THE SCIENCE OF CAN AND CAN'T

A PHYSICIST'S JOURNEY THROUGH THE LAND OF COUNTERFACTUALS

CHIARA MARLETTO

VIKING
An imprint of Penguin Random House LLC
penguinrandomhouse.com

Illustrations drawn by Vlatko Vedral

LIBRARY OF CONGRESS CATALOGING-IN-PUBLICATION DATA
Names: Marletto, Chiara, author.
Title: The science of can and can't : a physicist's journey through the land of counterfactuals / Chiara Marletto.
Description: New York : Viking, [2021] |
Includes bibliographical references and index.
Identifiers: LCCN 2020037751 (print) | LCCN 2020037752 (ebook) |
ISBN 9780525521921 (hardcover) | ISBN 9780525521938 (ebook)
Subjects: LCSH: Science—Philosophy. | Natural selection. | Counterfactuals (Logic)
Classification: LCC QH375 .M37 2021 (print) |
LCC QH375 (ebook) | DDC 501—dc23
LC record available at https://lccn.loc.gov/2020037751
LC ebook record available at https://lccn.loc.gov/2020037752

Printed in the United States of America
1 3 5 7 9 10 8 6 4 2

Designed by Alexis Farabaugh

Contents

Foreword

This is an intensely rational, transformative, and delightfully humane book about the power of taking *counterfactual* explanations of the world seriously. Those are explanations about what physical events could or could not be *made to happen*.

This is a major departure from the traditional conception of physics and science more generally, which takes for granted that scientific theories can only be about what *must* happen in the universe (or what is likely to), given what *has* happened, and which rejects such intangibles as causation, free will, and choice as being mere psychological props, or even mystical. And it even classifies such essential laboratory concepts as temperature, information, and computation as being incompatible with any exact description of nature, and convenient only at the level of human sensory experience.

But none of that is true. Those are no more than arbitrary limitations on our ability to understand the world, adopted only by custom and habit. Fortunately, they are widely flouted both in

everyday life and in theoretical science—albeit often guiltily and apologetically. If something is incompatible with the traditional conception, that does not make it incompatible with exact scientific description. It's just that in those cases, exact descriptions require a departure from the traditional conception—it requires counterfactuals. Something can hold *information* only if its state *could have been otherwise*: A computer memory is useless if all the changes in its contents over time are predetermined in the factory. The user could store nothing in it. And the same holds if you replace 'factory' with the Big Bang.

In this book, you will read why escaping from the traditional conception, and incorporating counterfactuals on an equal footing with factual statements into fundamental physics, is so promising—how it sheds a scientific light on much more of the world, informing a deeper conception of it and ourselves, and how it could facilitate further discoveries.

But there's more to it than that. Not only can counterfactuals provide new explanations; they are the foundation of a new *mode* of explanation. In the nineteenth and early twentieth centuries, not only were many new scientific explanations discovered, but new modes of explanation and understanding were being invented, too—such as evolution by natural selection, force fields, curved spacetime, quantum superposition, and the universality of computation. In the past few decades, in contrast, there have been none. Although new types of elementary particles have been discovered—and, for instance, the discovery of the Higgs particle was undoubt-

edly a triumph of both experiment and theoretical explanation—no new mode of explanation about physical phenomena has been discovered. In the first half of the twentieth century, however, the very idea of a particle as previously conceived had been swept away and replaced by a new, more deeply explanatory one.

With far fewer physicists, there were triumphs unmatched by anything in recent decades. Though the overall rate of scientific discovery has greatly increased by almost any measure, the discovery of *fundamentally* new truths about nature has, paradoxically, become less frequent. In fundamental physics in particular, there has been less and less exploration of transformative ideas—and new modes of explanation are not even being attempted.

This has happened for all sorts of more or less accidental reasons. But the net effect is a cautious and risk-averse culture in science: a preference for incremental over fundamental innovation, and for research with modest but foreseeable outcomes. In regard to fundamental progress itself, pessimism and fatalism have become the norm.

