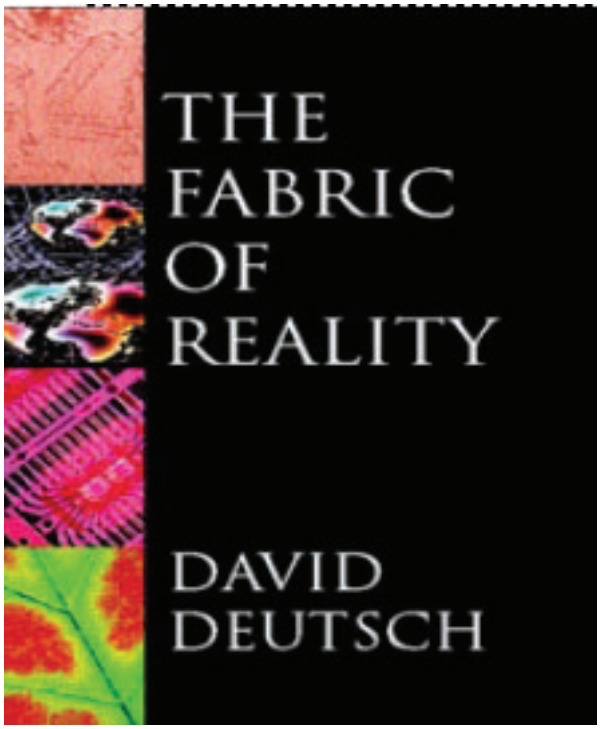

The Fabric of Reality

David Deutch

Deutsch’s pioneering and accessible book integrates recent advances in theoretical physics and computer science to explain and connect many topics at the leading edge of current research and thinking, such as quantum computers, and physics of time travel, and the ultimate fate of the universe.

David Deutch

The Fabric of Reality

PENGUIN BOOKS

THE FABRIC OF REALITY

Born in Haifa, Israel, David Deutsch was educated at Cambridge University and Oxford University. He is a member of the Quantum Computation and Cryptography Research Group at the Clarendon Laboratory, Oxford University. His papers on quantum computation laid the foundations for that field, and he is an authority on the theory of parallel universes.

Dedicated to the memory of Karl Popper, Hugh Everett and Alan Turing, and to Richard Dawkins. This book takes their ideas seriously.

Preface

If there is a single motivation for the world-view set out in this book, it is that thanks largely to a succession of extraordinary scientific discoveries, we now possess some extremely deep theories about the structure of reality. If we are to understand the world on more than a superficial level, it must be through those theories and through reason, and not through our preconceptions, received opinion or even common sense. Our best theories are not only truer than common sense, they make far more sense than common sense does. We must take them seriously, not merely as pragmatic foundations for their respective fields but as explanations of the world. And I believe that we can achieve the greatest understanding if we consider them not singly but jointly, for they are inextricably related.

It may seem odd that this suggestion — that we should try to form a rational and coherent world-view on the basis of our best, most fundamental theories — should be at all novel or controversial. Yet in practice it is. One reason is that each of these theories has, when it is taken seriously, very counter-intuitive implications. Consequently, all sorts of attempts have been made to avoid facing those implications, by making *ad hoc* modifications or reinterpretations of the theories, or by arbitrarily narrowing their domain of applicability, or simply by using them in practice but drawing no wider conclusions from them. I shall criticize some of these attempts (none of which, I believe, has much merit), but only when this happens to be a convenient way of explaining the theories themselves. For this book is not primarily a defence of these theories: it is an investigation of what the fabric of reality would be like if they were true.

1

The Theory of Everything

I remember being told, when I was a small child, that in ancient times it was still possible for a very learned person to know *everything that was known.* I was also told that nowadays so much is known that no one could conceivably learn more than a tiny fraction of it, even in a long lifetime. The latter proposition surprised and disappointed me. In fact, I refused to believe it. I did not know how to justify my disbelief. But I knew that I did not want things to be like that, and I envied the ancient scholars.

It was not that I wanted to memorize all the facts that were listed in the world's encyclopaedias: on the contrary, I hated memorizing facts. That is not the sense in which I expected it to be possible to know everything that was known. It would not have disappointed me to be told that more publications appear every day than anyone could read in a lifetime, or that there are 600,000 known species of beetle. I had no wish to track the fall of every sparrow. Nor did I imagine that an ancient scholar who supposedly knew everything that was known would have known everything of that sort. I had in mind a more discriminating idea of what should count as being known. By 'known', I meant *understood.*

The idea that one person might understand everything that is understood may still seem fantastic, but it is distinctly less fantastic than the idea that one person could memorize every known fact. For example, no one could possibly memorize all known observational data on even so narrow a subject as the motions of the planets, but many astronomers *understand* those motions to the full extent that they are understood. This is possible because understanding does not depend on knowing a lot of facts as such, but on having the right concepts, explanations and theories. One comparatively simple and comprehensible theory can cover an infinity of indigestible facts. Our best theory of planetary motions is Einstein's *general theory of relativity,* which early in the twentieth century superseded Newton's theories of gravity and motion. It correctly predicts, in principle, not only all planetary motions but also all other effects of gravity to the limits of accuracy of our best measurements. For a theory to predict something 'in principle' means that the predictions follow logically from the theory, even if in practice the amount of computation that would be needed to generate some of the predictions is too large to be technologically feasible, or even too large for it to be physically possible for us to carry it out in the universe as we find it.

Being able to predict things or to describe them, however accurately, is not at all the same thing as understanding them. Predictions and descriptions in physics are often expressed as mathematical formulae. Suppose that I memorize the formula from which I could, if I had the time and the inclination, calculate any planetary position that has been recorded in the astronomical archives. What exactly have I gained, compared with memorizing those archives directly? The formula is easier to remember — but then, looking a number up in the archives may be even easier than calculating it from the formula. The real advantage of the formula is that it can be used in an infinity of cases beyond the archived data, for instance to predict the results of future observations. It may also yield the historical positions of the planets more accurately, because the archived data contain observational errors.

Yet even though the formula summarizes infinitely more facts than the archives do, knowing it does not amount to understanding planetary motions. Facts cannot be understood just by being summarized in a formula, any more than by being listed on paper or committed to memory. They can be understood only by being explained. Fortunately, our best theories embody deep explanations as well as accurate predictions. For example, the general theory of relativity explains gravity in terms of a new, four-dimensional geometry of curved space and time. It explains precisely how this geometry affects and is affected by matter. That explanation is the entire content of the theory; predictions about planetary motions are merely some of the consequences that we can deduce from the explanation.

