo was last seen in 1662, the last great auk was killed by collectors in 1844 and the last passenger pigeon died in a zoo in 1914. Hunters shot the last quagga, a zebra-like wild horse, in the 1870s and the last thylacine – or Tasmanian tiger – died in captivity in 1936.

These examples, however, tell us little about the scale of the crisis. For that we have to aggregate known historical extinctions. Unfortunately the records are not good, but we do know that 130 species of bird were driven to extinction by hunting between 1500 and 2000. This gives us a starting point.

There are currently some 10,000 bird species, so these extinctions represent a loss of 1.3 per cent of species in 500 years, or 26 extinctions per million species per year – much greater than the background rate of extinction (see page 96).

Even this could be an underestimate because many other bird species might have become extinct in that time without being recorded. What is more, extinction rates have arguably risen in recent years due to habitat destruction. Taking these factors into account has yielded an alternative figure of about 100 extinctions per million species per year.

If we assume this applies to all of the estimated 10 million species on Earth, total losses might now be 1000 species per year, or three species every day. This is a very rough estimate but it suggests claims of a sixth mass extinction are not exaggerated.

It could of course be objected that this rate of loss cannot proceed inexorably. The optimist might argue, for example, that most of the species so far driven to extinction were already rare or vulnerable, and that they were hunted without mercy in less enlightened times. There is surely some truth in these assertions: it is unlikely that globally distributed species such as sparrows, rats or mice would be so easy to exterminate as the dodo. Further, no nation would allow hunters to slaughter animals as systematically as was done by Victorian-age hunting parties.

However, despite tighter controls on hunting and increasing conservation efforts, pressure on natural habitats has never been more extreme.

While it is frustratingly hard to put precise figures on current rates of species loss, uncertainties should not be seen as a reason for complacency. The fossil record shows how devastating mass extinctions are and that, although life does recover, it takes millions of years to do so. The study of mass extinctions, and comparisons with the modern world, show that we are almost certainly responsible for another mass extinction, and the living world could soon be a much-diminished place.

RECOMMENDED READING

Mass Extinctions and their Aftermath by Tony Hallam and Paul Wignall (Oxford University Press)

When Life Nearly Died: The greatest mass extinction of all time by Michael J. Benton (Thames & Hudson)

T. rex and the Crater of Doom by Walter Alvarez (Princeton University Press)

Alvarez's original paper *Science* (vol 208, p1085) *Vanishing Life: The mystery of mass extinctions* by Jeff Hecht (Prentice Hall & IBD)

Cover image: Jonathan Blair/NGS

COMING SOON

NewScientist
THE COLLECTION

VOL THREE / ISSUE ONE

DISCOVERING SPACE

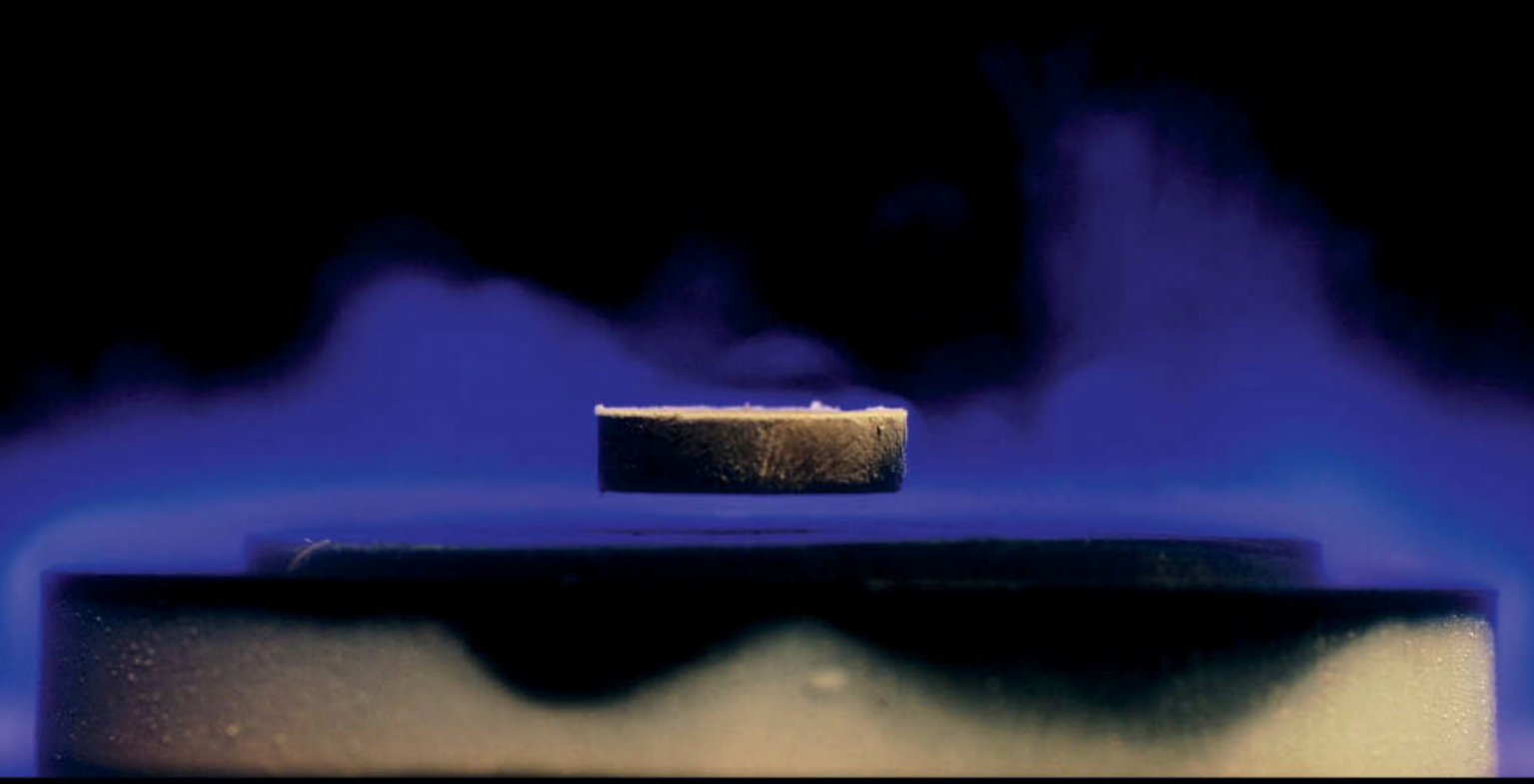
THE WONDERS OF OUR SOLAR SYSTEM, THE MILKY WAY AND BEYOND

ON SALE 17 FEBRUARY

To buy back issues of *New Scientist: The Collection*, visit *newscientist.com/TheCollection*

The Big Questions / The Unknown Universe / The Scientific Guide to a Better You
The Human Story / The Human Brain / Medical Frontiers / Being Human / Our Planet

NASA/JPL



CHAPTER FIVE

SHAPING THE FUTURE

SUPERCONDUCTIVITY

STEPHEN BLUNDELL

INSTANT
EXPERT

LOW-TEMPERATURE SUPERCONDUCTORS

In 1911, a Dutch physicist called Heike Kamerlingh Onnes and his team made a remarkable and completely unexpected discovery. They found that certain metals completely lose their electrical resistance when cooled to within a few degrees of absolute zero. A coil of wire made from such a metal could carry an electrical current round and round forever, without needing a power source to drive the current. No one had predicted this phenomenon and at the time no one could explain it. The effect was named superconductivity – the prefix "super" implying that it was in a completely different league from anything seen before. It took more than half a century to figure out how superconductivity might work and how to make it useful. More recently, though, we have come to realise that we understand the phenomenon far less well than we thought.

THE BIG DISCOVERY

Heike Kamerlingh Onnes spent most of his scientific career in the quest to achieve the lowest temperatures possible. His big breakthrough came in 1908 when he managed to make liquid helium, the last of the gases to be liquefied because it requires temperatures as low as 4 degrees above absolute zero – the lowest possible temperature, equivalent to 0 kelvin or -273.15 °C.

Kamerlingh Onnes was one of the first people to really understand that advances in low-temperature physics critically depended on having first-rate technicians, expert glassblowers and skilled craftspeople to build and maintain the delicate equipment. It was not enough to potter around alone in a ramshackle laboratory, as many of his contemporaries did. For this reason Kamerlingh Onnes's team at the University of Leiden in the Netherlands beat physicist James Dewar at the Royal Institution in the race to make liquid helium.

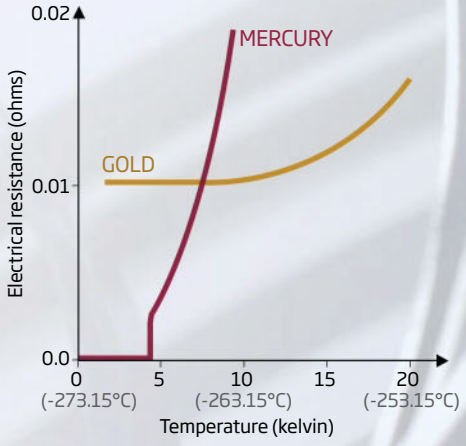
Once he had the technology to achieve the lowest temperatures available anywhere, Kamerlingh Onnes started exploring how matter behaves at low temperature. One of the questions he wanted to answer was how the electrical resistance of a metal changes as it approaches absolute zero. Experiments had shown that resistance falls as you cool from room temperature, but what happens when you get really cold? Some had speculated that resistance would fall steadily to zero, others that it would rise sharply as the electrons froze, or even simply flatten off at a constant value.

The first studies indicated that it flattened off, but Kamerlingh Onnes realised that this was actually down to the presence of impurities in the metal causing electrons to scatter. He needed a metal that was really pure and so he chose to focus on mercury, a liquid at room temperature that could be repeatedly distilled to make it as pure as possible. To make wires of mercury, his technician filled very narrow U-shaped glass capillaries and then carefully froze them. The capillaries had electrodes at either end so it was possible to pass a current through them and measure the resistance at various temperatures.

On 8 April 1911, Kamerlingh Onnes and his team observed that the resistance of the mercury wire suddenly disappeared below around 4 kelvin (see graph, right). They had discovered superconductivity. Subsequent experiments showed that mercury would behave this way even at lower purity, and that some other metals exhibited superconductivity, including tin and lead.

SCIENCE PHOTO LIBRARY

The discovery by Heike Kamerlingh Onnes took everyone by surprise



Superconductors repel magnetic fields, giving them the ability to levitate

CHARLES O'REAR/CORBIS-AMERICAN SUPERCONDUCTORS

LEVITATION

Superconductivity has another trick up its sleeve besides the loss of electrical resistance. In 1933, German physicists Walther Meissner and Robert Ochsenfeld performed an experiment at Germany's national measurement laboratory in Berlin to study what happens to magnetic fields near a superconducting material as it is cooled.

Magnetic fields are normally quite content to pass through any material, but as soon as the temperature fell low enough for superconductivity to occur, Meissner and Ochsenfeld found that the magnetic field was expelled from their material. It couldn't pass through, but instead was forced to pass around the superconductor, doing a circuitous detour as if sensing some kind of invisible "NO ENTRY" sign.

The Meissner effect, as it is now known, is responsible for the incredible ability of superconductors to levitate above magnets – or indeed for magnets to levitate above superconductors.