I don't agree with those who say that physics has already discovered all the "low-hanging fruit", and that all that remains is to harvest the rest, stolidly and mechanically. Nor with those who say that we apes are incapable of comprehending anything more fundamental than our current best theories, such as quantum theory and general relativity. On the contrary, in reality there has never been a time when there have been more blatant contradictions, gaps, and unresolved vagueness in our deepest under-

standing of nature, or more exciting prospects to explore them. Sometimes this will require us to adopt radically different modes of explanation.

The Science of Can and Can't sets out in nontechnical terms a new, counterfactual mode of explanation based on scientific and philosophical ideas that the author, Chiara Marletto, and I have pioneered. They provide new tools and new principles to address a number of notorious problems in physics and beyond. With a light but sure touch, Chiara Marletto argues for an emerging new theory, including a corpus of new and updated laws of nature—principles that will inform not just the next generation of atom-scale heat engines and nanorobots, but also artificial intelligence. This book goes through these topics with great enthusiasm and precision, punctuating the nonfiction in the chapters with short fictional stories that, in a manner reminiscent of Douglas Hofstadter's *Gödel, Escher, Bach*, elaborate the ideas of the book, to give the reader space to reflect.

In Chiara's land of counterfactuals, you will find new concepts (such as laws about information and knowledge) and old ones (such as work and heat) expressed in a radically different way. The Science of Can and Can't *can* enrich your understanding of the world, and of understanding itself.

David Deutsch

A Note on How to Read This Book

The details I remember most vividly from my childhood books are their illustrations: a colour plate depicting a monstrous whale in a memorable edition of *Pinocchio*; a black-and-white drawing of Captain Flint walking down a dark, narrow alley in *Treasure Island*; Quentin Blake's delicate illustrations for *Matilda*; Gustave Doré's terrifying beasts in Dante's *Inferno*; and so on. As I was writing this book, I thought I'd add a little something to each chapter that could work like illustrations do in books. Something as light and intriguing as a good conjuring trick; something that would capture the basic elements of the chapter and make them more memorable—turning them into a story. These stories are not essential to understand the scientific content of the book—you can skip them if you are in a hurry, and come back to them at some later point. They are intended as places where you can rest for a while, should you feel like having a pause from the scientific discourse that takes place within the chapters. I hope they will provide a fun addition to your exploration.

Prelude

If you could soar high in the sky, as red kites often do in search of prey, and look down at the domain of all things known and yet to be known, you would see something very curious: a vast class of things that science has so far almost entirely neglected. These things are central to our understanding of physical reality, both at the everyday level and at the level of the most fundamental phenomena in physics—yet they have traditionally been regarded as impossible to incorporate into fundamental scientific explanations. They are facts not about what is—the 'actual'—but about what *could or could not be*. In order to distinguish them from the actual, they are called *counterfactuals*.

Suppose that some future space mission visited a remote planet in another solar system, and that they left a stainless-steel box there, containing among other things the critical edition of, say, William Blake's poems. That the poetry book is subsequently sitting somewhere on that planet is a *factual* property of it. That the words in it *could be read* is a counterfactual property, which is true

regardless of whether those words will ever be read by anyone. The box may be never found; and yet that those words could be read would still be true—and laden with significance. It would signify, for instance, that a civilisation visited the planet, and much about its degree of sophistication.

To further grasp the importance of counterfactual properties, and their difference from actual properties, imagine a computer programmed to produce on its display a string of zeroes. That is a factual property of the computer, to do with its actual state—with what is. The fact that it *could be reprogrammed* to output other strings is a counterfactual property of the computer. The computer may never be so programmed; but the fact that it *could* is an essential fact about it, without which it would not qualify as a computer.