What makes the general theory of relativity so important is not that it can predict planetary motions a shade more accurately than Newton's theory can, but that it reveals and explains previously unsuspected aspects of reality, such as the curvature of space and time. This is typical of scientific explanation. Scientific theories explain the objects and phenomena of our experience in terms of an underlying reality which we do not experience directly. But the ability of a theory to explain what we experience is not its most valuable attribute. Its most valuable attribute is that it explains the fabric of reality itself. As we shall see, one of the most valuable, significant and also useful attributes of human thought generally is its ability to reveal and explain the fabric of reality.

Yet some philosophers — and even some scientists — disparage the role of explanation in science. To them, the basic purpose of a scientific theory is not to explain anything, but to predict the outcomes of experiments: its entire content lies in its predictive formulae. They consider that any consistent explanation that a theory may give for its predictions is as good as any other — or as good as no explanation at all — so long as the predictions are true. This view is called *instrumentalism* (because it says that a theory is no more than an 'instrument' for making predictions). To instrumentalists, the idea that science can enable us to understand the underlying reality that accounts for our observations is a fallacy and a conceit. They do not see how anything a scientific theory may say beyond predicting the outcomes of experiments can be more than empty words. Explanations, in particular, they regard as mere psychological props: a sort of fiction which we incorporate in theories to make them more easily remembered and entertaining. The Nobel prize-winning physicist Steven Weinberg was in instrumentalist mood when he made the following extraordinary comment about Einstein's explanation of gravity:

The important thing is to be able to make predictions about images on the astronomers' photographic plates, frequencies of spectral lines, and so on, and it simply doesn't matter whether we ascribe these predictions to the physical effects of gravitational fields on the motion of planets and photons [as in pre-Einsteinian physics] or to a curvature of space and time. (*Gravitation and Cosmology,* p. 147)

Weinberg and the other instrumentalists are mistaken. What we ascribe the images on astronomers' photographic plates to *does* matter, and it matters not only to theoretical physicists like myself, whose very motivation for formulating and studying theories is the desire to understand the world better. (I am sure that this is Weinberg's motivation too: he is not really

driven by an urge to predict images and spectra!) For even in purely practical applications, the explanatory power of a theory is paramount and its predictive power only supplementary. If this seems surprising, imagine that an extraterrestrial scientist has visited the Earth and given us an ultra-high-technology ‘oracle’ which can predict the outcome of any possible experiment, but provides no explanations. According to instrumentalists, once we had that oracle we should have no further use for scientific theories, except as a means of entertaining ourselves. But is that true? How would the oracle be used in practice? In some sense it would contain the knowledge necessary to build, say, an interstellar spaceship. But how exactly would that help us to build one, or to build another oracle of the same kind — or even a better mousetrap? The oracle only predicts the outcomes of experiments. Therefore, in order to use it at all we must first know what experiments to ask it about. If we gave it the design of a spaceship, and the details of a proposed test flight, it could tell us how the spaceship would perform on such a flight. But it could not design the spaceship for us in the first place. And even if it predicted that the spaceship we had designed would explode on take-off, it could not tell us how to prevent such an explosion. That would still be for us to work out. And before we could work it out, before we could even begin to improve the design in any way, we should have to *understand,* among other things, how the spaceship was supposed to work. Only then would we have any chance of discovering what might cause an explosion on take-off. Prediction — even perfect, universal prediction — is simply no substitute for explanation.

Similarly, in scientific research the oracle would not provide us with any new theory. Not until we already had a theory, and had thought of an experiment that would test it, could we possibly ask the oracle what would happen if the theory were subjected to that test. Thus, the oracle would not be replacing theories at all: it would be replacing experiments. It would spare us the expense of running laboratories and particle accelerators. Instead of building prototype spaceships, and risking the lives of test pilots, we could do all the testing on the ground with pilots sitting in flight simulators whose behaviour was controlled by the predictions of the oracle.

The oracle would be very useful in many situations, but its usefulness would always depend on people’s ability to solve scientific problems in just the way they have to now, namely by devising explanatory theories. It would not even replace all experimentation, because its ability to predict the outcome of a particular experiment would in practice depend on how easy it was to describe the experiment accurately enough for the oracle to give a useful answer, compared with doing the experiment in reality. After all, the oracle would have to have some sort of ‘user interface’. Perhaps a description of the experiment would have to be entered into it, in some standard language. In that language, some experiments would be harder to specify than others. In practice, for many experiments the specification would be too complex to be entered. Thus the oracle would have the same general advantages and disadvantages as any other source of experimental data, and it would be useful only in cases where consulting it happened to be more convenient than using other sources. To put that another way: there already is one such oracle out there, namely the physical world. It tells us the result of any possible experiment if we ask it in the right language (i.e. if we do the

experiment), though in some cases it is impractical for us to ‘enter a description of the experiment’ in the required form (i.e. to build and operate the apparatus). But it provides no explanations.

In a few applications, for instance weather forecasting, we may be almost as satisfied with a purely predictive oracle as with an explanatory theory. But even then, that would be strictly so only if the oracle’s weather forecast were complete and perfect. In practice, weather forecasts are incomplete and imperfect, and to make up for that they include explanations of how the forecasters arrived at their predictions. The explanations allow us to judge the reliability of a forecast and to deduce further predictions relevant to our own location and needs. For instance, it makes a difference to me whether today’s forecast that it will be windy tomorrow is based on an expectation of a nearby high-pressure area, or of a more distant hurricane. I would take more precautions in the latter case. Meteorologists themselves also need explanatory theories about weather so that they can guess what approximations it is safe to incorporate in their computer simulations of the weather, what additional observations would allow the forecast to be more accurate and more timely, and so on.

Thus the instrumentalist ideal epitomized by our imaginary oracle, namely a scientific theory stripped of its explanatory content, would be of strictly limited utility. Let us be thankful that real scientific theories do not resemble that ideal, and that scientists in reality do not work towards that ideal.

An extreme form of instrumentalism, called *positivism* (or logical positivism), holds that all statements other than those describing or predicting observations are not only superfluous but meaningless. Although this doctrine is itself meaningless, according to its own criterion, it was nevertheless the prevailing theory of scientific knowledge during the first half of the twentieth century! Even today, instrumentalist and positivist ideas still have currency. One reason why they are superficially plausible is that, although prediction is not the purpose of science, it is part of the characteristic *method* of science. The scientific method involves postulating a new theory to explain some class of phenomena and then performing a *crucial experimental test,* an experiment for which the old theory predicts one observable outcome and the new theory another. One then rejects the theory whose predictions turn out to be false. Thus the outcome of a crucial experimental test to decide between two theories does depend on the theories’ predictions, and not directly on their explanations. This is the source of the misconception that there is nothing more to a scientific theory than its predictions. But experimental testing is by no means the only process involved in the growth of scientific knowledge. The overwhelming majority of theories are rejected because they contain bad explanations, not because they fail experimental tests. We reject them without ever bothering to test them. For example, consider the theory that eating a kilogram of grass is a cure for the common cold. That theory makes experimentally testable predictions: if people tried the grass cure and found it ineffective, the theory would be proved false. But it has never been tested and probably never will be, because it contains no explanation — either of how the cure would work, or of anything else. We rightly presume it to be false. There are always infinitely many possible theories of that sort, compatible with existing observations and making new predictions, so we could never have the time

or resources to test them all. What we test are new theories that seem to show promise of explaining things better than the prevailing ones do.