Magnetic fields are unable to go through the superconductor due to currents that run across its surface. These are known as screening currents, because they act to screen the interior of the superconductor from the externally applied magnetic field. The currents themselves also produce their own magnetic field around the superconductor that repels an external one. This repulsive force balances the force of gravity and allows the superconductor to hover eerily in mid-air.

Later in the 1930s, Fritz and Heinz London, two brothers working at the University of Oxford, developed a theoretical framework for understanding the Meissner effect. Their results hinted that the response of the superconductor to a magnetic field is due to electrons in the superconductor being locked into a peculiar configuration that Fritz London christened a "macroscopic quantum state".

THE QUANTUM CONNECTION

Theoretical physicist Fritz London was the first to show that superconductivity is a quantum phenomenon. We can understand what led to his idea with a thought experiment involving two black boxes. One box contains a permanent magnet and the other a coil of wire connected to a battery. At first the two boxes are indistinguishable – the current produced by the battery creates the same pattern of magnetic field lines as the permanent magnet. The similarity doesn't end there because the permanent magnet's field comes from microscopic electric currents within it.

Yet there is a difference between the boxes. Leave them for a while and the magnetic field around the box containing the circuit will disappear when the battery goes flat. In contrast, the permanent magnet retains its magnetism and its field pattern is unchanged. Why is it that the currents driving the permanent magnet are not subject to wear and tear, while those driven by batteries are?

This puzzle is related to the behaviour of an atom's electrons. Electrons orbit around the nucleus, creating microscopic currents, and because they normally stay in their orbits their energy levels do not wear out or run down. In the early 20th century, quantum mechanics began to explain this in terms of electrons residing in what are called stationary states and this is the origin of permanent magnetism.

London grasped that similar reasoning could apply to the behaviour of a superconductor. The current in a superconducting coil behaves much more like the currents created by electrons around atoms in a magnet than those in an ordinary coil of wire. Indeed, if a third black box contained a coil of superconducting wire, a current started around it, and its resulting magnetic field, would last indefinitely.

This is why London thought of superconductivity as a macroscopic quantum state. He realised that a superconducting wire was just like a big atom, with the current going round and round the wire just like the electrons orbiting round and round the atom. Both are immune from wear, tear and decay.

HIGH-TEMPERATURE SUPERCONDUCTORS

Within 50 years of the discovery of superconductivity, an elegant theory was in place that explained all of its effects. Superconductivity was essentially a solved problem. Then, in 1986, one discovery upset everything. Its implications are still unravelling and we are yet to find a theory to account for it. What we really want, however, is a superconductor that operates at room temperature. Might one be within our grasp?

THE THEORY OF BARDEEN, COOPER AND SCHRIEFFER

SCIENCE PHOTO LIBRARY

The most important breakthrough in understanding superconductivity near absolute zero came from the work of John Bardeen, Leon Cooper and Robert Schrieffer in 1957. Bardeen (pictured left) already had one Nobel prize in physics for his part in the invention of the transistor, and the work on superconductivity would earn him his second, shared with Cooper and Schrieffer.

The ideas they worked on together are now known as BCS theory and provide a description of the superconducting state in terms of interactions between pairs of electrons. Because they have the same negative charge, electrons tend to repel each other, but this can change in certain materials that have a crystal lattice. The lattice vibrates with more or less energy depending on the temperature. When it is very cold, the gentle vibrations can push electrons together, producing a net attractive force that drives them to pair up. BCS theory shows how this tendency to pair up can result in superconductivity.

A current is essentially a flow of electrons. This flow can be knocked off course by lattice vibrations and impurities that scatter the electrons, and this is the source of electrical resistance in normal metals.

Scattering occurs all the time in superconductors too. But at superconducting temperatures, a scattered pair of electrons will stay together and keep the same net momentum even though the momenta of the individual electrons may change. Rather like two people running along holding hands, the pairs keep to the same direction and hence so does the current.

BCS theory explains pretty much all of the properties observed in superconductors up to the mid-1980s, by which time we had learned how to make and deploy technological applications such as superconducting magnets. It also suggested that superconductivity is purely a low-temperature phenomenon. However, all that was about to change.

THE DISCOVERY OF HIGH-TEMPERATURE SUPERCONDUCTORS

Fifty years after its discovery, superconductivity had been found in a number of elemental metals and also in alloys. No one expected to see the effect in an oxide. Oxidation means rust or tarnishing – something scientists working with metals generally don't want. Yet in the late 1960s, superconductivity turned up in an oxide called strontium titanate, cooled to well below 1 kelvin. Alex Müller (pictured below), working at IBM's research laboratory in Zurich, Switzerland, was one of very few people to suspect that this discovery heralded an exciting new possibility. Together with colleague Georg Bednorz, Müller beavered away at preparing various oxides and studying them.

The breakthrough came when the pair spotted a report from a French research group on an oxide compound containing barium, lanthanum and copper. The French team had found their sample conducted like a metal, which was highly unusual for an oxide. But they only studied its properties at high temperatures; they were more interested in its possible use as a catalyst for certain chemical reactions.

IMAGES COURTESY OF IBM RESEARCH – ZURICH

Bednorz and Müller suspected there might be more to the unusual oxide. They immediately began preparing samples containing the same elements, though with different compositions, and cooled them to much lower temperatures. Bednorz and Müller's suspicions were correct. By January 1986, they had found superconductivity and with some further tuning saw it at 30 K – a dizzy new height for a superconductor.

The oxide broke all records, but once the idea was out it was just a matter of optimising the chemical composition. The basic formula seemed to be to keep a structure with planes of copper and oxygen and vary the other atoms. By following this, new superconductors were discovered that worked at higher temperatures. Within a year, yttrium barium copper oxide – also known as YBCO and pronounced ibb-ko – was found to superconduct at 93 K. The temperature record hit 135 K by 1993 and even 150 K when the compound was squeezed to high pressure.

The era of high-temperature superconductivity had begun.

NEW AVENUES

With the discovery of superconductivity in oxides, of all things, a concerted attempt ensued to find other materials exhibiting the phenomenon. This field of research is a bit like prospecting during the great US gold rush of the 19th century: there is a lot of hacking through various regions of unpromising rock, but once somebody in an isolated valley stumbles on signs of a rich seam, everyone else quickly arrives with their hammers and starts to hunt nearby.

In this way, researchers recently found a new family of superconducting compounds containing iron. The initial discovery seems unexciting: the compound containing lanthanum, oxygen, iron and phosphorus (LaOFeP) transformed into a superconductor at around 3 kelvin, which is hardly dramatic in terms of temperature. However, the presence of iron raised a few eyebrows. Iron atoms are magnetic and not the sort of constituent you would expect to see in a superconductor. That's because magnetic fields usually rip apart the electron pairs needed for superconductivity.

So researchers were keen to find out more about this most unusual superconductor. As they switched lanthanum for samarium, arsenic for phosphorus, and replaced some of the oxygen atoms with fluorine, they succeeded in finding a compound that superconducted at 55 K.

What makes it so exciting is that it sets a record for a superconductor that does not contain copper, hitherto a crucial constituent of high-temperature superconductors. Like other high-temperature superconductors, the new family have a layered structure; in this case layers of iron and arsenic are interleaved with samarium-oxygen layers.

Sometimes the discovery is completely accidental, as Jun Akimitsu at Aoyama Gakuin University in Tokyo, Japan, found. In 2000, he was trying to isolate a complicated compound when he found that his sample had an impurity in it that seemed to be superconducting up to 39 K. He isolated the impurity and found it to be magnesium diboride (MgB_2).

This incredibly simple compound has been known since the 1950s and sits in a chemical jar in pretty much every chemistry lab in the world. Nobody had ever thought to measure its electrical conductivity at low temperature and so this extremely good superconductor had remained undiscovered. It seems like MgB_2 is going to be very useful. It superconducts at the relatively balmy temperature of 39 K, it can be made into wires that carry large currents and is relatively inexpensive.

In 2015, the record transition temperature passed to hydrogen sulphide (H_2S) subjected to a pressure of more than a million atmospheres. Although best known as a gas that smells of bad eggs, at extremely high pressures hydrogen sulphide becomes a superconducting crystal with a transition temperature of 203 K (an astonishing -70 °C). This high transition temperature is a direct result of the light mass of the hydrogen atom, ensuring that the hydrogen-sulphur bond vibrates at a high frequency that drives up the temperature scale of the superconductivity.

One uncomfortably large sticking point remains, though: how do these superconductors work? Earlier theories based on electron pairs only explain the phenomenon at very low temperatures, and finding the right conceptual framework for understanding these "unconventional" superconductors has proved to be very tough – although incidentally MgB_2 and H2S appear to superconduct along the lines that Bardeen, Cooper and Schrieffer prescribed. Solid-state chemists are therefore left to carry on hacking away in the mines looking for the elusive chemical composition that will herald the next gold rush.

APPLICATIONS OF SUPERCONDUCTORS

Superconductivity is not only fascinating, it is also incredibly useful. Superconductors are already used in applications as diverse as seeing inside the human body and discovering the origin of mass. As important as these achievements are, their promise for future revolutionary technologies may be even greater.

SEEING INSIDE THE HUMAN BODY

Heike Kamerlingh Onnes realised that one of the most important applications of superconductors would be in making powerful electromagnets. Superconducting wire can carry immense electrical currents with no heating, which allows it to generate large magnetic fields. An electromagnet with non-superconducting copper windings would melt with the same current.

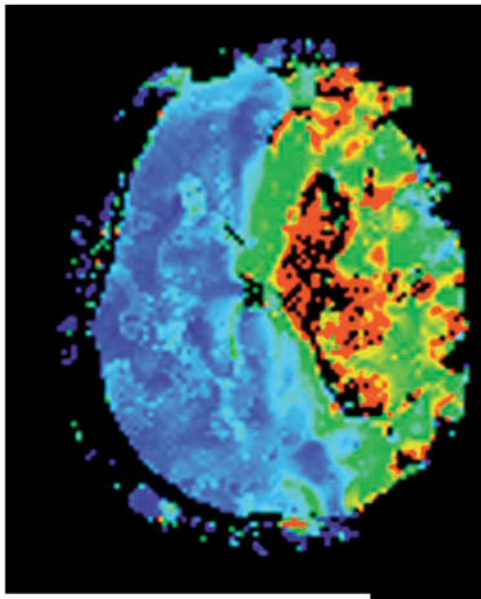
Unfortunately, the superconductors available to Kamerlingh Onnes could only carry small currents producing correspondingly small magnetic fields and so he never realised this possibility in his lifetime. It took until the late 1950s and early 1960s for the right materials to be identified and the relevant technology developed.

One of the most important applications of superconducting magnets is in medicine, with the development of magnetic resonance imaging. MRI is the best way to see inside the body without invasive surgery.

Today MRI is used to examine the body's soft tissues and is especially valuable for detecting tumours, examining neurological functions and revealing disorders in joints, muscles, the heart and blood vessels. It shows up the water content of tissue, which varies throughout the body and is altered by pathological processes in many diseases.

Water molecules are composed of hydrogen and oxygen atoms, and it is the hydrogen nuclei that MRI probes. Hydrogen nuclei behave like little spinning tops and placing them in a magnetic field causes the spin axis to tip over and rotate at a well-defined frequency called the precession frequency. The nuclei can interact with a second electromagnetic field whose frequency is tuned to match the precession frequency. When this happens, the nuclei absorb energy, allowing you to work out how much water is present and where.