The counterfactuals that matter to science and physics, and that have so far been neglected, are facts about what *could or could not be made to happen* to physical systems; about what is *possible* or *impossible*. They are fundamental because they express essential features of the laws of physics—the rules that govern every system in the universe. For instance, a counterfactual property imposed by the laws of physics is that it is *impossible* to build a perpetual motion machine. A perpetual motion machine is not simply an object that moves forever once set into motion: it must also generate some useful sort of motion. If this device could exist, it would produce energy out of no energy. It could be harnessed to make your car run forever without using fuel of any sort. Any sequence of transformations turning something without energy into some-

thing with energy, without depleting any energy supply, is impossible in our universe: it could not be made to happen, because of a fundamental law that physicists call the *principle of conservation of energy*.

Another significant counterfactual property of physical systems, central to thermodynamics, is that a steam engine is *possible*. A steam engine is a device that transforms energy of one sort into energy of a different sort, and it can perform useful tasks, such as moving a piston, without ever violating that principle of conservation of energy. Actual steam engines (those that have been built so far) are factual properties of our universe. The *possibility* of building a steam engine, which existed long before the first one was actually built, is a counterfactual.

So the fundamental types of counterfactuals that occur in physics are of two kinds: one is the *impossibility* of performing a transformation (e.g., building a perpetual motion machine); the other is the *possibility* of performing a transformation (e.g., building a steam engine). Both are cardinal properties of the laws of physics; and, among other things, they have crucial implications for our endeavours: no matter how hard we try, or how ingeniously we think, we cannot bring about transformations that the laws of physics declare to be impossible—for example, creating a perpetual motion machine. However, by thinking hard enough, we can come up with more and better ways of performing a possible transformation—for instance, that of constructing a steam engine—which can then improve over time.

In the prevailing scientific worldview, counterfactual properties of physical systems are unfairly regarded as second-class citizens, or even excluded altogether. Why? It is because of a deep misconception, which, paradoxically, originated within my own field, theoretical physics. The misconception is that once you have specified everything that exists in the physical world and what happens to it—all the actual stuff—then you have explained everything that can be explained. Does that sound indisputable? It may well. For it is easy to get drawn into this way of thinking without ever realising that one has swallowed a number of substantive assumptions that are unwarranted. For you can't explain what a computer is solely by specifying the computation it is actually performing at a given time; you need to explain what the *possible* computations it *could* perform are, if it were programmed in *possible* ways. More generally, you can't explain the presence of a lifeboat aboard a pirate ship only in terms of an actual shipwreck. Everyone knows that the lifeboat is there because of a shipwreck that *could happen* (a counterfactual explanation). And that would still be the reason even if the ship never did sink!

Despite regarding counterfactuals as not fundamental, science has been making rapid, relentless progress, for example, by developing new powerful theories of fundamental physics, such as quantum theory and Einstein's general relativity; and novel explanations in biology—with genetics and molecular biology—and in neuroscience. But in certain areas, it is no longer the case. The assumption that all fundamental explanations in science must be

expressed only in terms of what happens, with little or no reference to counterfactuals, is now getting in the way of progress. For counterfactuals are essential to a number of things that are currently explained only vaguely in science, or not explained at all. Counterfactuals are central to an exact, unified theory of heat, work, and information (both classical and quantum); to explain matters such as the appearance of design in living things; and to a scientific explanation of knowledge. As I shall explain in this book, some of these things, such as information, heat, and work, already have some explanation in physics, but it is insufficient: it is only approximate, unlike more fundamental theories of physics, such as quantum theory and general relativity. Some others, such as knowledge creation, do not even have a fully fledged explanation yet. All these entities must be understood, without approximations, for science to make new progress in all sorts of fields—from fundamental physics to biology, computer science, and even artificial intelligence. Counterfactuals are essential to understand them all.

This book is an exploratory journey through the land of counterfactuals. It charts the territory beyond the boundary that has been currently set by the traditional conception of physics. Reading this book will feel a bit like going on an expedition in forbidden seas—like that of Darwin on the *Beagle*. You are going to explore alien lands with diverse beasts and creatures, taking note of what they are, and how they behave. At the end, you will have learnt something new about how to approach counterfactuals and

how they are key to address long-unsolved problems. Most important, you will see that a barrier has been erected that prevents us from understanding those entities; that each counterfactual property is at the heart of fields where physics and science more broadly are currently unable to make actual progress; and that it is vital to trespass the boundary in order to incorporate them into physics and science.