To say that prediction is the purpose of a scientific theory is to confuse means with ends. It is like saying that the purpose of a spaceship is to burn fuel. In fact, burning fuel is only one of many things a spaceship has to do to accomplish its real purpose, which is to transport its payload from one point in space to another. Passing experimental tests is only one of many things a theory has to do to achieve the real purpose of science, which is to explain the world.

As I have said, explanations are inevitably framed partly in terms of things we do not observe directly: atoms and forces; the interiors of stars and the rotation of galaxies; the past and the future; the laws of nature. The deeper an explanation is, the more remote from immediate experience are the entities to which it must refer. But these entities are not fictional: on the contrary, they are part of the very fabric of reality.

Explanations often yield predictions, at least in principle. Indeed, if something is, in principle, predictable, then a sufficiently complete explanation must, in principle, make complete predictions (among other things) about it. But many intrinsically unpredictable things can also be explained and understood. For example, you cannot predict what numbers will come up on a fair (i.e. unbiased) roulette wheel. But if you understand what it is in the wheel's design and operation that makes it fair, then you can explain why predicting the numbers is impossible. And again, merely knowing that the wheel is fair is not the same as understanding what makes it fair.

It is understanding, and not mere knowing (or describing or predicting), that I am discussing. Because understanding comes through explanatory theories, and because of the generality that such theories may have, the proliferation of recorded facts does not necessarily make it more difficult to understand everything that is understood. Nevertheless most people would say — and this is in effect what was being said to me on the occasion I recalled from my childhood — that it is not only recorded facts which have been increasing at an overwhelming rate, but also the number and complexity of the theories through which we understand the world. Consequently (they say), whether or not it was ever possible for one person to understand everything that was understood at the time, it is certainly not possible now, and it is becoming less and less possible as our knowledge grows. It might seem that every time a new explanation or technique is discovered that is relevant to a given subject, another theory must be added to the list that anyone wishing to understand that subject must learn; and that when the number of such theories in any one subject becomes too great, specializations develop. Physics, for example, has split into the sciences of astrophysics, thermodynamics, particle physics, quantum field theory, and many others. Each of these is based on a theoretical framework at least as rich as the whole of physics was a hundred years ago, and many are already fragmenting into sub-specializations. The more we discover, it seems, the further and more irrevocably we are propelled into the age of the specialist, and the more remote is that hypothetical ancient time when a single person's understanding might have encompassed all that was understood.

Confronted with this vast and rapidly growing menu of the collected theories of the human race, one may be forgiven for doubting that an individual could so much as taste every dish in a lifetime, let alone, as might once have been possible, appreciate all known recipes. Yet explanation is a strange sort of food — a larger portion is not necessarily harder to swallow. A theory may be superseded by a new theory which explains more, and is more accurate, but is also easier to understand, in which case the old theory becomes redundant, and we gain more understanding while needing to learn less than before. That is what happened when Nicolaus Copernicus's theory of the Earth travelling round the Sun superseded the complex Ptolemaic system which had placed the Earth at the centre of the universe. Or a new theory may be a simplification of an existing one, as when the Arabic (decimal) notation for numbers superseded Roman numerals. (The theory here is an implicit one. Each notation renders certain operations, statements and thoughts about numbers simpler than others, and hence it embodies a theory about which relationships between numbers are useful or interesting.) Or a new theory may be a unification of two old ones, giving us more understanding than using the old ones side by side, as happened when Michael Faraday and James Clerk Maxwell unified the theories of electricity and magnetism into a single theory of electromagnetism. More indirectly, better explanations in any subject tend to improve the techniques, concepts and language with which we are trying to understand other subjects, and so our knowledge as a whole, while increasing, can become structurally more amenable to being understood.

Admittedly, it often happens that even when old theories are thus subsumed into new ones, the old ones are not entirely forgotten. Even Roman numerals are still used today for some purposes. The cumbersome methods by which people once calculated that XIX times XVII equals CCCXXIII are never applied in earnest any more, but they are no doubt still known and understood somewhere — by historians of mathematics for instance. Does this mean that one cannot understand 'everything that is understood' without knowing Roman numerals and their arcane arithmetic? It does not. A modern mathematician who for some reason had never heard of Roman numerals would nevertheless already possess in full the understanding of their associated mathematics. By learning about Roman numerals, that mathematician would be acquiring no new understanding, only new facts — historical facts, and facts about the properties of certain arbitrarily defined symbols, rather than new knowledge about numbers themselves. It would be like a zoologist learning to translate the names of species into a foreign language, or an astrophysicist learning how different cultures group stars into constellations.

It is a separate issue whether knowing the arithmetic of Roman numerals might be necessary in the understanding of *history.* Suppose that some historical theory — some explanation — depended on the specific techniques used by the ancient Romans for multiplication (rather as, for instance, it has been conjectured that their specific plumbing techniques, based on lead pipes, which poisoned their drinking water, contributed to the decline of the Roman Empire). Then we should have to know what those techniques were if we wanted to understand history, and therefore also if we wanted to understand everything that is understood. But in the event, no

current explanation of history draws upon multiplication techniques, so our records of those techniques are mere statements of facts. Everything that is understood can be understood without learning those facts. We can always look them up when, for instance, we are deciphering an ancient text that mentions them.

In continually drawing a distinction between understanding and ‘mere’ knowing, I do not want to understate the importance of recorded, non-explanatory information. This is of course essential to everything from the reproduction of a micro-organism (which has such information in its DNA molecules) to the most abstract human thinking. So what distinguishes understanding from mere knowing? What is an explanation, as opposed to a mere statement of fact such as a correct description or prediction? In practice, we usually recognize the difference easily enough. We know when we do not understand something, even if we can accurately describe and predict it (for instance, the course of a known disease of unknown origin), and we know when an explanation helps us to understand it better. But it is hard to give a precise definition of ‘explanation’ or ‘understanding’. Roughly speaking, they are about ‘why’ rather than ‘what’; about the inner workings of things; about how things really are, not just how they appear to be; about what must be so, rather than what merely happens to be so; about laws of nature rather than rules of thumb. They are also about coherence, elegance and simplicity, as opposed to arbitrariness and complexity, though none of those things is easy to define either. But in any case, understanding is one of the higher functions of the human mind and brain, and a unique one. Many other physical systems, such as animals’ brains, computers and other machines, can assimilate facts and act upon them. But at present we know of nothing that is capable of understanding an explanation — or of wanting one in the first place — other than a human mind. Every discovery of a new explanation, and every act of grasping an existing explanation, depends on the uniquely human faculty of creative thought.