In an MRI scanner, a superconducting magnet provides the magnetic field that starts the nuclei precessing. To produce high-resolution images typically requires a field between 1 and 3 tesla, tens of thousands of times larger than the magnetic field at the Earth's surface. The magnet also needs to be large enough for a person to slide inside its bore.



SIMON FRASER/SCIENCE PHOTO LIBRARY

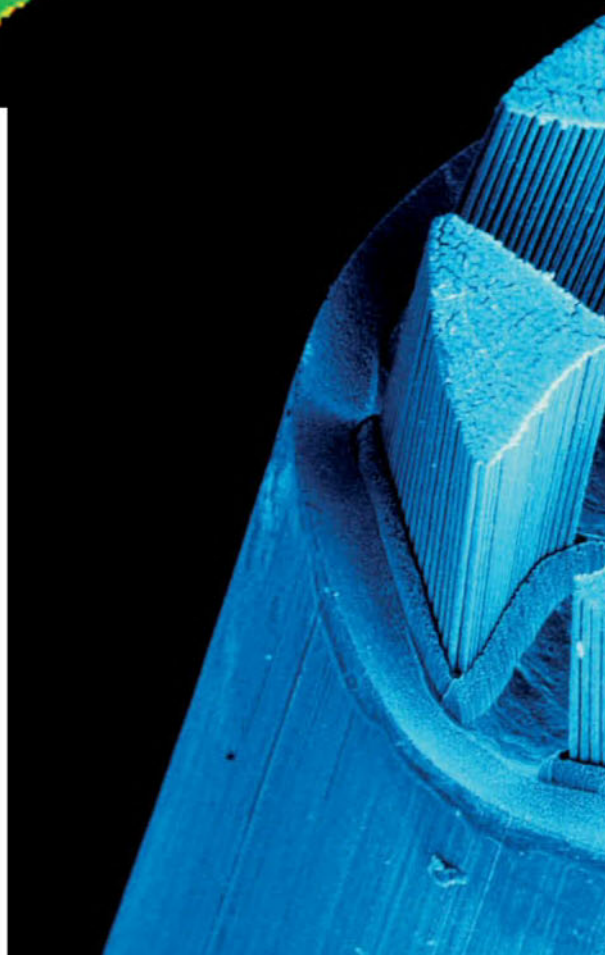
fMRI reveals what happened to the brain of a person who had a stroke

The exquisite 3D images of the body are accomplished using a sophisticated sequence of electromagnetic pulses and a magnetic field gradient – techniques that won the British physicist Peter Mansfield and American chemist Paul Lauterbur the Nobel prize for medicine or physiology in 2003.

There are now several variants of MRI, including functional MRI (fMRI) that monitors processes such as blood flow in the brain in response to particular stimuli.

Superconductors have given the medical profession something even better than X-ray spectacles, and hundreds of thousands of people a year get a much better medical diagnosis because of them.

"Superconductors have given the medical profession something even better than X-ray spectacles"



HUNTING THE HIGGS BOSON

Particle accelerators need magnets to manipulate beams of highly energetic particles. As the energy of these beams has increased over the decades, so has the need for ever stronger magnetic fields. Only superconducting magnets can do the job.

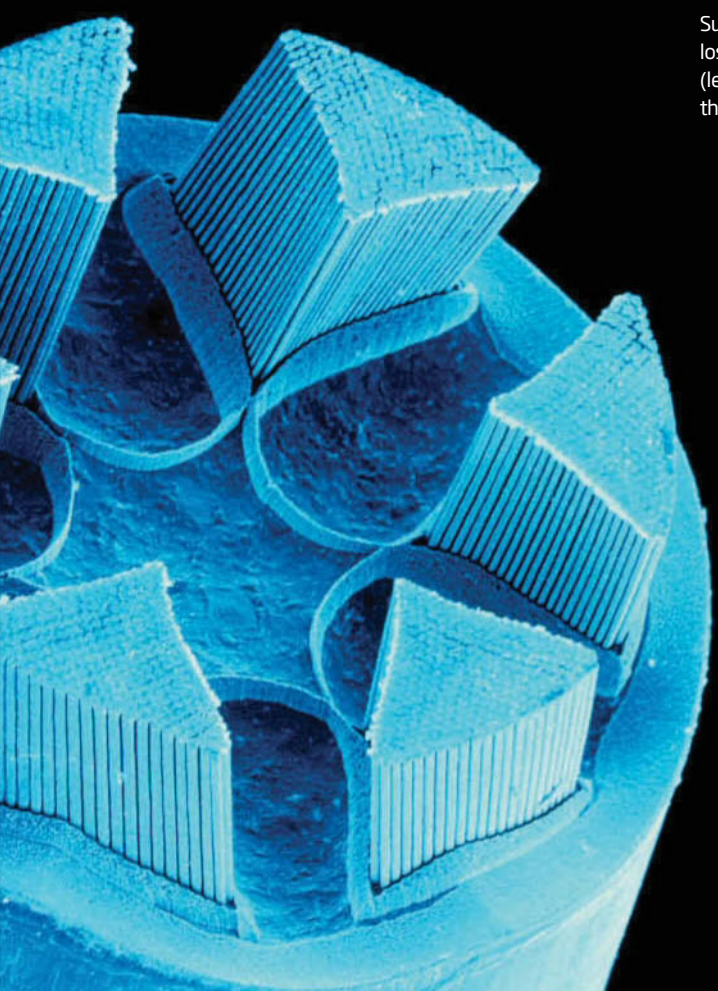
In fact, accelerators have been using superconducting magnets since the 1970s and the Large Hadron Collider at CERN near Geneva, Switzerland, is no exception. Particle physicists designed the LHC to explain the origin of mass by searching for a particle called the Higgs boson among the debris of collisions between high-energy protons.

Beams of protons are accelerated in opposite directions around a circular tunnel 27 kilometres round under the French-Swiss border. To steer the protons requires a large magnetic field all the way around the ring. The LHC comprises 1232 superconducting magnets, each 15 metres long and weighing 35 tonnes. The magnets contain coils of superconducting cables made from niobium and titanium and cooled to a little over 1 kelvin using 100 tonnes of liquid helium.

The stored energy in the magnets at the LHC is approximately 15 gigajoules, more than the kinetic energy of a fully laden Boeing 747 at cruising speed. In September 2008, days after the LHC was switched on, an electrical fault caused the superconducting magnets to fail. Around 100 were affected and repairs delayed the LHC's operation for over a year.

CERN

These events are now forgotten in the afterglow of the discovery of the Higgs boson. Increases in the energy and number of collisions at the LHC rely on producing even larger magnetic fields. Niobium-titanium magnets cannot deliver magnetic fields stronger than 8 tesla, so present research is focused on niobium-tin magnets that could double this figure. Thus future advances in particle physics depend on a deeper understanding of superconducting technology.



Superconductors form lossless power cables (left) and magnets for the LHC (above)

MANFRED KAGE/SCIENCE PHOTO LIBRARY

NEW APPLICATIONS

Superconductors have such remarkable properties that they find many diverse applications. Magnetically levitated (maglev) trains, for example, exploit a superconductor's ability to repel magnetic fields. At the Yamanashi test line in Japan, a superconducting coil attached to the train keeps it floating above the magnetic track thus avoiding the slowing effects of friction. The train has reached a top speed of 581 kilometres per hour – making it the fastest land-based mass transportation system in the world. However, it remains a prototype because it is so expensive to run.

Meanwhile, circuits containing a superconducting coil provide unprecedented sensitivity to tiny magnetic fields. Any magnetic field passing through the coil generates a screening current that subtly

alters the current flowing in the circuit. This idea is at the heart of SQUIDs (superconducting quantum interference devices), which are used to detect the very small magnetic fields generated by currents in the heart and brain. SQUIDs have also been used in research for sensing X-rays, gamma rays and various exotic particles. They are even playing an important part in the search for dark matter in the universe.

The properties of superconductors have led to novel energy storage systems. One is based on a large flywheel driven to high speeds by electricity during the night when it is cheaper. The flywheel has frictionless superconducting bearings, so it doesn't lose its stock of rotational kinetic energy. As a result, it can store energy until it is needed.

A second system stores current in a superconducting magnet whose energy is released when needed by connecting the magnet to a circuit. While both of these systems are attractive in principle, and several commercial systems are available, the high cost of refrigeration has not yet made this sort of technology economically viable.

The lack of electrical resistance means that superconductivity has obvious applications in power transmission. Several companies have made cables out of high-temperature superconductors that can carry currents of 2000 amps without any losses. Similar resistance-free cables are used in electrical circuits in some MRI scanners, radio telescopes and cellphone masts.

If we can do all this with today's superconductors, we can only speculate what applications may become possible with the next generation of materials.

Stephen Blundell

Stephen Blundell is professor of physics at the University of Oxford. He works on various problems in magnetism and superconductivity, particularly molecule-based systems, using implanted muons and high-magnetic fields.

SUPERCONDUCTIVITY 100 YEARS ON

A little over a century after the discovery of superconductivity, what have we learned? One big lesson is that the time from the discovery of a phenomenon to important applications is long. It took the best part of 50 years to work out how to make a superconducting magnet and half as much time again to develop the first killer application, magnetic resonance imaging. The long timescale is not for want of trying – it is just that getting all the conceptual and technological pieces in place for an application is very difficult.

Superconducting materials are still surprising us. The iron-based superconductors discovered three years ago were utterly unexpected, and this highlights how much we still have to learn. Like a master chef searching for that elusive perfect recipe, we are still working by something only marginally more informed than trial and error to search for a revolutionary room-temperature superconductor.

Part of the key to this may be in properly understanding how high-temperature superconductors work. Despite their discovery some 30 years ago, as yet we have no universally accepted theory. The problem is very complex because it involves the interplay of many factors: the crystal lattice, magnetism, orbitals and electrons. Different theories stress the role of different components of the problem but we seem to be missing the right way of looking at it. The next few years could well see these issues starting to be resolved.

Applications may have been slow in coming, yet how can you fail to be impressed when you look at an MRI scanner? Aside from transforming medicine, it is an extraordinary feat of engineering and large enough to fit a person inside. The sense of wonder only increases when you realise that existing all around the circumference of its gigantic coil there is a persistent current going round and round.

This current is described by a superconducting wave function, a fully coherent macroscopic quantum object, as untarnished by everyday wear and tear as an electron in an atom performing its stately, never-ending orbit around the nucleus.