To clarify how to do that, I shall describe a few key open problems in physics that can be resolved fully by using counterfactuals. I shall start with the most fundamental phenomena—classical and quantum information—and then proceed to the theories of life and knowledge, and finally consider thermodynamics. All these phenomena have something in common: they are currently at best only approximately expressed in physics, but a counterfactual-based science can provide a unified explanation of them all, while also revealing unexpected connections between them.

I shall also explain how new scientific theories about counterfactuals can be formulated; I shall outline a whole new way of describing the workings of the universe, which constitutes what I call the Science of Can and Can't. This counterfactual-based approach to science can dramatically overhaul our current way of looking at the world, making it sharper and more powerful. It is a mind-blowing step with the potential to unlock centuries-old secrets.

1.

Such Stuff As Dreams Are Made On

Where I explain how to look at the laws of physics in a far broader way, including **counterfactuals** (statements about what transformations are possible or impossible); and you become acquainted with **knowledge**—defined objectively, via counterfactuals, as information that is capable of perpetuating its own existence.

Most things in our universe are impermanent. Rocks are inexorably abraded away; the pages of books tear and turn yellow; living things—from bacteria, to elephants, to humans—age and die. Notable exceptions are the elementary constituents of matter—such as electrons, quarks, and other fundamental particles. While the systems they constitute do change, those elementary constituents stay unchanged.

Entirely responsible for both the permanence and the imper-

manence are the laws of physics. They put formidable constraints on everything in our universe: on all that has occurred so far and all that will occur in the future. The laws of physics decree how planets move in their orbits; they govern the expansion of the universe, the electric currents in our brains and in our computers; they also control the inner workings of a bacterium or a virus; the clouds in the sky; the waves in the ocean; the fluid, molten rock in the glowing interior of our planet. Their dominion extends even beyond what actually happens in the universe to encompass what *can*, and *cannot*, be made to happen. Whatever the laws of physics forbid cannot be brought about—no matter how hard one tries to realise it. No machine can be built that would cause a particle to go faster than the speed of light, for instance. Nor, as I have mentioned, could one build a perpetual motion machine, creating energy out of no energy—because the laws of physics say that the total energy of the universe is conserved.

The laws of physics are the primary explanation for that natural tendency for things to be impermanent. The reason for impermanence is that the laws of physics are not especially suited for preserving things other than *elementary components*. They apply to the primitive constituents of matter, without being specially crafted, or designed, to preserve certain special aggregates of them. Electrons and protons attract each other—it is a fundamental interaction; this simple fact is the foundation of the complex chemistry of our body, but no trace of that complexity is to be found in the laws of physics. Laws of physics, such as those of our uni-

verse, that are *not* specially *designed*, or tailored, to preserve anything in particular, aside from that elementary stuff, I shall call *no-design laws*. Under no-design laws, complex aggregates of atoms, such as rocks, are constantly modified by their interactions with their surroundings, causing continuous small changes in their structure.

From the point of view of preserving the structure, most of these interactions introduce errors, in the form of small glitches, causing any complex structure to be corrupted over time. Unless something intervenes to prevent and correct those errors, the structure will eventually fade away or collapse. The more complex and different from elementary stuff a system is, the harder it is to counteract errors and keep it in existence. Think of the ancient practice of preserving manuscripts by hand-copying them. The longer and more complex the manuscript, the higher the chance that some error may be performed while copying, and the harder it is for the scribe to counteract errors—for instance, by double-checking each word after having written it.

Given that the laws of physics are no-design, the capacity of a system to maintain itself in existence (in an otherwise changing environment) is a rare, noteworthy property in our universe. Because it is so important, I shall give it a name: *resilience*.