One can think of what happened to Roman numerals as a process of ‘demotion’ of an explanatory theory to a mere description of facts. Such demotions happen all the time as our knowledge grows. Originally, the Roman system of numerals did form part of the conceptual and theoretical framework through which the people who used them understood the world. But now the understanding that used to be obtained in that way is but a tiny facet of the far deeper understanding embodied in modern mathematical theories, and implicitly in modern notations.

This illustrates another attribute of understanding. It is possible to understand something without knowing that one understands it, or even without having specifically heard of it. This may sound paradoxical, but of course the whole point of deep, general explanations is that they cover unfamiliar situations as well as familiar ones. If you were a modern mathematician encountering Roman numerals for the first time, you might not instantly realize that you already understood them. You would first have to learn the facts about what they are, and then think about those facts in the light of your existing understanding of mathematics. But once you had done that, you would be able to say, in retrospect, ‘Yes, there is nothing new to me in the Roman system of numerals, beyond mere facts.’ And that is what it means to say that Roman numerals, in their explanatory role, are fully

obsolete.

Similarly, when I say that I understand how the curvature of space and time affects the motions of planets, even in other solar systems I may never have heard of, I am not claiming that I can call to mind, without further thought, the explanation of every detail of the loops and wobbles of any planetary orbit. What I mean is that I understand the theory that contains all those explanations, and that I could therefore produce any of them in due course, given some facts about a particular planet. Having done so, I should be able to say in retrospect, 'Yes, I see nothing in the motion of that planet, other than mere facts, which is not explained by the general theory of relativity.' We understand the fabric of reality only by understanding theories that explain it. And since they explain more than we are immediately aware of, we can understand more than we are immediately aware that we understand.

I am not saying that when we understand a theory it *necessarily* follows that we understand everything it can explain. With a very deep theory, the recognition that it explains a given phenomenon may itself be a significant discovery requiring independent explanation. For example, quasars — extremely bright sources of radiation at the centre of some galaxies — were for many years one of the mysteries of astrophysics. It was once thought that new physics would be needed to explain them, but now we believe that they are explained by the general theory of relativity and other theories that were already known before quasars were discovered. We believe that quasars consist of hot matter in the process of falling into black holes (collapsed stars whose gravitational field is so intense that nothing can escape from them). Yet reaching that conclusion has required years of research, both observational and theoretical. Now that we believe we have gained a measure of understanding of quasars, we do not think that this understanding is something we already had before. Explaining quasars, albeit through existing theories, has given us genuinely new understanding. Just as it is hard to define what an explanation is, it is hard to define when a subsidiary explanation should count as an independent component of what is understood, and when it should be considered as being subsumed in the deeper theory. It is hard to define, but not so hard to recognize: as with explanations in general, in practice we know a new explanation when we are given one. Again, the difference has something to do with creativity. Explaining the motion of a particular planet, when one already understands the general explanation of gravity, is a mechanical task, though it may be a very complex one. But using existing theory to account for quasars requires creative thought. Thus, to understand everything that is understood in astrophysics today, you would have to know the theory of quasars explicitly. But you would not have to know the orbit of any specific planet.

So, even though our stock of known theories is indeed snowballing, just as our stock of recorded facts is, that still does not necessarily make the whole structure harder to understand than it used to be. For while our specific theories are becoming more numerous and more detailed, they are continually being 'demoted' as the understanding they contain is taken over by deep, general theories. And those theories are becoming fewer, deeper and more general. By 'more general' I mean that each of them says more, about a wider range of situations, than several distinct theories did

previously. By ‘deeper’ I mean that each of them explains more — embodies more understanding — than its predecessors did, combined.

Centuries ago, if you had wanted to build a large structure such as a bridge or a cathedral you would have engaged a master builder. He would have had some knowledge of what it takes to give a structure strength and stability with the least possible expense and effort. He would not have been able to express much of this knowledge in the language of mathematics and physics, as we can today. Instead, he relied mainly on a complex collection of intuitions, habits and rules of thumb, which he had learned from his apprentice-master and then perhaps amended through guesswork and long experience. Even so, these intuitions, habits and rules of thumb were in effect *theories,* explicit and inexplicit, and they contained real knowledge of the subjects we nowadays call engineering and architecture. It was for the knowledge in those theories that you would have hired him, pitifully inaccurate though it was compared with what we have today, and of very narrow applicability. When admiring centuries-old structures, people often forget that we see only the surviving ones. The overwhelming majority of structures built in medieval and earlier times have collapsed long ago, often soon after they were built. That was especially so for innovative structures. It was taken for granted that innovation risked catastrophe, and builders seldom deviated much from designs and techniques that had been validated by long tradition. Nowadays, in contrast, it is quite rare for any structure — even one that is unlike anything that has ever been built before — to fail because of faulty design. Anything that an ancient master builder could have built, his modern colleagues can build better and with far less human effort. They can also build structures which he could hardly have dreamt of, such as skyscrapers and space stations. They can use materials which he had never heard of, such as fibreglass or reinforced concrete, and which he could hardly have used even if he could somehow have been given them, for he had only a scanty and inaccurate understanding of how materials work.

Progress to our current state of knowledge was not achieved by accumulating more theories of the same kind as the master builder knew. Our knowledge, both explicit and inexplicit, is not only much greater than his but structurally different too. As I have said, the modern theories are fewer, more general and deeper. For each situation that the master builder faced while building something in his repertoire — say, when deciding how thick to make a load-bearing wall — he had a fairly specific intuition or rule of thumb, which, however, could give hopelessly wrong answers if applied to novel situations. Today one deduces such things from a theory that is general enough for it to be applied to walls made of any material, in all situations: on the Moon, underwater, or wherever. The reason why it is so general is that it is based on quite deep explanations of how materials and structures work. To find the proper thickness of a wall that is to be made from an unfamiliar material, one uses the same theory as for any other wall, but starts the calculation by assuming different facts — by using different numerical values for the various parameters. One has to look up those facts, such as the tensile strength and elasticity of the material, but one needs no additional understanding.

That is why, despite understanding incomparably more than an ancient master builder did, a modern architect does not require a longer or more

arduous training. A typical theory in a modern student's syllabus may be harder to understand than any of the master builder's rules of thumb; but the modern theories are far fewer, and their explanatory power gives them other properties such as beauty, inner logic and connections with other subjects which make them easier to learn. Some of the ancient rules of thumb are now known to be erroneous, while others are known to be true, or to be good approximations to the truth, and we know why that is so. A few are still in use. But none of them is any longer the source of anyone's understanding of what makes structures stand up.