RECOMMENDED READING

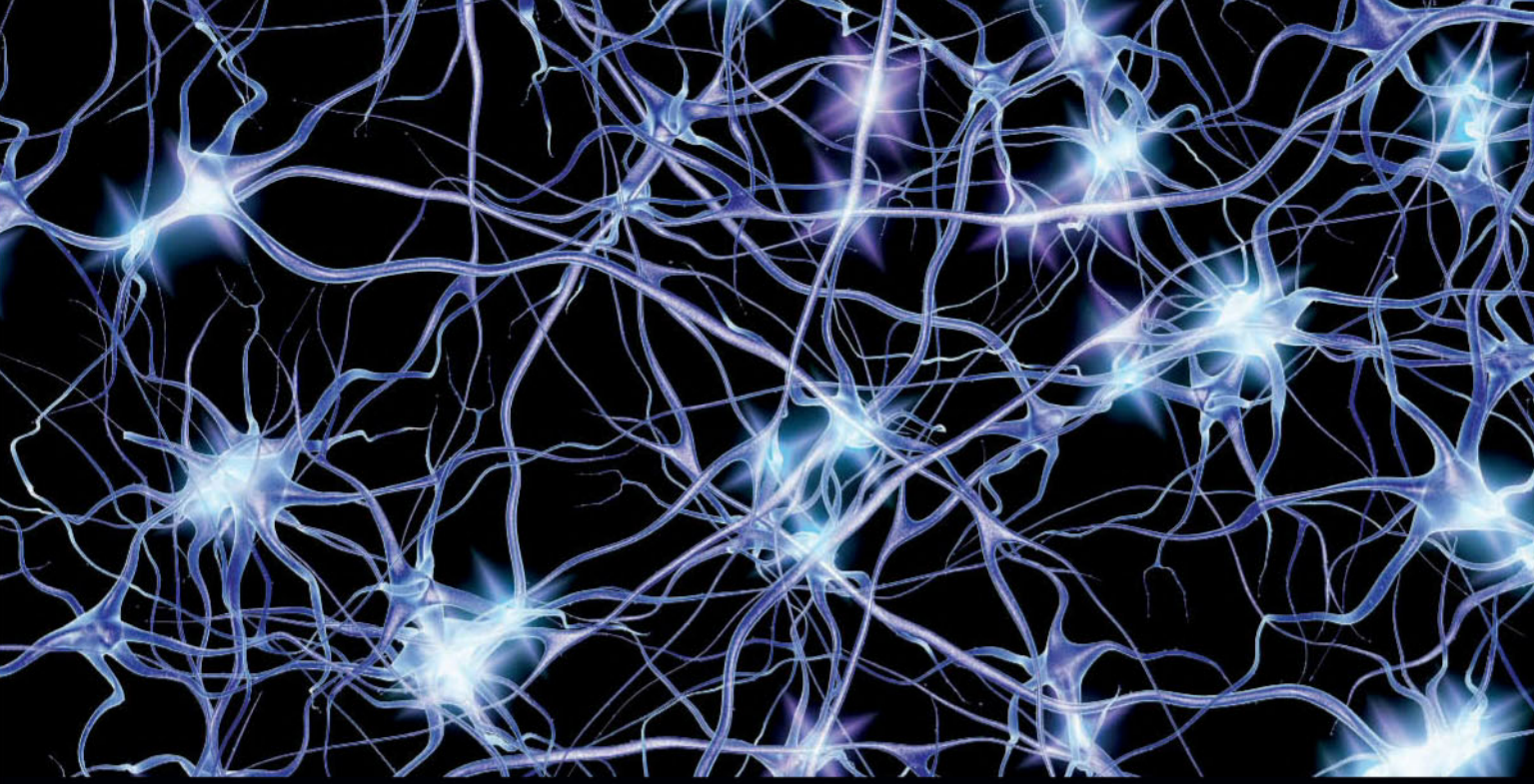
Superconductivity: A very short introduction by S.J. Blundell, Oxford University Press, 2008

Superconductivity, Superfluids and Condensates by J.F. Annett, Oxford University Press, 2004

"Superconductivity, the first 100 years", *Physics World*, vol 24, p 17

"Happy 100th, superconductivity", *Science*, vol 332, p 189

Cover image: D. Parker/University of Birmingham High TC Consortium/SPL



ARTIFICIAL INTELLIGENCE

PETER NORVIG

INSTANT
EXPERT

EARLY AMBITIONS

We have long suspected that intelligence is not exclusively a human quality, and that it is possible to build machines capable of reasoning and learning as well as a human can. But what seemed straightforward at first has turned out to be surprisingly difficult

WHAT IS ARTIFICIAL INTELLIGENCE?

The field of artificial intelligence (AI) is the science and engineering of machines that act intelligently. That raises a vexing question: what is "intelligent"? In many ways, "unintelligent" machines are already far smarter than we are. But we don't call a program smart for multiplying massive numbers or keeping track of thousands of bank balances; we just say it is correct. We reserve the word intelligent for uniquely human abilities, such as recognising a familiar face, negotiating rush-hour traffic, or mastering a musical instrument.

Why is it so difficult to program a machine to do these things? Traditionally, a programmer will start off knowing what task they want a computer to do. The knack in AI is getting a computer to do the right thing when you don't know what that might be.

In the real world, uncertainty takes many forms. It could be an opponent trying to prevent you from reaching your goal, say. It could be that the repercussions of one decision do not become apparent until later – you might swerve your car to avoid a collision without knowing if it is safe to do so – or that new information becomes available during a task. An intelligent program must be capable of handling all this input and more.

To approximate human intelligence, a system must not only model a task, but also model the world in which that task is undertaken. It must sense its environment and then act on it, modifying and adjusting its own actions accordingly. Only when a machine can make the right decision in uncertain circumstances can it be said to be intelligent.

The future of AI is ever-changing in popular culture

After ambitious beginnings, the discipline of artificial intelligence quickly split into several interrelated subfields

1950
Alan Turing proposes that a digital computer could be programmed to answer questions as accurately as a human can

1956
The term "artificial intelligence" is coined at Dartmouth University by the creators of the nascent field

1958
Allen Newell and Herbert Simon predict that "within 10 years a digital computer will be the world's chess champion". It actually takes 40

1961
Computer program solves calculus problems at the first-year university level

1965
The first chatbot, a psychotherapy program called ELIZA, carries on rudimentary conversations

1967
The STUDENT program solves high-school-level algebra problem, written in words

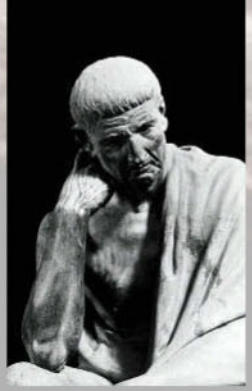
1973
Freddy Robot uses visual perception to locate and assemble models

1974
"AI winter" sets in as government funding agencies cut back on their investment in AI research

1975
Stanford University's Meta-Dendral program discovers previously unknown rules about molecules, published in the *Journal of the American Chemical Society*

1980
Autonomous vehicles drive themselves at the University of Munich, hitting speeds of about 90 km/hour

- Roots of AI
- Representation and reasoning
- Understanding natural language
- Robotics
- Machine learning



PHILOSOPHICAL ORIGINS

The roots of artificial intelligence predate the first computers by centuries. Aristotle described a method of formal, mechanical logic called a syllogism that allows us to draw conclusions from premises. One of his rules sanctioned the following argument: *Some swans are white. All swans are birds. Therefore, some birds are white.* That form of argument – *Some S are W; All S are B; Therefore some B are W* – can be applied to any S, W, and B to arrive at a valid conclusion, regardless of the meaning of the words that make up the sentence. According to this formulation, it is possible to build a mechanism that can act intelligently despite lacking an entire catalogue of human understanding.

Aristotle's proposal set the stage for extensive enquiry into the nature of machine intelligence. It wasn't until the mid-20th century, though, that computers finally became sophisticated enough to test these ideas. In 1948, Grey Walter at the University of Bristol, UK, built a set of autonomous mechanical "turtles" that could move, react to light, and learn. One of these, called Elsie, reacted to her environment for example by decreasing her sensitivity to light as her battery drained. This complex behaviour made her unpredictable, which Walter compared to the behaviour of animals.

In 1950, Alan Turing suggested that if a computer could carry on a conversation with a person, then we should, by "polite convention", agree that the computer "thinks".

But it wasn't until 1956 that the term artificial intelligence was coined. At a summer workshop held at Dartmouth College in Hanover, New Hampshire, the founders of the nascent field laid out their vision: "Every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it."

The expectations had been set for a century of rapid progress. Human-level machine intelligence seemed inevitable.

BACKGROUND: THE KOBAL COLLECTION; LEFT: ALAMY/MOVIESTORE COLLECTION; TOP: BETTMAN/CORBIS

A FIELD OF FRAGMENTS

In the 1960s, most leading artificial intelligence researchers were confident that they would meet their goal within a few decades. After all, aeronautic engineering had gone from the first jet aircraft to an astronaut on the moon in 30 years. Why couldn't AI take off in a similar way?

The difference is that there are no simple formulas for intelligence; the discipline lacks its own $F = ma$ or $E = mc^2$. By the 1980s, AI researchers realised that they had neither sufficient hardware nor knowledge to simulate everything a human can do and the field fragmented. Instead of working towards a single human-equivalent computer intelligence, research groups splintered off to investigate specific aspects of the larger problem: speech recognition, for example, computer vision, probabilistic inference – even chess.

Each of these subdisciplines saw significant successes. In 1997, IBM's Deep Blue computer beat the world chess champion, Garry Kasparov. Deep Blue could evaluate 200 million chess positions per second in its search for the right move. This allowed it to quickly look ahead at many different sequences to see where they might lead.

Deep Blue scored an impressive victory in a game that demands intellectual rigour. However, the machine had a very narrow range of expertise. It could win a game of chess, but it could not discuss the strategy it had employed, nor could it play any other game. No one would mistake its intelligence for human.

1988
Dominant approach to AI is based on probabilistic reasoning using uncertain data (replaces prior focus on logic)

1989
NASA's AutoClass program discovers several previously unknown classes of stars

1997
IBM's Deep Blue supercomputer beats chess champion Garry Kasparov

1998
First AI pet – Furby, produced by Hasbro – is sold in the US

NASA's Remote Agent is first fully autonomous program to control a spacecraft in flight

2000
Nomad robot explores remote regions of Antarctica, looking for meteorite samples

2004
A program quickly learns to pilot a remote-control helicopter better than expert human pilots

2007
AI program at the University of Alberta completely solves the game of draughts/checkers

2011
iRobot sells its 6 millionth Roomba robotic vacuum cleaner

2011
Apple's Siri voice-recognition software lets people converse with their iPhones

2012
Google's statistical machine-translation program does more translations than all professional human translators combined

2014
Google launches its Inception neural network, enabling people to search for photos of specific dog breeds and high-level concepts such as "hugs"

EVERYDAY AI

It might not be obvious, but you interact with AIs every day. They route your phone calls, approve your credit card transactions, prevent fraud and automatically trade stocks in your mutual fund. What's more, they can recognise faces in your digital photos, or your gestures when you play a video game, and even help your doctor interpret test results. But you won't think of these programs as having a human-like intelligence

SPAM HUNTERS

Over 90 per cent of all email sent today is spam. If that were reflected in the contents of your inbox, email would be unusable. Your main protection against the purveyors of miracle pills and work-at-home schemes comes from spam filters that rely on machine learning. As the term implies, the spam filter learns from its environment and from how people treat their emails.

Individual email users provide the gold standard by correctly identifying and labelling messages in their inboxes as "spam" or "not spam". The program uses this information to break down each message into features. A feature can be an individual word, a two-word or multiword phrase, the time of day the message was sent, or the computer that sent it. Those features can then help the program decide whether or not an incoming message is spam. For example, suppose it contains the phrases "lowest prices" and "discreet packaging". The AI will refer to global statistics that tell it that these phrases appear in 8 per cent and 3 per cent of spam, respectively, but only in 0.1 per cent and 0.3 per cent of legitimate messages. After making some assumptions about the independence of features – and applying a formula called Bayes' rule, which assesses the probability of one event happening based on observations of associated events – it concludes that the message is 99.9 per cent likely to be spam.