That resilience is hard to come by has long been considered a cruel fact of nature, about which many poets and writers have expressed their resigned disappointment. Here is a magisterial example from a speech by Prospero in Shakespeare's *Tempest*:

Our revels now are ended. These our actors,
As I foretold you, were all spirits, and
Are melted into air, into thin air:
And like the baseless fabric of this vision,
The cloud-capp'd tow'rs, the gorgeous palaces,
The solemn temples, the great globe itself,
Yea, all which it inherit, shall dissolve,
And, like this insubstantial pageant faded,
Leave not a rack behind. We are such stuff
As dreams are made on; and our little life
Is rounded with a sleep.

Now, those lines have such a delightful form and rhythm that, on first reading, something important may go unnoticed. They present only a narrow, one-sided view of reality, which neglects fundamental facts about it. If we take these other facts into consideration, we see that Prospero's pessimistic tone and conclusion are misplaced. But those facts are not immediately evident. In order to see them, we need to contemplate something more than what spontaneously happens in our universe (such as impermanence, occasional resilience, planets, and the cloud-capped towers of our cities). We shall have to consider what can, and cannot, be made to happen: the *counterfactuals*—which, too, as I said, are ultimately decided by the laws of physics.

The most important element that Prospero's speech neglects is that even under no-design laws, resilience *can* be achieved. There

is no guarantee that it shall be achieved, since the laws are not designed for that; but it *can* be achieved because the laws of physics do not forbid that. An immediate way to see this is to look around a bit more carefully than was possible in Shakespeare's time. There are indeed entities that are resilient to some degree; even more importantly, some are more resilient than others. Some of them very much more. These are not, contrary to what proverbs and conventional wisdom might suggest, rocks and stones, but living entities.

Living things in general stand out as having a much greater aptitude to resilience than things like rocks. An animal that is injured can often repair itself, whereas a rock cannot; an individual animal will ultimately die, but its species may survive for much longer than a rock can.

Consider bacteria, for example. They have remained almost unchanged on Earth for more than three billion years (while also evolving!). More precisely, what has remained almost unchanged are some of the particular sequences of instructions that code for how to generate a bacterium out of elementary components, which are present in every bacterial cell: a *recipe*. That recipe is embodied in a DNA molecule, which is the core part of any cell. It is a string of chemicals, of four different kinds. The string works exactly like a long sequence of words composed of an alphabet of four letters: each word corresponds roughly to an instruction in the recipe. Groups of these elementary instructions are called 'genes' by biologists.

It is the particular structure, or pattern, of bacterial DNA that has remained almost the same over such a long time. In contrast, during the same period, the arrangement and structure of rocks on Earth have profoundly changed; entire continents have been rearranged by inner movements taking place underneath the Earth's crust. Suppose some aliens had landed on Earth early in prehistory, collected DNA from certain organisms (say, blue-green algae), and had also taken a picture of our planet from space; and that they were to come back now to do the same. In the pictures of the planet, everything would have changed. The very arrangement of continents and oceans would be utterly different. But the structure of the DNA from those organisms would be *almost unchanged*. So, after all, certain things in our universe, like recipes encoded in DNA, *can* achieve a rather remarkable degree of resilience.

The other element that Prospero's speech disregards is that living entities can operate on the environment, transform it, and (crucially) preserve the ability to do so again and again, thus leaving behind much more than 'a rack'. The Earth still bears the signs of bacterial activity from a billion years ago (for instance, in the form of fossil carbon). Plants have caused a dramatic change in the composition of the atmosphere by releasing gaseous oxygen as a side effect of converting the sun's light into chemical energy via photosynthesis. Humans, too, are capable of transforming the environment in a wide set of conditions. Contrary to Prospero's view, palaces, temples, and cloud-capped towers can achieve resilience—

because they are products of civilisation. Humans can restore them by following a blueprint—or rather, again, a recipe—of how they were initially built, guaranteeing that they will endure much longer than their constituent materials. In principle, a 3-D printer provided with such a recipe could reconstruct from scratch any ancient palace that happened to be completely destroyed.