I am not, of course, denying that specialization is occurring in many subjects in which knowledge is growing, including architecture. This is not a one-way process, for specializations often disappear too: wheels are no longer designed or made by wheelwrights, nor ploughs by ploughwrights, nor are letters written by scribes. It is nevertheless quite evident that the deepening, unifying tendency I have been describing is not the only one at work: a continual *broadening* is going on at the same time. That is, new ideas often do more than just supersede, simplify or unify existing ones. They also extend human understanding into areas that were previously not understood at all — or whose very existence was not guessed at. They may open up new opportunities, new problems, new specializations and even new subjects. And when that happens it may give us, at least temporarily, more to learn in order to understand it all.

The science of medicine is perhaps the most frequently cited case of increasing specialization seeming to follow inevitably from increasing knowledge, as new cures and better treatments for more diseases are discovered. But even in medicine the opposite, unifying tendency is also present, and is becoming stronger. Admittedly, many functions of the body are still poorly understood, and so are the mechanisms of many diseases. Consequently some areas of medical knowledge still consist mainly of collections of recorded facts, together with the skills and intuitions of doctors who have experience of particular diseases and particular treatments, and who pass on these skills and intuitions from one generation to the next. Much of medicine, in other words, is still in the rule-of-thumb era, and when new rules of thumb are discovered there is indeed more incentive for specialization. But as medical and biochemical research comes up with deeper explanations of disease processes (and healthy processes) in the body, understanding is also on the increase. More general concepts are replacing more specific ones as common, underlying molecular mechanisms are found for dissimilar diseases in different parts of the body. Once a disease can be understood as fitting into a general framework, the role of the specialist diminishes. Instead, physicians coming across an unfamiliar disease or a rare complication can rely increasingly on explanatory theories. They can look up such facts as are known. But then they may be able to apply a general theory to work out the required treatment, and expect it to be effective even if it has never been used before.

Thus the issue of whether it is becoming harder or easier to understand everything that is understood depends on the overall balance between these two opposing effects of the growth of knowledge: the increasing *breadth* of our theories, and their increasing *depth.* Breadth makes it harder; depth makes it easier. One thesis of this book is that, slowly but surely, depth is

winning. In other words, the proposition that I refused to believe as a child is indeed false, and practically the opposite is true. We are not heading away from a state in which one person could understand everything that is understood, but towards it.

It is not that we shall soon understand *everything.* That is a completely different issue. I do not believe that we are now, or ever shall be, close to understanding *everything there is.* What I am discussing is the possibility of understanding *everything that is understood.* That depends more on the structure of our knowledge than on its content. But of course the structure of our knowledge — whether it is expressible in theories that fit together as a comprehensible whole — does depend on what the fabric of reality, as a whole, is like. If knowledge is to continue its open-ended growth, and if we are nevertheless heading towards a state in which one person could understand everything that is understood, then the depth of our theories must continue to grow fast enough to make this possible. That can happen only if the fabric of reality is itself highly unified, so that more and more of it can become understood as our knowledge grows. If that happens, then eventually our theories will become so general, deep and integrated with one another that they will effectively become a single theory of a unified fabric of reality. This theory will still not explain every aspect of reality: that is unattainable. But it will encompass all known explanations, and will apply to the whole fabric of reality in so far as it is understood. Whereas all previous theories related to particular subjects, this will be a theory of all subjects: a *Theory of Everything.*

It will not, of course, be the last such theory, only the first. In science we take it for granted that even our best theories are bound to be imperfect and problematic in some ways, and we expect them to be superseded in due course by deeper, more accurate theories. Such progress is not brought to a halt when we discover a universal theory. For example, Newton gave us the first universal theory of gravity and a unification of, among other things, celestial and terrestrial mechanics. But his theories have been superseded by Einstein's general theory of relativity which additionally incorporates geometry (formerly regarded as a branch of mathematics) into physics, and in so doing provides far deeper explanations as well as being more accurate. The first fully universal theory — which I shall call the Theory of Everything — will, like all our theories before and after it, be neither perfectly true nor infinitely deep, and so will eventually be superseded. But it will not be superseded through unifications with theories about other subjects, for it will already be a theory of all subjects. In the past, some great advances in understanding came about through great unifications. Others came through structural changes in the way we were understanding a particular subject — as when we ceased to think of the Earth as being the centre of the universe. After the first Theory of Everything, there will be no more great unifications. All subsequent great discoveries will take the form of changes in the way we understand the world as a whole: shifts in our world-view. The attainment of a Theory of Everything will be the last great unification, and at the same time it will be the first across-the-board shift to a new world-view. I believe that such a unification and shift are now under way. The associated world-view is the theme of this book. I must stress immediately that I am not referring merely to the 'theory of everything' which some particle physicists hope they

will soon discover. *Their* 'theory of everything' would be a unified theory of all the basic forces known to physics, namely gravity, electromagnetism and nuclear forces. It would also describe all the types of subatomic particles that exist, their masses, spins, electric charges and other properties, and how they interact. Given a sufficiently precise description of the initial state of any isolated physical system, it would in principle predict the future behaviour of the system. Where the exact behaviour of a system was intrinsically unpredictable, it would describe all possible behaviours and predict their probabilities. In practice, the initial states of interesting systems often cannot be ascertained very accurately, and in any case the calculation of the predictions would be too complicated to be carried out in all but the simplest cases. Nevertheless, such a unified theory of particles and forces, together with a specification of the initial state of the universe at the Big Bang (the violent explosion with which the universe began), would in principle contain all the information necessary to predict everything that can be predicted (Figure 1.1).

But prediction is not explanation. The hoped-for 'theory of everything', even if combined with a theory of the initial state, will at best provide only a tiny facet of a real Theory of Everything. It may *predict* everything (in principle). But it cannot be expected to *explain* much more than existing theories do, except for a few phenomena that are dominated by the nuances of subatomic interactions, such as collisions inside particle accelerators, and the exotic history of particle transmutations in the Big Bang. What motivates the use of the term 'theory of everything' for such a narrow, albeit fascinating, piece of knowledge? It is, I think, another mistaken view of the nature of science, held disapprovingly by many critics of science and (alas) approvingly by many scientists, namely that science is essentially *reductionist.* That is to say, science allegedly explains things reductively — by analysing them into components. For example, the resistance of a wall to being penetrated or knocked down is explained by regarding the wall as a vast aggregation of interacting molecules. The properties of those molecules are themselves explained in terms of their constituent atoms, and the interactions of these atoms with one another, and so on down to the smallest particles and most basic forces. Reductionists think that all scientific explanations, and perhaps all sufficiently deep explanations of any kind, take that form.