But the most important thing spam filters do is update their models over time based on experience. Every time a user corrects a mistake, perhaps by rescuing a legitimate message from the junk-mail folder – the system updates itself to reflect the new reality. So programmers do not need to specify step-by-step instructions for identifying a spam message. They need only build a general learning system and expose it to examples of spam and genuine emails. The software does the rest.

"90 per cent of email is spam, but you are protected from the purveyors of pills by a branch of AI called machine learning"

Stanford University's autonomous car, Stanley, won the first driverless contest

DATA MINERS

In 2011, IBM introduced the world to Watson, a question-answering supercomputer with the ability to make sense of questions posed in everyday language, and answer them accurately. Watson's 3000 networked computers were loaded with millions of documents that it could access instantly to answer almost any question.

The company set this computational brawn loose on *Jeopardy!*, an American quiz show famous for posing questions that are far from straightforward. The game is much more complex than chess: *Jeopardy!* requires not only the sum of all human knowledge, but also the ability to understand the puns and wordplay that constitute the game's questions.

The branch of AI primarily responsible for Watson's smarts is probabilistic reasoning: the ability to extract full understanding from a combination of incomplete information. Before the contest began, Watson was loaded with text from encyclopedias, web pages, other reference works and previous *Jeopardy!* games.

IBM then divided the Watson program into a committee of about 100 subprograms, each in charge of its own specialised area, for example, famous authors. After the experts had scoured their own databases, Watson pooled the knowledge of all its component experts and selected the answer with the highest probability of being correct. The machine defeated two human champions.

But *Jeopardy!* championships are not Watson's true calling. IBM plans to spin off the supercomputer's game show success into more serious work, such as providing doctors, businesses, farmers, and others with time-critical information.

BACKGROUND: ALAMY; FAR LEFT: WALTERZERLA/GETTY; LEFT: STANFORD; TOP: BEN HIDER/GETTY

AI HITS THE ROAD

If, on your way to Las Vegas, you pass a car with red licence plates and an infinity symbol, be warned that the car is driving itself. In May 2012, the state of Nevada issued the first licence for an autonomous car.

Will self-driving vehicles catch on in the rest of the world? Until now, driving has been a task best left to humans precisely because it involves so many variables: is the approaching car going at 60 or 70 kilometres per hour? Could there be another vehicle out of sight around the corner? If I attempt to pass the car ahead will its driver speed up? And many more.

For AI, the actual driving has been the easy part, at least on highways. In 1994, two driverless Mercedes-Benz cars fitted with cameras and on-board AI drove 1000 kilometres on highways around Paris.

However, most driving takes place in cities, and that's where it becomes tough for AI, which until recently was unable to negotiate the unwritten rules of city traffic. For example, when Google's researchers programmed an autonomous vehicle to faithfully give way at an intersection as specified in the driver's manual, they found that the self-driving car would often never get a chance to go. So they changed the car's behaviour to make it inch forward after it had waited a while, signalling its intent to move ahead.

Another major source of uncertainty for a self-driving car is locating itself in space. It can't rely solely on GPS, which can be off by several metres, so the AI compensates by simultaneously keeping track of feedback from cameras, radar, and a range-finding laser, crosschecked against GPS data. An average of these imperfect locations provides a highly accurate measurement.

But AI is not restricted to driving. In recent model cars, an AI program automatically adjusts the fuel flow to make it more efficient and the brakes to make them more effective.

Cutting-edge self-driving cars bring together many branches of AI, and people are starting to accept the idea. With special permission, fleets of self-driving Google cars have already negotiated more than a million kilometres on California's highways and busy city streets with no human intervention. Four states, California, Florida, Nevada and Michigan have now legalised autonomous cars, and others may soon follow.

UNIVERSAL TRANSLATOR

On any given day, Google translates more text than all the professional human translators in the world decipher in a year. Google Translate handles 90 languages, translating in either direction between any pair – over 8000 possibilities altogether. It could not do this without a field of AI called probabilistic reasoning.

In the early days, linguists built translation systems based on bilingual dictionaries and codified grammar rules. But these fell short because such rules are inflexible. For example, adjectives come after the noun in French and before the noun in English – except when they don't, as in the phrase "the light fantastic".

In the last decade, translation has shifted from rules that are hand-written by human experts to probabilistic guidelines that are automatically learned from real examples.

Another key aspect of machine translation is the computer-human partnership. Modern machine translation systems start by gathering millions of documents from across the internet that have already been previously translated by humans.

While machine translation is not yet perfect, it is improving at a steady pace as accuracy increases and more languages are added. Google has a translator app called Translate on iPhone and Android. Speak into the phone's microphone, and the app will read back what you've said, translated into your chosen language. The person you are talking to can reply in their own language and Google automatically figures out which language is being spoken.

A NEW FUTURE

More than half a century after the introduction of AI, three key developments could augur the emergence of machine intelligence. New insights from neuroscience and cognitive science are leading to new hardware and software designs. The development of the internet provides access to a vast store of global data. It may even be possible for AI to evolve on its own

BRAINY MACHINES

If we want to build a computer that performs with human-level intellect, why not just copy the brain? Humans, after all, are our best example of intelligence. In the last decade, neuroscience has provided many new insights about how we process and store information.

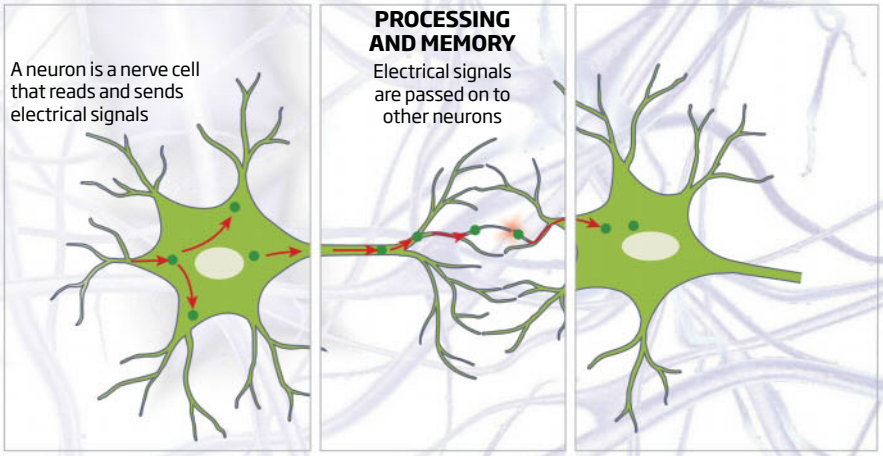
The human brain is a network of 100 trillion synapses that connect 100 billion neurons, most of which change their state between 10 and 100 times per second. Our brain's layout makes us good at tasks like recognising objects in an image.

A supercomputer, on the other hand, has about 100 trillion bytes of memory and its transistors can perform operations 100 million times faster than a brain. This architecture makes a computer better for quickly handling highly defined, precise tasks.

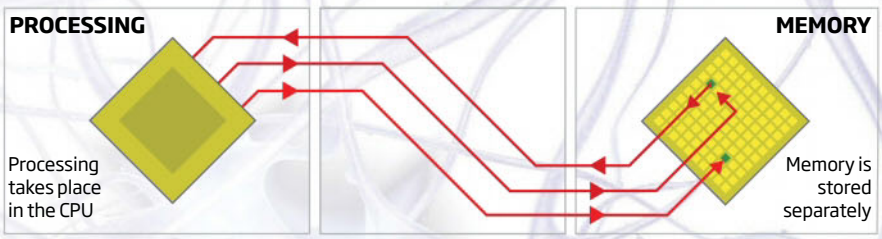
But some tasks would benefit from more brain-like processing, even with the attendant trade-offs. For example, uncertain tasks like recognising faces don't necessarily require highly accurate circuitry in which processing follows a precise path.

Some researchers are looking into brain-like hardware architectures to mimic the brain's low-power requirements. The brain does all of its computations on roughly 20 watts, the equivalent of a very dim light bulb. A supercomputer capable of roughly analogous computations requires 200,000 watts. Other groups are interested in learning from the brain's ability to process information and store it in the same place (see diagram, right). For these reasons, projects are underway to build novel computer circuits inspired by the brain: more parallel rather than serial, more analogue rather than digital, slower, and consuming far less power.

In the mammalian brain, processing and memory storage occur in the same places, making brains better at handling many different tasks at once

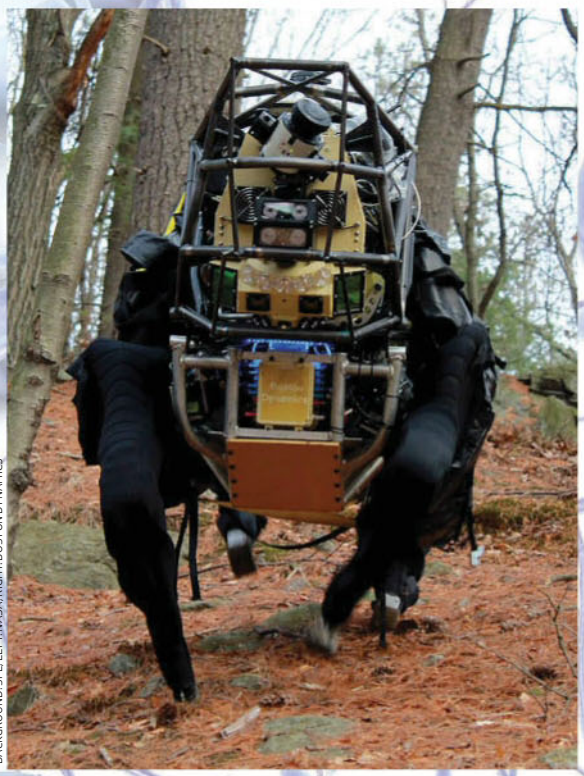


In traditional microchip architecture, memory and processing are separated, limiting speed



New, more brain-like, microchip architectures may lead to chips better at multitasking

"If an autonomous robot tries the same action a few times and fails repeatedly, a 'frustration' circuit would be an effective way of prompting it to explore a new path"



BACKGROUND: SPL; LEFT: NASA; RIGHT: BOSTON DYNAMICS

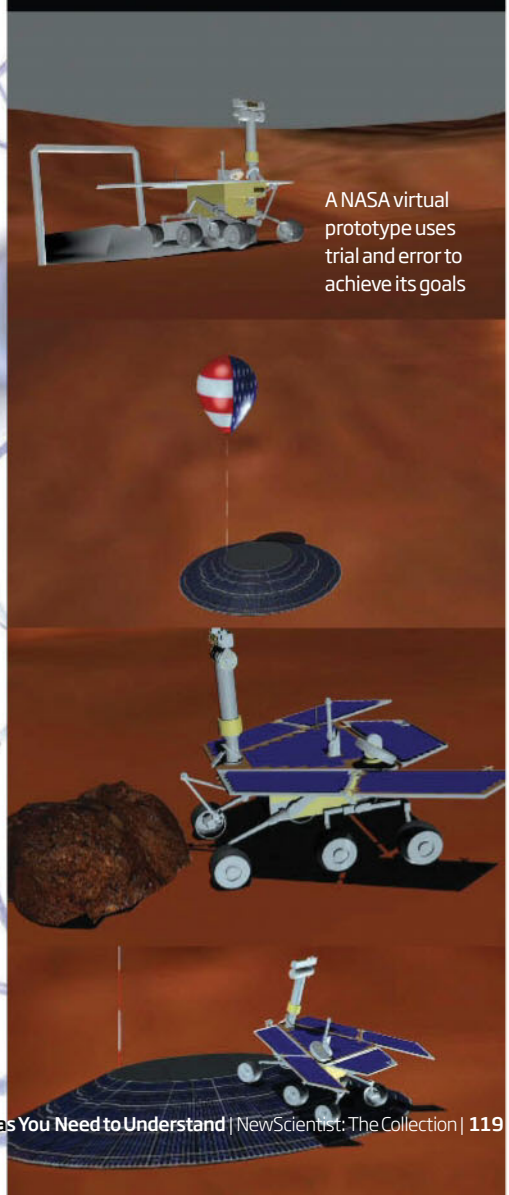
Big Dog uses AI to scramble over rough ground – just don't throw it a stick

EVOLUTION

Most modern AI systems are too complex to program by hand. One alternative is to allow the systems to evolve themselves. In this approach, an iterative process tries out variations of the program in a virtual environment, and chooses the variations that are most successful and make the best decisions, using trial and error.