The human life span may be still constrained, but technology has already extended it well beyond that of our ancestors. By changing the naturally occurring environment, human civilisation is tentatively improving and growing. We now have the knowledge to produce warm (or cooled) houses, powerful medications, efficient transport on Earth and even into space, and tools to save ourselves labour, to lengthen our lives and make them more enjoyable. We have majestic works of art and literature, music, and science. Those very words in Prospero's speech are an example of our literary heritage, and they have therefore survived—together with countless other wondrous outputs of human intellectual activity. So, rather than fading away, this pageant we have set up, which sustains us, has been under way for centuries. The rest of life's show on Earth has endured even longer, for billions of years.

Of course, the resilience of our civilisation is constantly threatened by severe problems, which crop up as we try to move forward. Some of them, such as global warming and fast-spreading pandemics, are in fact a by-product of the very progress I have described. These problems present considerable challenges and could easily wipe out several aspects of the progress we have made.

mentary components of matter, served in the form of an amorphous bubbling soup, and nothing more. So how can living entities, and the resilient recipes coding for the biological adaptations in their structure, have come about in the absence of a designer?

What Darwin discovered, and what Paley could not quite see, is that there is no need for any intentional design process: biological adaptations in animals can be created out of elementary components of matter, such as simple chemicals, via a nonpurposeful process called *natural selection*. That process needs only enough time and elementary resources, such as simple chemicals and so on. It is an undirected mechanism, and yet it can produce purposeful complexity, starting from scratch under laws of physics that are simple and no-design themselves.

There are two key concepts in Darwin's powerful explanation (as it is understood today). One is that of a *replicator*—whose key role in evolution has been exposed with superb clarity by Richard Dawkins in the celebrated book *The Selfish Gene*. Think of the bacterial example again. Each instruction in the recipe to build a bacterium is embodied in a particular pattern of a portion of bacterial DNA; that portion is called a 'gene'. Now, genes have a special property. Every time the bacterium cell self-reproduces and creates a new instance of itself, each gene's pattern gets replicated, or copied, accurately; then the rest of the new cell is constructed by executing the recipe in the DNA. Since they are capable of being replicated, those patterns are called 'replicators'. Incidentally, their replication is a step-wise, 'letter-by-letter' process, similar to that

used by monastic scribes to copy the content of ancient manuscripts; and it can be error-corrected via a similar method, which in bacteria is implemented by the cell once the replication has happened. In this way, the structure of bacterial DNA has survived for long: by being copied from generation to generation and potentially preserved for a much longer time than the bacterium's life span thanks to error-correction enacted by the cell. It is interesting that what's passed on from generation to generation, via replication, is the particular pattern that codes for a gene, or an elementary instruction: every time it is copied it changes its physical support, while retaining all its properties as a pattern. It is the same as what happens to the sequence of words copied by the scribes: the ink and bits of parchment embodying those words change, but the copied words are, if no typos occur, the same as those in the source manuscript. Patterns with this particular counterfactual property, that of being *copiable* from one physical support to another while retaining all their defining properties, are a special case of 'information'—of which I shall give a precise explanation (based on counterfactuals) in chapter 3.

So, the resilient recipes we see in animals around us must be constituted by some kind of information. To understand what kind, you must consider the second powerful concept of Darwin's theory: variation and selection.

While the copying process occurs from generation to generation, since the physical laws are no-design, errors can happen as a result of the interaction with the environment: these result in non-

purposeful changes ('variations') in the replicators. When errors happen at the right pace, not too often, and not too rarely, they produce novel variants of the genes in the newly formed bacterial cell, coding for a different trait—they produce new recipes. Sometimes, this means that individuals with that variant are able to cope better with the environment and become more successful—thus granting the perpetuation of that variant gene, and the recipe it codes for, to the detriment of the others. Less successful variants eventually go extinct as a result of the competition with the more successful variants in that particular environment. This phenomenon is 'natural selection': the blind process that can bring about something as graceful as a winged butterfly, and as elegant as an inky black panther, without having any clue as to what they should be like, just because replicating molecules replicate.