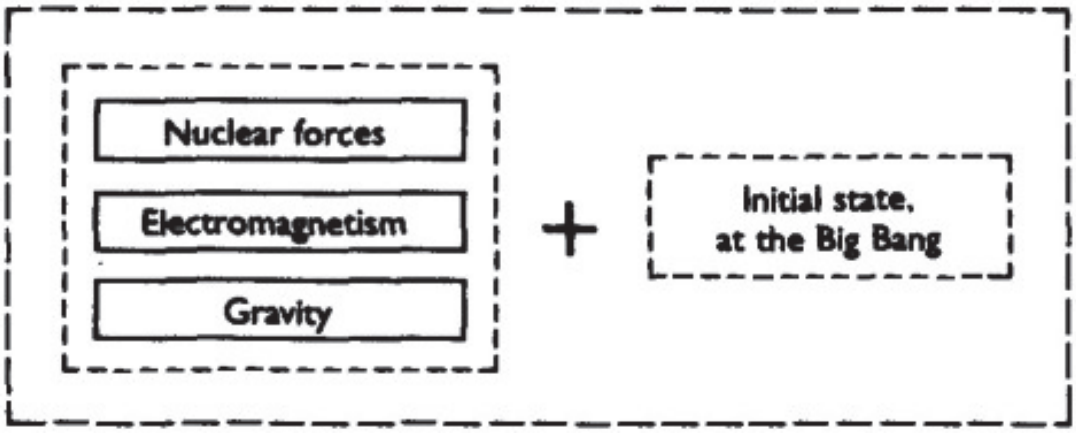

Figure 1.1. *An inadequate conception of the 'theory of everything'.*

The reductionist conception leads naturally to a classification of objects and theories in a hierarchy, according to how close they are to the 'lowest-level' predictive theories that are known. In this hierarchy, logic and mathematics form the immovable bedrock on which the edifice of science is built. The foundation stone would be a reductive 'theory of everything', a universal theory of particles, forces, space and time, together with some theory of what the initial state of the universe was. The rest of physics forms the first few storeys. Astrophysics and chemistry are at a higher level, geology even higher, and so on. The edifice branches into many towers of increasingly high-level subjects like biochemistry, biology and genetics. Perched at the tottering, stratospheric tops are subjects like the theory of evolution, economics, psychology and computer science, which in this picture are almost inconceivably derivative. At present, we have only approximations to a reductive 'theory of everything'. These can already predict quite accurate laws of motion for individual subatomic particles. From these laws, present-day computers can calculate the motion of any isolated group of a few interacting particles in some detail, given their initial state. But even the smallest speck of matter visible to the naked eye contains trillions of atoms, each composed of many subatomic particles, and is continually interacting with the outside world; so it is quite infeasible to predict its behaviour particle by particle. By supplementing the exact laws of motion with various approximation schemes, we can predict some aspects of the gross behaviour of quite large objects — for instance, the temperature at which a given chemical compound will melt or boil. Much of basic chemistry has been reduced to physics in this way. But for higher-level sciences the reductionist programme is a matter of principle only. No one expects actually to deduce many principles of biology, psychology or politics from those of physics. The reason why higher-level subjects can be studied at all is that under special circumstances the stupendously complex behaviour of vast numbers of particles resolves itself into a measure of simplicity and comprehensibility. This is called *emergence:* high-level simplicity 'emerges' from low-level complexity. High-level phenomena about which there are comprehensible facts that are not simply deducible from lower-level theories are called *emergent phenomena.* For example, a wall might be strong because its builders feared that their enemies might try to force their way through it. This is a high-level explanation of the wall's strength, not deducible from (though not incompatible with) the low-level explanation I gave above. 'Builders', 'enemies', 'fear' and 'trying' are all emergent phenomena. The purpose of high-level sciences is to enable us to understand emergent phenomena, of which the most important are, as we shall see, *life, thought* and *computation.*

By the way, the opposite of reductionism, *holism* — the idea that the only legitimate explanations are in terms of higher-level systems — is an even greater error than reductionism. What do holists expect us to do? Cease our search for the molecular origin of diseases? Deny that human beings are made of subatomic particles? Where reductive explanations exist, they are just as desirable as any other explanations. Where whole sciences are reducible to lower-level sciences, it is just as incumbent upon us as scientists to find those reductions as it is to discover any other knowledge.

A reductionist thinks that science is about analysing things into components. An instrumentalist thinks that it is about predicting things. To either of them, the existence of high-level sciences is merely a matter of convenience. Complexity prevents us from using fundamental physics to make high-level predictions, so instead we guess what those predictions would be if we could make them — emergence gives us a chance of doing that successfully — and supposedly that is what the higher-level sciences are about. Thus to reductionists and instrumentalists, who disregard both the real structure and the real purpose of scientific knowledge, the base of the predictive hierarchy of physics is by definition the 'theory of everything'. But to everyone else scientific knowledge consists of explanations, and the structure of scientific explanation does not reflect the reductionist hierarchy. There are explanations at every level of the hierarchy. Many of them are autonomous, referring only to concepts at that particular level (for instance, 'the bear ate the honey because it was hungry'). Many involve deductions in the opposite direction to that of reductive explanation. That is, they explain things not by analysing them into smaller, simpler things but by regarding them as components of larger, more complex things — about which we nevertheless have explanatory theories. For example, consider one particular copper atom at the tip of the nose of the statue of Sir Winston Churchill that stands in Parliament Square in London. Let me try to explain why that copper atom is there. It is because Churchill served as prime minister in the House of Commons nearby; and because his ideas and leadership contributed to the Allied victory in the Second World War; and because it is customary to honour such people by putting up statues of them; and because bronze, a traditional material for such statues, contains copper, and so on. Thus we explain a low-level physical observation — the presence of a copper atom at a particular location — through extremely high-level theories about emergent phenomena such as ideas, leadership, war and tradition. There is no reason why there should exist, even in principle, any lower-level *explanation* of the presence of that copper atom than the one I have just given. Presumably a reductive 'theory of everything' would in principle make a low-level *prediction* of the probability that such a statue will exist, given the condition of (say) the solar system at some earlier date. It would also in principle describe how the statue probably got there. But such descriptions and predictions (wildly infeasible, of course) would explain nothing. They would merely describe the trajectory that each copper atom followed from the copper mine, through the smelter and the sculptor's studio, and so on. They could also state how those trajectories were influenced by forces exerted by surrounding atoms, such as those comprising the miners' and sculptor's bodies, and so predict the existence and shape of the statue. In fact such a prediction would have to refer to atoms all over the planet, engaged in the complex motion we call the Second World War, among other things. But even if you had the superhuman capacity to follow such lengthy predictions of the copper atom's being there, you would still not be able to say, 'Ah yes, now I understand why it is there.' You would merely know that its arrival there in that way was inevitable (or likely, or whatever), given all the atoms' initial configurations and the laws of physics. If you wanted to understand why, you would still have no option but to take a further step. You would have to inquire into what it was about that configuration of atoms, and those trajectories, that gave them the propensity to deposit a copper atom at this location. Pursuing

this inquiry would be a creative task, as discovering new explanations always is. You would have to discover that certain atomic configurations support emergent phenomena such as leadership and war, which are related to one another by high-level explanatory theories. Only when you knew those theories could you understand fully why that copper atom is where it is.