First, the designers build a simulation of the program's environment – perhaps a desert environment or a pool. Then they set several different AI designs loose inside the simulation, and measure the quality of their decisions – in other words, their fitness. Perhaps one variant of the AI program quickly finds rewards but another variant is slow. Unfit designs are discarded, modifications are made to the fit designs, and the process repeats.

Modifications that change some parameter of a single individual design can be likened to random mutations in natural selection. Modifications that combine two different designs to produce a third "offspring" design are not unlike sexual reproduction. For that reason, they are referred to as genetic algorithms.



A NASA virtual prototype uses trial and error to achieve its goals

COMPUTERS CALL ON INTUITION

Humans persistently fail to live up to the ideal of rationality. We make common errors in our decision-making processes and are easily influenced by irrelevant details. And when we rush to a decision without reasoning through all the evidence, we call this trusting our intuition. We used to think the absence of such human quirks made computers better, but recent research in cognitive science tells us otherwise.

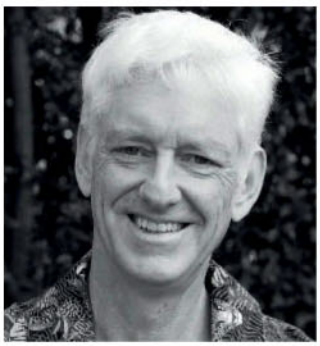
Humans appear to have two complementary decision-making processes, one slow, deliberate and mostly rational, the other fast, impulsive, and able to match the present situation to prior experience, enabling us to reach a quick conclusion. This second mode seems to be key to making human intelligence so effective.

While it is deliberative and sound, the rational part requires more time and energy. Say an oncoming car starts to drift into your lane. You need to act immediately: sound the horn, hit the brakes, or swerve, rather than start a lengthy computation that would determine the optimal but possibly belated act. Such shortcuts are also beneficial when there is no emergency. Expend too much brain power computing the optimal solution to details like whether to wear the dark blue or the midnight blue shirt, and you'll quickly run out of time and energy for the important decisions.

So should AI incorporate an intuitive component? Indeed, many modern AI systems do have two parts, one that reacts instantly to the situation, and one that does more deliberative reasoning. Some robots have been built with a "subsumption" architecture, in which the lowest layers of the system are purely reactive, and higher levels serve to inhibit the reactions and organise more goal-directed behaviour. This approach has proved to be useful, for example, for getting a legged robot to negotiate rough terrain.

There has been a similar push to motivate AIs to make better decisions by giving them emotions. For example, if an autonomous robot tries the same action a few times and fails repeatedly, a "frustration" circuit would be an effective way of prompting it to explore a new path.

Creating machines that simulate emotions is a complicated undertaking. Marvin Minsky, one of the founders of AI, has argued that emotions arise not as a single thing that brains do, but as an interaction involving many parts of the brain, and communication between the brain and the body. Emotions motivate us to choose certain decisions over others, and thinking of the parts of a computer program as if they were motivated by emotions may help us to pave the way for more human-like intelligence.



Peter Norvig

Peter Norvig is director of research at Google Inc. He teaches an open online class at Stanford University in California that provides an introduction to AI. Previously, he directed the Computational Sciences Division at NASA's Ames Research Center, which pioneered the use of autonomous planning, scheduling, and fault-identification on-board the Deep Space 1 spacecraft.

THE COMING SUPERINTELLIGENCE

The notion of the ultra-intelligent machine – one that can surpass human thinking on any subject – was introduced in 1965 by mathematician I. J. Good, who worked with Alan Turing at Bletchley Park, the UK's centre for coding and code-breaking during the second world war. Good noted that "the first ultra-intelligent machine is the last invention that man need ever make", because from then on, the machines would be designing other, ever-better machines, and there would be no work left for humans to do.

In response, some have fearfully predicted that these intelligent machines will dispense with useless humans – mirroring the plot of *The Matrix* – while others see a utopian future filled with endless leisure.

Focusing on these equally unlikely outcomes has distracted the conversation from the very real societal effects already brought about by the increasing pace of technological change. For 100,000 years, we relied on the hard labour of small bands of hunter-gatherers. A scant 200 years ago we moved to an industrial society that shifted most manual labour to machines. And then, just one generation ago, we made the transition into the digital age. Today much of what we manufacture is information, not physical objects – bits, not atoms. Computers are ubiquitous tools, and much of our manual labour has been replaced by calculations.

A similar acceleration is taking place in robotics. The robots you can buy today to vacuum your floor appeal mainly to technophiles. But within a decade there will be an explosion of uses for robots in the office and home. Some will be completely autonomous, others will be tele-operated by a human. Science fiction author Robert Heinlein predicted this development in 1942; remote human operators of his Waldo robots wore gloves that translated their every motion to make a remote robot perform tasks ranging from micro-surgery to maintenance.

Personally, I think that the last invention we need ever make is the partnership of human and tool. Paralleling the move from mainframe computers in the 1970s to personal computers today, most AI systems went from being standalone entities to being tools that are used in a human-machine partnership.

Our tools will get ever better as they embody more intelligence. And we will become better as well, able to access ever more information and education. We may hear less about AI and more about IA, that is to say "intelligence amplification". In movies we will still have to worry about the machines taking over, but in real life humans and their sophisticated tools will move forward together.

RECOMMENDED READING

Artificial Intelligence: A Modern Approach by Stuart Russell and Peter Norvig (Third edition, Prentice Hall, 2009)

"Computing Machinery and Intelligence" by Alan Turing (*Mind*, vol 49, p 433)

Machines Who Think by Pamela McCorduck (First published in 1979. Second edition published in 2004 by A. K. Peters/CRC Press)

The Sciences of the Artificial by Herbert A. Simon (Third edition, 1996, MIT Press)

Cover image

Science Photo Library

QUANTUM INFORMATION

VLATKO VEDRAL

INSTANT EXPERT

PROMISE OF QUANTUM INFORMATION

Processing information in quantum states, rather than in the electrical currents of conventional computer chips, exploits strange quantum effects such as superposition and entanglement. It offers the prospect of peerlessly powerful, economical and secure number crunching. That is the well-developed theory, at least. The challenge is to make it a reality

HISTORY OF AN IDEA

CERN/SCIENCE PHOTO LIBRARY

Physicist Richard Feynman was the first to recognise the potential of quantum computing

The decade or so after physicist Richard Feynman first floated the idea of a quantum computer saw the theory of quantum information bloom.

1981 Feynman argues that modelling the correlations and interactions of particles in complex quantum physics problems can only be tackled by a universal quantum simulator that exploits those same properties.

1982 The no cloning theorem threatens hopes for quantum computing. It states that you cannot copy quantum bits, so there is no way to back up information. The plus side is that this makes intercepting data difficult – a boon for secure transmission of quantum information.

1984 Charles Bennett of IBM and Gilles Brassard of the University of Montreal in Canada develop BB84, the first recipe for secure encoding and transfer of information in quantum states (see "Quantum security", opposite).

1985 David Deutsch at the University of Oxford shows how a universal quantum computer might, in theory, emulate classical logic gates and perform all the functions of quantum logic.

1992 Superdense coding theory shows how a sender and receiver can communicate two classical bits of information by sharing only one entangled pair of quantum states.

1993 In fact you do not need to transmit quantum states at all to exploit their power, as quantum teleportation protocols prove: it is sufficient to possess entangled quantum states and communicate using classical bits.

1994 Shor's algorithm indicates how a quantum computer might factorise numbers faster than any classical computer.

1995 US physicist Benjamin Schumacher coins the term qubit for a quantum bit.

1996 Grover's algorithm gives a recipe by which quantum computers can outperform classical computers in an extremely common task: finding an entry in an unsorted database (see "Number crunching", page 127).

1996 Quantum error correction theory finally overcomes the no-cloning problem. Quantum information cannot be copied – but it can be spread over many qubits.

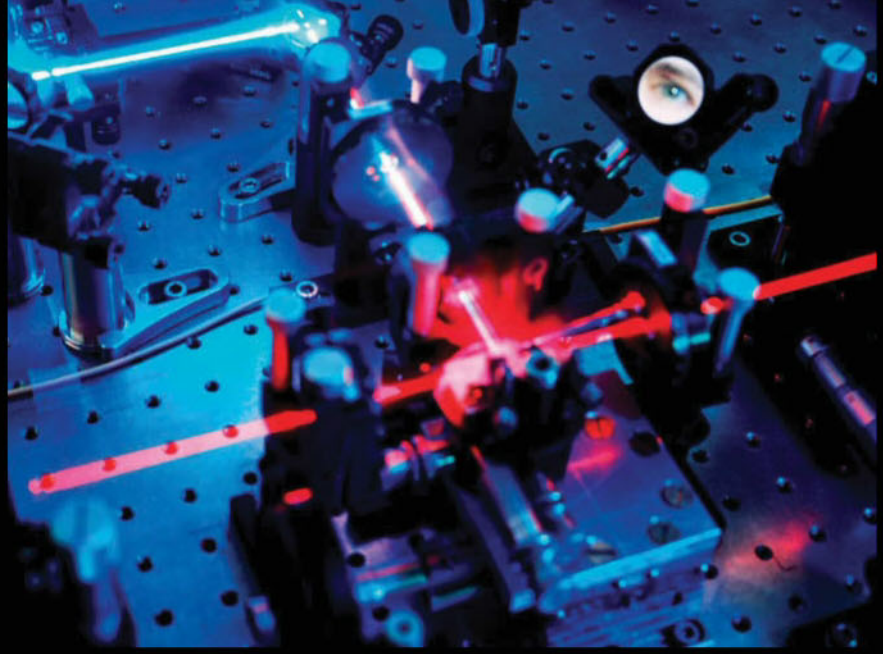
With the problem of copying quantum bits finally solved, the main theoretical tools for quantum information processing were in place. All that remained was to make something practical with them.

QUANTUM POWER

So what makes quantum computers so different? Conventional, classical computers process information using the presence or absence of electrical charge or current. These classical bits have two positions, on (1) and off (0). Semiconductor switches – transistors – flip these bits, making logic gates such as AND, OR and NOT. Combining these gates, we can compute anything that is in principle computable.