Natural selection gives us the key to explaining what makes the information in the surviving recipes worthy of attention. Since natural selection is blind, only few, particular changes are valuable and generate replicators that are capable of keeping themselves in existence: most of them are not, and lead to extinction. For example, in a forest where the trees have dark-coloured barks, only certain changes to the genes coding for the pigment of a moth's wing would be advantageous: for example, those that make the pigment closer to the bark's colour, so as to make the moth carrier of that trait less visible to predators. What distinguishes helpful changes in the recipe from unhelpful ones? It is a particular kind of infor-

mation: information that is *capable* of keeping itself instantiated in physical systems. It is resilient information.

I shall call this resilient information, which is the ingredient in successful recipes, 'knowledge' (and I shall talk about it extensively in chapter 5): for adaptations, it is knowledge of some features of the environment, such as that trees have dark-coloured bark. Knowledge in this sense does not have to be known by anyone: the moth does not know its wings are black. 'Knowledge' merely denotes a particular kind of information, which has the capacity to perpetuate itself and stay embodied in physical systems—in this case by encoding some facts about the environment. Natural selection is a process that, by nonpurposefully selecting for biological adaptations, can accidentally create knowledge. It is a non-directed, blind process, which with enough time and generic resources can bring about things that look as if they had been designed. But that is an illusion: no designer is needed.

The other kind of recipes I mentioned is those that maintain our civilisation in existence—by coding for how to build things like palaces, factories, cars, and robots.

Such recipes contain knowledge, too: they consist of information that can perpetuate itself, embodied in physical supports such as our brains, bits of papers, books, documentaries, historical records, scientific papers, conference proceedings, the internet, and so on. However, this kind of knowledge is brought about via a process different from natural selection: it is produced by thinking,

and it can reach further than knowledge that emerges directly by natural selection.

It is primarily via this kind of knowledge that humans have been able to construct a civilisation that is tentatively improving and growing, despite also often making bad mistakes. Such knowledge consists of thoughts. It is made of the same stuff as dreams are made on. Yet rather than fading away like fog in the morning sun, as Prospero suggests, knowledge is the key to resilience. The knowledge in his speech survives to this day. In fact, knowledge is the most resilient stuff that can exist in our universe.

Given that knowledge has such an essential role in the survival of complex entities, it is essential to understand the process by which new knowledge is created from scratch in our mind. Fortunately, this process was elucidated by the philosopher Karl Popper in the mid-twentieth century. He argued that knowledge creation always starts with a problem, which we can think of as a clash between different ideas someone has about reality. Incidentally, this suggests a rather positive, uplifting interpretation of conflicting states of mind where contrary impulses clash and fight. These conflicts are all examples of problems: but luckily problems can lead to new discoveries. For example, when writing a story, the clash in the author's mind might be between the desire to use elegant, lyrical language and the necessity of keeping the attention of

So far, I have explained that recipes are key to resilience; that they are made of a special kind of information—knowledge—which has the ability to keep itself in existence. I've also explained how the two known processes of creating knowledge work: by conjecture and criticism, in the mind; by variation and natural selection, in the wild.

An important point is that the laws of physics allow for knowledge creation, but they do not guarantee it. So knowledge creation may stop at any point. For example, natural selection can sometimes enter stagnation—which can result in events such as mass extinctions (like those that took place in prehistoric ice ages). The reason is primarily that natural selection, unlike conjecture and criticism, cannot perform jumps: each of the recipes that leads to a new resilient recipe must itself be resilient—i.e., it must code for a successful variant of a trait of the particular animal in question that permits the animal's survival for long enough to allow replication of that recipe, via reproduction. But there may be viable, resilient recipes coding for useful traits that can never be realised because they would require a sequence of nonresilient recipes to be realised first, which is impossible, as those recipes produce animals that cannot survive and cannot pass on their genes.