In the reductionist world-view, the laws governing subatomic particle interactions are of paramount importance, as they are the base of the hierarchy of all knowledge. But in the real structure of scientific knowledge, and in the structure of our knowledge generally, such laws have a much more humble role.

What is that role? It seems to me that none of the candidates for a 'theory of everything' that has yet been contemplated contains much that is new by way of explanation. Perhaps the most innovative approach from the explanatory point of view is *superstring theory,* in which extended objects, 'strings', rather than point-like particles, are the elementary building blocks of matter. But no existing approach offers an entirely new mode of explanation — new in the sense of Einstein's explanation of gravitational forces in terms of curved space and time. In fact, the 'theory of everything' is expected to inherit virtually its entire explanatory structure — its physical concepts, its language, its mathematical formalism and the form of its explanations — from the existing theories of electromagnetism, nuclear forces and gravity. Therefore we may look to this underlying structure, which we already know from existing theories, for the contribution of fundamental physics to our overall understanding.

There are two theories in physics which are considerably deeper than all others. The first is the general theory of relativity, which as I have said is our best theory of space, time and gravity. The second, *quantum theory,* is even deeper. Between them, these two theories (and not any existing or currently envisaged theory of subatomic particles) provide the detailed explanatory and formal framework within which all other theories in modern physics are expressed, and they contain overarching physical principles to which all other theories conform. A unification of general relativity and quantum theory — to give a *quantum theory of gravity* — has been a major quest of theoretical physicists for several decades, and would have to form part of any theory of everything in either the narrow or the broad sense of the term. As we shall see in the next chapter, quantum theory, like relativity, provides a revolutionary new mode of explanation of physical reality. The reason why quantum theory is the deeper of the two lies more outside physics than within it, for its ramifications are very wide, extending far beyond physics — and even beyond science itself as it is normally conceived. Quantum theory is one of what I shall call the *four main strands* of which our current understanding of the fabric of reality is composed.

Before I say what the other three strands are, I must mention another way in which reductionism misrepresents the structure of scientific knowledge. Not only does it assume that explanation always consists of analysing a system into smaller, simpler systems, it also assumes that all explanation is of later events in terms of earlier events; in other words, that the only way of explaining something is to state its *causes.* And this implies that the earlier the events in terms of which we explain something, the better the

explanation, so that ultimately the best explanations of all are in terms of the initial state of the universe.

A ‘theory of everything’ which excludes a specification of the initial state of the universe is not a complete description of physical reality because it provides only laws of motion; and laws of motion, by themselves, make only conditional predictions. That is, they never state categorically what happens, but only what will happen at one time given what was happening at another time. Only if a complete specification of the initial state is provided can a complete description of physical reality in principle be deduced. Current cosmological theories do not provide a complete specification of the initial state, even in principle, but they do say that the universe was initially very small, very hot and very uniform in structure. We also know that it cannot have been perfectly uniform because that would be incompatible, according to the theory, with the distribution of galaxies we observe across the sky today. The initial variations in density, ‘lumpiness’, would have been greatly enhanced by gravitational clumping (that is, relatively dense regions would have attracted more matter and become denser), so they need only have been very slight initially. But, slight though they were, they are of the greatest significance in any reductionist description of reality, because almost everything that we see happening around us, from the distribution of stars and galaxies in the sky to the appearance of bronze statues on planet Earth, is, from the point of view of fundamental physics, a consequence of those variations. If our reductionist description is to cover anything more than the grossest features of the observed universe, we need a theory specifying those all-important initial deviations from uniformity.

Let me try to restate this requirement without the reductionist bias. The laws of motion for any physical system make only conditional predictions, and are therefore compatible with many possible histories of that system. (This issue is independent of the limitations on predictability that are imposed by quantum theory, which I shall discuss in the next chapter.) For instance, the laws of motion governing a cannon-ball fired from a gun are compatible with many possible trajectories, one for every possible direction and elevation in which the gun could have been pointing when it was fired (Figure 1.2). Mathematically, the laws of motion can be expressed as a set of equations called the *equations of motion.* These have many different solutions, one describing each possible trajectory. To specify which solution describes the actual trajectory, we must provide *supplementary data* — some data about what actually happens. One way of doing that is to specify the initial state, in this case the direction in which the gun was pointing. But there are other ways too. For example, we could just as well specify the final state — the position and direction of motion of the cannon-ball at the moment it lands. Or we could specify the position of the highest point of the trajectory. It does not matter what supplementary data we give, so long as they pick out one particular solution of the equations of motion. The combination of any such supplementary data with the laws of motion amounts to a theory that describes everything that happens to the cannon-ball between firing and impact.

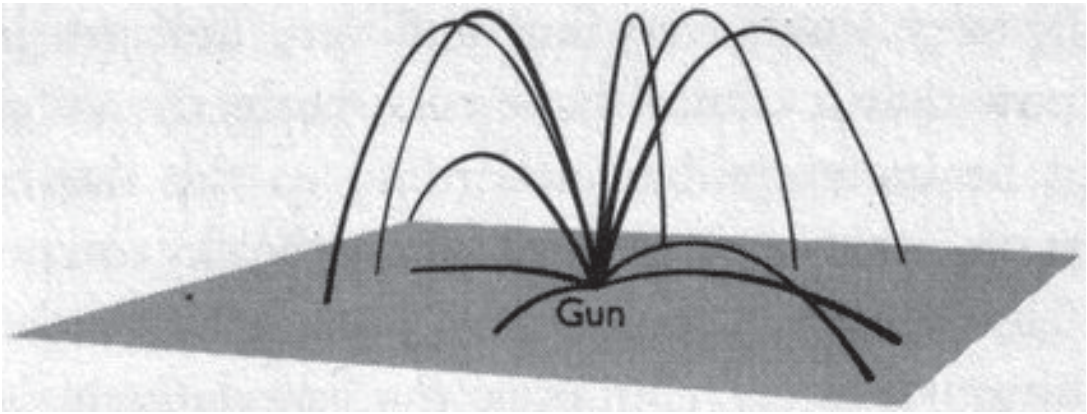

FIGURE 1.2. *Some possible trajectories of a cannon-ball fired from a gun. Each trajectory is compatible with the laws of motion, but only one of them is the trajectory on a particular occasion.*

Similarly, the laws of motion for physical reality as a whole would have many solutions, each corresponding to a distinct history. To complete the description, we should have to specify which history is the one that has actually occurred, by giving enough supplementary data to yield one of the many solutions of the equations of motion. In simple cosmological models at least, one way of giving such data is to specify the initial state of the universe. But alternatively we could specify the final state, or the state at any other time; or we could give some information about the initial state, some about the final state, and some about states in between. In general, the combination of enough supplementary data of any sort with the laws of motion would amount to a complete description, in principle, of physical reality.