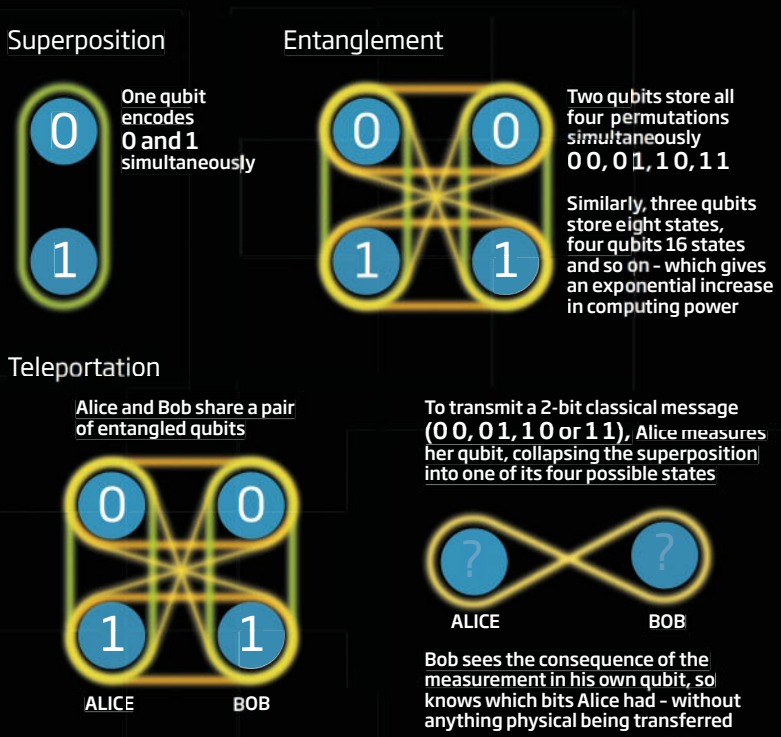
In quantum computation, the switching is between quantum states. Quantum objects can generally be in many states at once: an atom may simultaneously be in different locations or energy states, a photon in more than one state of polarisation, and so on. In general, quantum theory allows us to distinguish two of these states at any one time. In essence, a quantum bit, or qubit, is in a "superposition" that stores 0 and 1 at the same time.

That already suggests an enhanced computational capacity, but the real basis of a quantum computer's power is that the states of many qubits can be "entangled" with one another. This creates a superposition of all the possible combinations of the single-qubit states. Different operations performed on different parts of the superposition at the same time effectively make a massively powerful parallel processor. The power increase is exponential: n qubits have the information processing capacity of 2^n classical bits (see diagram, right). A 400-qubit quantum computer would be like a classical computer with 10^{120} bits – far more than the number of particles estimated to exist in the universe.



Information encoded in polarised laser light can be transmitted entirely securely

Quantum superposition and entanglement combine to allow information to be processed more efficiently and teleported over distances



"Quantum computers could make all problems if not easy, then at least less hard"

QUANTUM ECONOMY

In recent decades, we have done very well in cramming ever more transistors into classical computer chips. But the density of heat generated by constantly resetting these physical on-off switches now represents a fundamental barrier to further miniaturisation. Quantum computation can sidestep that barrier.

That's because by using the right manipulations you can flick between quantum states such as the polarisations of photons without expending any heat. This is no blank cheque for low-power computing, however. Reading and writing information to a quantum memory entails making measurements akin to flipping a classical switch, and will still generate some heat.

QUANTUM SECURITY

Quantum information can be used to make communication totally secure. The trick is that before exchanging an encrypted message by classical means, a sender and receiver (conventionally called Alice and Bob) first share an encryption key imprinted in quantum states.

The first implementation of this "quantum key distribution", the BB84 protocol (see "History of an idea", opposite), proved the principle using an encryption key imprinted on photon polarisation states in superpositions. An attempt by an eavesdropper (Eve) to intercept the key collapses the superpositions. Provided the system has been designed carefully enough, if no such effect is seen, Alice and Bob can be sure only they have the key, which can be used to decrypt the classically exchanged message.

Entanglement-based quantum cryptography, devised by Artur Ekert of the University of Oxford in 1991, relies further on "monogamy of entanglement". In this implementation, Alice and Bob's shared key is made of maximally entangled qubits with the greatest degree of correlation allowed by quantum theory. Eve cannot then entangle any of her own qubits with the key's qubits, and so cannot become party to the information they contain.

QUANTUM LIMITS

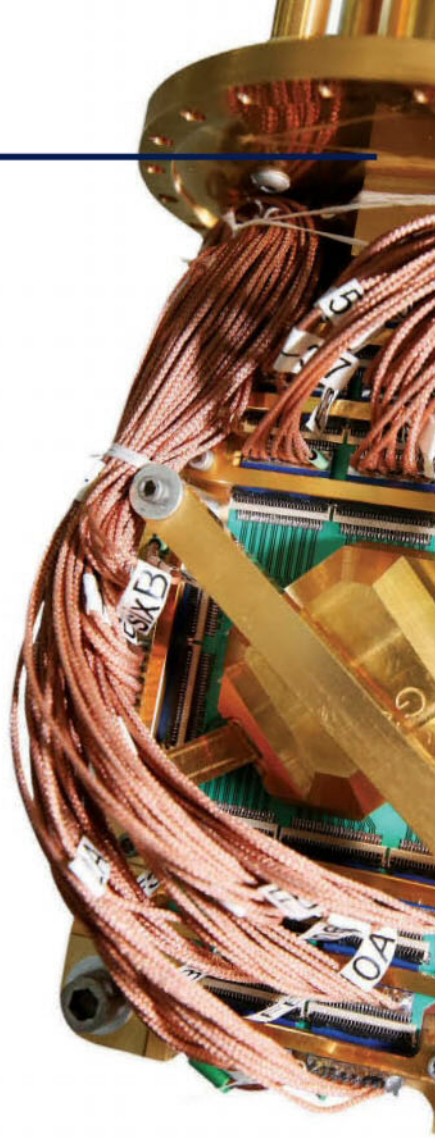
Computer scientists divide problems into "easy" problems, where the computational resources needed to find a solution scale as some power of the number of variables involved, and "hard" problems, where the resources needed increase in a much steeper exponential curve. Hard problems rapidly get beyond the reach of classical computers. The exponentially scaling power of a quantum computer could bring more firepower to bear, if not making the problems easy, then at least less hard.

A quantum speed-up is not a given, however: we first need a specific algorithm that tells us how to achieve it. For important, hard tasks such as factoring large numbers or searching a database, we already have recipes such as Shor's and Grover's algorithms (see "Number crunching", page 127). But for a mundane task such as listing all the entries in a database, the time or processing power required to solve the problem will always scale with the number of entries, and there will be no appreciable quantum speed-up.

All this presumes a quantum computer big enough to make a difference. As we shall see on the following pages, however, making one is easier said than done.

BUILDING A QUANTUM COMPUTER

There are many ways of making the "qubits" for a quantum computer to crunch, from polarising light to cooling atoms to taming the collective motions of electrons. But any qubit must fulfil some stringent criteria, particularly in proving robust, or "coherent", in the face of buffeting from its surrounding classical environment. No single sort of qubit has yet ticked all the boxes



WHAT MAKES FOR A GOOD QUBIT?

In 1997, David DiVincenzo of IBM wrote down some desirable conditions that remain a rough, though not exhaustive, checklist for what any practical quantum computer must achieve.

SCALABILITY To out-gun a classical computer, a quantum computer must entangle and manipulate hundreds of qubits. Quantum computers built so far have just a handful. Scaling up is a big hurdle: the larger the system, the more prone it is to "decohere" in the face of environmental noise, losing its essential quantumness.

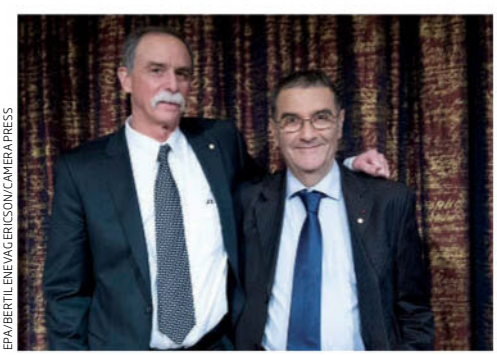
INITIALISATION We must be able to reliably set all the qubits to the same state (to zero, say) at the beginning of a computation.

COHERENCE The time before decoherence kicks in must be a lot longer than the time to switch a quantum logic gate – preferably, several tens of times. In most practical implementations so far this requires an operating temperature near absolute zero to limit the effects of environmental interference.

ACCURACY The results of manipulations must be reproduced accurately by the qubit, even when many manipulations are applied in sequence.

STABLE MEMORY There must be a reliable way to set a qubit's state, keep it in that state, and reset it later.

"To out-gun a classical computer, we must entangle hundreds of qubits. So far we have managed a handful"



EPA/BERTIL ENEVAG ERICSON/CAMERA PRESS

Quantum computing pioneers David Wineland (left) and Serge Haroche shared the 2012 Nobel prize in physics

QUBIT: Photons

The position, polarisation or just number of photons in a given space can be used to encode a qubit. Though initialising their states is easy, photons are slippery: they are easily lost and do not interact very much with each other. That makes them good for communicating quantum information, but to store that information we need to imprint photon states on something longer-lived, such as an atomic energy state.

If we can nail that, quantum computing with photons is a promising concept, not least because the processing can be done at room temperature. In 2012, a team at the University of Vienna, Austria, used four entangled photons to perform the first blind quantum computation. Here a user sends quantum-encoded information to a remote computer that does not itself "see" what it is crunching. This may be a future paradigm – totally secure quantum cloud computing.

RIGHT: Y. COLOMBE/NIST; TOP: IBM RESEARCH

QUBIT: Superconducting states

At temperatures of a few kelvin, electrons in some superconducting materials form entangled pairs, known as Cooper pairs, which flow without resistance and are peculiarly resilient against decoherence.

Superconducting quantum interference devices (SQUIDs) already exploit this effect to make incredibly sensitive measurements of electromagnetic fields. But electron movements and magnetic field states within a SQUID can also be manipulated using external fields to form the bits of a quantum logic device. SQUID qubits offer good initialisation and decoherence times, typically 10 or so times greater than the time taken to switch a gate.

With large numbers of qubits, however, heating due to the external fields used for manipulation becomes an issue, and the largest verified system remains at just three qubits. The company D-Wave Systems in Burnaby, British Columbia, Canada, claimed since 2011 to have a 128-qubit computer in operation (pictured, left); but it remains controversial whether this device is fully quantum, and if it can implement all quantum logic operations. In 2015, the company released what it claims to be a 1152-qubit computer.

QUBIT: Trapped ions

Trapping ions is perhaps the most advanced method of making a quantum computer's qubits. Positively charged ions are caught in electromagnetic fields and cooled to a nanokelvin or so to reduce their vibrations and limit decoherence. Information is then encoded in the ions' energy levels and manipulated using laser light. That brings excellent initialisation success (99.99 per cent), accuracy (over 99 per cent) and stable memory storage (years).

In 1995 David Wineland and his colleagues at the US National Institute of Standards and Technology based in Boulder, Colorado, used trapped ions to create the first quantum logic gate – a controlled NOT (C-NOT) gate for disentangling entangled ions. In 2011, physicists from the University of Innsbruck, Austria, developed a 6-qubit trapped-ion computer that fulfilled the specifications for a universal quantum simulator that Richard Feynman had set out in 1981.