The thinking process, in contrast, can perform jumps. As we all know very well, the sequence of ideas leading to a good idea need not consist entirely of good, viable ideas. Nonetheless, knowledge creation in the mind, too, can enter stagnation and stop progressing. We must be wary of not entering such states both as

individuals and as societies. Particularly detrimental to knowledge creation are the immutable limitations imposed by dogmas, as they restrain the ability to conjecture and criticise.

The stupendous enterprise of knowledge creation has been unfolding over the centuries, producing brilliant works of art, music, literature, and powerful scientific theories. Looking back, we are comforted by seeing the progress made since the early beginnings of our civilisation. Looking ahead, it would seem that there might be several different fields in which progress can occur, especially in solving urgent and dramatic problems that humanity is now facing—such as the issues related to climate change and the open challenges in medicine and macroeconomics.

But if we could zoom out and see the arena of knowledge creation from above, a rather different scenario would appear within the domain of fundamental science, and of physics specifically. A boundary has been generated that affects and constrains the way criticism and conjectures can occur—a boundary that is keeping out certain kinds of explanations from the allowed set. These are explanations that involve counterfactuals. The boundary grew up because of a phenomenon that has been going on for some time, silently, largely unnoticed—like water seeping into a ship whose hull has a hidden hole below the waterline. To see what it is, we must start where it all began. We must start with physics.

It is perhaps ironic that this boundary-generating phenomenon started in physics, because physics is one of the clearest examples of how thinking can produce knowledge and make rapid progress. At a glance, from what one is taught in elementary courses at school, physics may appear like a collection of tools to solve irrelevant problems, of the kind you get in weekly physics tests: What is the time of flight of an apple that falls from a tree from a certain height? How long will it take for a bathtub of such a volume to be filled with water if the water is flowing in at this rate? And so on. Compared with other disciplines, such as literature or philosophy, physics may not seem to be about deep things at all. Who cares, after all, about how an apple falls? Isn't that fantastically narrow in scope?

This first impression is very far from the truth. Physics is a dazzling firework display; it is profound, beautiful, and illuminating; a source of never-ending delight. Physics is about solving problems in our understanding of reality by formulating explanations that fill gaps in our previous understanding. The point of physics is not the particular calculation about the fall of an apple. It is the explanation behind it, which unifies all motions—that of the apple with that of a planet in the solar system, and beyond. The dazzling stuff consists of explanations: for they surprise us by revealing things that were previously unknown and very distant from our intuition, with the aim of solving a particular problem. As I said, problems always consist of a contrast or clash between ideas about the world. For example, in the past, people believed

that the Earth was at the centre of the universe; but that notion clashed irremediably with observations, such as those about the apparent movements of the stars, of the other planets, and of the Moon. This led Copernicus and Galileo to conjecture that the sun, not the Earth, was at the centre of the solar system. The Copernican Revolution was an astonishing change of perspective, which allowed us to make formidable progress in understanding astronomy and celestial mechanics, and eventually led, via a series of further steps, to our current space exploration enterprises.

By solving problems of that kind, physicists have gradually uncovered entirely unsuspected worlds, telling us a deeper layer of the story of how things are. These layers are beyond the immediate reach of our senses, but our mind can visualise them in the light of explanations.

In existing physics, all explanations have some primitive elements, in terms of which the physical reality to be explained is expressed. The appearance of the dark sky at night is a perfect example of that. It can be explained in terms of unexpected underlying phenomena involving things like photons, the remarkable fact that the universe is expanding, and so on. None of those elements is apparent in the sky itself, but they are all part of the explanation for why it looks as it does, in terms of what is really out there. Explanations are accounts of what is seen in terms of mostly unseen elements.

There is no limitation, in principle, to how deep one can go in finding even more primitive elements. The primitive elements of