For the cannon-ball, once we have specified, say, the final state it is straightforward to calculate the initial state, and vice versa, so there is no practical difference between different methods of specifying the supplementary data. But for the universe most such calculations are intractable. I have said that we infer the existence of ‘lumpiness’ in the initial conditions from observations of ‘lumpiness’ today. But that is exceptional: most of our knowledge of supplementary data — of what specifically happens — is in the form of high-level theories about emergent phenomena, and is therefore by definition not practically expressible in the form of statements about the initial state. For example, in most solutions of the equations of motion the initial state of the universe does not have the right properties for life to evolve from it. Therefore our knowledge that life *has* evolved is a significant piece of the supplementary data. We may never know what, specifically, this restriction implies about the detailed structure of the Big Bang, but we can draw conclusions from it directly. For example, the earliest accurate estimate of the age of the Earth was made on the basis of the biological theory of evolution, contradicting the best physics of the day. Only a reductionist prejudice could make us feel that this was somehow a less valid form of reasoning, or that in general it is more ‘fundamental’ to theorize about the initial state than about emergent features of reality.

Even in the domain of fundamental physics, the idea that theories of the initial state contain our deepest knowledge is a serious misconception. One reason is that it logically excludes the possibility of explaining the initial state

itself — why the initial state was what it was — but in fact we have explanations of many aspects of the initial state. And more generally, no theory of *time* can possibly explain it in terms of anything 'earlier'; yet we do have deep explanations, from general relativity and even more from quantum theory, of the nature of time (see Chapter 11).

Thus the character of many of our descriptions, predictions and explanations of reality bear no resemblance to the 'initial state plus laws of motion' picture that reductionism leads to. There is no reason to regard high-level theories as in any way 'second-class citizens'. Our theories of subatomic physics, and even of quantum theory or relativity, are in no way privileged relative to theories about emergent properties. None of these areas of knowledge can possibly subsume all the others. Each of them has logical implications for the others, but not all the implications can be stated, for they are emergent properties of the other theories' domains. In fact, the very terms 'high level' and 'low level' are misnomers. The laws of biology, say, are high-level, emergent consequences of the laws of physics. But logically, some of the laws of physics are then 'emergent' consequences of the laws of biology. It could even be that, between them, the laws governing biological and other emergent phenomena would entirely determine the laws of fundamental physics. But in any case, when two theories are logically related, logic does not dictate which of them we ought to regard as determining, wholly or partly, the other. That depends on the explanatory relationships between the theories. The truly privileged theories are not the ones referring to any particular scale of size or complexity, nor the ones situated at any particular level of the predictive hierarchy — but the ones that contain the deepest explanations. The fabric of reality does not consist only of reductionist ingredients like space, time and subatomic particles, but also of life, thought, computation and the other things to which those explanations refer. What makes a theory more fundamental, and less derivative, is not its closeness to the supposed predictive base of physics, but its closeness to our deepest explanatory theories.

Quantum theory is, as I have said, one such theory. But the other three main strands of explanation through which we seek to understand the fabric of reality are all 'high level' from the point of view of quantum physics. They are the *theory of evolution* (primarily the evolution of living organisms), *epistemology* (the theory of knowledge) and the *theory of computation* (about computers and what they can and cannot, in principle, compute). As I shall show, such deep and diverse connections have been discovered between the basic principles of these four apparently independent subjects that it has become impossible to reach our best understanding of any one of them without also understanding the other three. The four of them taken together form a coherent explanatory structure that is so far-reaching, and has come to encompass so much of our understanding of the world, that in my view it may already properly be called the first real Theory of Everything. Thus we have arrived at a significant moment in the history of ideas — the moment when the scope of our understanding begins to be fully universal. Up to now, all our understanding has been about some aspect of reality, untypical of the whole. In the future it will be about a unified conception of reality: all explanations will be understood against the backdrop of universality, and every new idea will automatically tend to illuminate not just

a particular subject, but, to varying degrees, all subjects. The dividend of understanding that we shall eventually reap from this last great unification may far surpass that yielded by any previous one. For we shall see that it is not only physics that is being unified and explained here, and not only science, but also potentially the far reaches of philosophy, logic and mathematics, ethics, politics and aesthetics; perhaps everything that we currently understand, and probably much that we do not yet understand.

What conclusion, then, would I address to my younger self, who rejected the proposition that the growth of knowledge was making the world ever less comprehensible? I would agree with him, though I now think that the important issue is not really whether what our particular species understands can be understood by *one* of its members. It is whether the fabric of reality itself is truly unified and comprehensible. There is every reason to believe that it is. As a child, I merely knew this; now I can explain it.

TERMINOLOGY

epistemology The study of the nature of knowledge and the processes that create it.

explanation (roughly) A statement about the nature of things and the reasons for things.

instrumentalism The view that the purpose of a scientific theory is to predict the outcomes of experiments.

positivism An extreme form of instrumentalism which holds that all statements other than those describing or predicting observations are meaningless. (This view is itself meaningless according to its own criterion.)

reductive A reductive explanation is one that works by analysing things into lower-level components.

reductionism The view that scientific explanations are inherently reductive.

holism The idea that the only legitimate explanations are in terms of higher-level systems; the opposite of reductionism.

emergence An emergent phenomenon is one (such as life, thought or computation) about which there are comprehensible facts or explanations that are not simply deducible from lower-level theories, but which may be explicable or predictable by higher-level theories referring directly to that phenomenon.

SUMMARY

Scientific knowledge, like all human knowledge, consists primarily of explanations. Mere facts can be looked up, and predictions are important only for conducting crucial experimental tests to discriminate between competing scientific theories that have already passed the test of being good explanations. As new theories supersede old ones, our knowledge is becoming both broader (as new subjects are created) and deeper (as our fundamental theories explain more, and become more general). Depth is winning. Thus we are not heading away from a state in which one person

could understand everything that was understood, but towards it. Our deepest theories are becoming so integrated with one another that they can be understood only jointly, as a single theory of a unified fabric of reality. This Theory of Everything has a far wider scope than the ‘theory of everything’ that elementary particle physicists are seeking, because the fabric of reality does not consist only of reductionist ingredients such as space, time and subatomic particles, but also, for example, of life, thought and computation. The *four main strands* of explanation which may constitute the first Theory of Everything are:

quantum physics Chapters 2, 9, 11, 12, 13, 14

epistemology Chapters 3, 4, 7, 10, 13, 14

the theory of computation Chapters 5, 6, 9, 10, 13, 14

the theory of evolution Chapters 8, 13, 14.

The next chapter is about the first and most important of the four strands, quantum physics.