Decoherence and scalability remain interrelated problems, however. With a few entangled qubits the decoherence time is 1000 times the gate-switching time, but this rapidly reduces as qubits are added.

Using lasers to trap ultracooled atomic ions is a well-developed way to make a small-scale quantum computer

BACKGROUND: RITA SCAGLIA/PICTURETANK; TOP: KIM STALLKNECHT/EYEVINE

QUBIT: Cold atoms

Collections of many hundreds of atoms might make for good qubits when trapped, cooled and arranged using lasers in a two-dimensional array known as an optical lattice. The energy states of these atoms can encode information that can be manipulated using further lasers, as with trapped ions (see below left). We've mastered the basic techniques, but making a true quantum computer from cold atoms awaits establishing reliable entanglement among these aloof bodies.

QUBIT: Nuclear spins

Nuclear spin states manipulated using magnetic fields were among the first qubits explored. In 1998, the first implementation of Grover's algorithm (see "Number crunching", page 125) used two nuclear magnetic resonance qubits to seek out one of four elements in a database.

The great advantage of spin states is that they make qubits at room temperature, albeit with a very low initialisation accuracy of about one in a million. But the disrupting effects of thermal noise on entanglement means that nuclear-spin computers are limited to about 20 qubits before their signal becomes washed out.

A variant on the spin theme exploits nitrogen impurities in an otherwise perfect diamond (carbon) lattice. These introduce electrons whose spin can be manipulated electrically, magnetically or with light – but scaling up to anything more than a couple of spins has proved difficult.

QUBIT: Atom-light hybrids

Cavity electrodynamics is a quantum computing approach that aims to combine stable cold atoms with agile photons. Light is trapped inside a micrometre-scale cavity and atoms sent flying through, with logical operations performed through the interactions of the two.

Initialisation is highly efficient, and the decoherence time allows 10 or so gate operations to be performed – although scaling up the technology awaits reliable ways of entangling trapped cold atoms. Serge Haroche of the Collège de France in Paris, one of the pioneers of this approach, shared the 2012 Nobel prize in physics with trapped-ion researcher David Wineland.

QUBIT: Topological states

This promising basis for a quantum computer has yet to get off the theoretical drawing board, because it depends on the existence of particles confined to two dimensions called anyons. These "topological" particles are peculiarly impervious to environmental noise, in principle making them excellent qubits. Particles such as Majorana fermions that fulfil some of the requirements of anyons have been fabricated in certain solids, but whether they are useful for practical quantum computing is still debatable.

KILLER QUANTUM APPS

The theory is in place, and we have no shortage of ideas as to how we can physically implement a quantum computer. But what might we use them for if we did? There are many suggestions – some practical, some highly fanciful

ULTRASECURE ENCODING

One quantum information technology is already up and running. Various small-scale quantum cryptographic systems for secure information transfer, typically using polarised photons as their qubits, have been implemented by labs and companies such as Toshiba, Hewlett Packard, IBM and Mitsubishi. In October 2007, a quantum cryptography system developed by Nicolas Gisin and his colleagues at the University of Geneva in Switzerland was used to transmit votes securely between the city's central polling station and the counting office during the Swiss national elections. A similar trial system developed by the researchers' company, ID Quantique, was used to transmit data securely during the 2010 Football World Cup in South Africa.

The distance over which quantum states can be transmitted along fibre-optic cables is limited to tens of kilometres owing to random diffusion. One promising way to get around this is akin to error correction protocols devised for quantum computers: to spread information over more than one qubit (see "History of an idea", page 122). But this might pose a security risk by giving more information for an eavesdropper to hack.

Transmission via air is an alternative. The world record in faithfully teleporting a qubit of information, held by Anton Zeilinger of the University of Vienna, Austria, and his colleagues, is over a distance of 143 kilometres between the Canary Islands of La Palma and Tenerife. This indicates that delicate quantum states can be transmitted significant distances through air without being disturbed – and suggests that a worldwide secure quantum network using satellites is a distinct possibility.

HARRY GRUYAERT/MAGNUM PHOTOS

Classical encryption methods depend on the difficulty of finding the prime factors of large numbers

QUANTUM SIMULATION

Richard Feynman's original motivation for thinking about quantum computers in 1981 was that they should be more effective than classical computers at simulating quantum systems – including themselves.

This sounds a little underwhelming, but many of science's thorniest practical problems, such as what makes superconductors superconduct or magnets magnetic, are difficult or impossible to solve with classical computers.

Quantum information theorists have already developed intricate algorithms for approximating complex, many-bodied quantum systems, anticipating the arrival of quantum computers powerful enough to deal with them.

The beauty is that such simulators would not be limited to existing physics: we could also use them to glean insights into phenomena not yet seen. Quantum simulations might tell us, say, where best to look in nature for Majorana particles, for example in complex many-bodied superconductor states. Since these particles, thought to be their own antiparticles, have properties that could make them ideally suited to making robust quantum memories (see "Qubit: Topological states", page 125), this raises the intriguing possibility of using quantum computers to suggest more powerful quantum computers.

BENJAMIN BECHET/PICTURETANK

NUMBER CRUNCHING

The promise of quantum computers rests largely on two algorithms. One, developed in 1994 by Peter Shor, then of Bell Laboratories, provides a way for a quantum computer to speedily find the prime factors of large numbers. Classical computers effectively have to try to divide the given number by all possible prime factors (2, 3, 5, 7, 11 and so on) in turn, whereas quantum computers can do these divisions simultaneously.

Conventional encryption methods rely on the fact that classical computers cannot factorise efficiently. If Shor's algorithm were ever implemented on a large scale, encrypted information such as the PIN for your bank card would be vulnerable to hacking – and quantum cryptography would be the only viable defence (see "Ultrasecure encoding", opposite). There is no need to worry just yet: demonstrations so far, for example using a 7-qubit nuclear-spin quantum computer, have been limited to demonstrating that the prime factors of 15 are 5 and 3.

In the longer term, an algorithm devised by physicist Lov Grover in 1996, also at Bell Labs, may become a quantum computer's greatest selling point. This provides a recipe by which a quantum computer can radically speed up how we access and search large bodies of data. Take the example of a database listing the contents of a library. Searching this database for a particular book with a classical computer takes a time that scales with the number of books, n; Grover's algorithm shows that for a quantum computer it scales with $\sqrt{n}$. For a library of a million books, this amounts to 1000 times faster.

Implementing such an algorithm has ubiquitous appeal: almost all computationally hard problems – for instance that of the travelling salesman who has to find the shortest route between a number of different cities – ultimately reduce to a search for the optimal solution. There's a way to go yet. The biggest Grover search yet performed, with 3 qubits, allows for a search of just 8 database elements.

"A quantum computer could search the database of a million-book library 1000 times faster than a classical computer"

METROLOGY

Making precise measurements is a potentially highly significant application of quantum computers. When we record sensitive measurements of physical quantities, such intervals in time or distances in space, the effects of classical noise mean that the best statistical accuracy we can achieve increases with the square-root of the number of bits used to make the recording.

Quantum uncertainty, meanwhile, is determined by the Heisenberg uncertainty principle and improves much more rapidly, simply with the number of measurements made. By encoding distances and time intervals using quantum information – probing them using polarised laser photons, for example – much greater accuracies can be achieved.

This principle is already being applied in giant "interferometers" that use long laser beams in a bid to detect the elusive gravitational waves predicted by Einstein's relativity, such as the LIGO detector in Livingston, Louisiana (pictured, left). In these cases we can think of gravity as noise that disturbs qubits – the qubits being the position and momentum of laser photons. By measuring this disturbance, we can estimate the waves' strength.

BOTTOM: LIGO; TOP: JOERG BUSCHMAN/MILLENIUM

ALICIA CANTER/GUARDIAN NEWS & MEDIA LTD

Vlatko Vedral

Vlatko Vedral is a professor of information theory at the Oxford Martin School of the University of Oxford and at the National University of Singapore. He has published over 200 papers on quantum physics, and is currently focused on bio-inspired quantum-information technologies.

ARE WE NEARLY THERE YET?

No overview of quantum computing would be complete without an attempt to answer the $64,000 (or possibly much more) question: are we likely to see working quantum computers in our homes, offices and hands any time soon?

That depends largely on finding a medium that can encode and process a number of qubits beyond the 10 or 20 that existing technologies can handle. But getting up to the few hundred qubits needed to outperform classical computers is largely a technological issue. Within a couple of decades, given improvements in cooling and trapping, as well as coupling with light, existing technologies of trapped ions and cold atoms may well be made stable enough in large enough quantities to achieve meaningful quantum computation.

The first large-scale quantum computers are likely to be just that: large-scale. They will probably require lasers for qubit manipulation and need supercooling, so are unlikely to appear in our homes. But if the future of much computing is in centralised clouds, perhaps this need not be a problem.

When it comes to anything smaller, the elephant in the room is entanglement, which is a fragile good at the best of times and becomes harder and harder to maintain as the quantum system grows. It would aid the progress of quantum computing if our assumption that entangled states are an essential, central feature turned out to be wrong. This intriguing possibility was raised in 1998 with the development of "single-qubit" algorithms. These can solve a large class of problems, including Shor's factorisation algorithm, without the need for many entangled qubits. That would be a remarkable trick if it could be pulled off in practice – although Grover's all-important database-search algorithm might still not be implementable in this way.

Some people believe that the fragility of quantum systems will never allow us to implement quantum computation in the sort of large, noisy, warm and wet environments in which we humans work. But we can draw hope from recent evidence that living systems – such as photosynthesis in bacteria and retinal systems for magnetic navigation in birds – might be employing some simple quantum information processing to improve their own efficiency.

If we can learn such secrets, a quantum computer on every desktop and in the palm of every hand no longer seems so fanciful an idea.

RECOMMENDED READING

Quantum Theory: Concepts and Methods by Asher Peres (Springer, 1993)

The Fabric of Reality by David Deutsch (Penguin, 1998)

Quantum Computation and Quantum Information by Michael A. Nielsen and Isaac L. Chuang (Cambridge University Press, 2000)

Introduction to Quantum Information Science by Vlatko Vedral (Oxford University Press, 2006)

Programming the Universe by Seth Lloyd (Alfred A. Knopf, 2006)

Decoding Reality by Vlatko Vedral (Oxford University Press, 2010)

Gifted

Give a gift that lasts this Christmas

Subscribe and save over 50%

Visit **newscientist.com/8307** or call
0330 333 9470 and quote 8307

NewScientist


Professor Dame Carol Robinson
2015 Laureate for United Kingdom
By Brigitte Lacombe
For Women in Science
FONDATION L'ORÉAL
Science needs women
L'ORÉAL UNESCO AWARDS
Dame Carol Robinson, Professor of Chemistry at Oxford University, invented a ground-breaking method for studying how membrane proteins function, which play a critical role in the human body.
Throughout the world, exceptional women are at the heart of major scientific advances.
For 17 years, L'Oréal has been running the L'Oréal-UNESCO For Women In Science programme, honouring exceptional women from around the world. Over 2000 women from over 100 countries have received our support to continue to move science forward and inspire future generations.
JOIN US ON FACEBOOK.COM/FORWOMENINSCIENCE