1. The Five Great Problems in Theoretical Physics

FROM THE BEGINNING of physics, there have been those who imagined they would be the last generation to face the unknown. Physics has always seemed to its practitioners to be almost complete. This complacency is shattered only during revolutions, when honest people are forced to admit that they don't know the basics. But even revolutionaries still imagine that the big idea—the one that will tie it all up and end the search for knowledge—lies just around the corner.

We live in one of those revolutionary periods, and have for a century. The last such period was the Copernican revolution, beginning in the early sixteenth century, during which Aristotelian theories of space, time, motion, and cosmology were overthrown. The culmination of that revolution was Isaac Newton's proposal of a new theory of physics, published in his *Philosophiae Naturalis Principia Mathematica* in 1687. The current revolution in physics began in 1900, with Max Planck's discovery of a formula describing the energy distribution in the spectrum of heat radiation, which demonstrated that the energy is not continuous but quantized. This revolution has yet to end. The problems that physicists must solve today are, to a large extent, questions that remain unanswered because of the incompleteness of the twentieth century's scientific revolution.

The core of our failure to complete the present scientific revolution consists of five problems, each famously intractable. These problems confronted us when I began my study of physics in the 1970s, and while we have learned a lot about them in the last three decades, they remain unsolved. One way or another, any proposed theory of fundamental physics must solve these five problems, so it's worth taking a closer look at each.

Albert Einstein was certainly the most important physicist of the twentieth century. Perhaps his greatest work was his discovery of general relativity, which is the best theory we have so far of space, time, motion, and gravitation. His profound insight was that gravity and motion are intimately related to each other and to the geometry of space and time. This idea broke with hundreds of years of thinking about the nature of space and time, which until then had been viewed as fixed and absolute. Being eternal and unchanging, they provided a background, which we used to define notions like position and energy.

In Einstein's general theory of relativity, space and time no longer provide a fixed, absolute background. Space is as dynamic as matter; it moves and morphs. As a result, the whole universe can expand or shrink, and time can even begin (in a Big Bang) and end (in a black hole).

Einstein accomplished something else as well. He was the first person to understand the need for a new theory of matter and radiation. Actually, the need for a break was implicit in Planck's formula, but Planck had not understood its implications deeply enough; he felt that it could be reconciled with Newtonian physics. Einstein thought otherwise, and he gave the first definitive argument for such a theory in 1905. It took twenty more years to invent that theory, known as the quantum theory.

These two discoveries, of relativity and of the quantum, each required us to break definitively with Newtonian physics. However, in spite of great progress over the century, they remain incomplete. Each has defects that point to the existence of a deeper theory. But the main

reason each is incomplete is the existence of the other.

The mind calls out for a third theory to unify all of physics, and for a simple reason. Nature is in an obvious sense "unified." The universe we find ourselves in is interconnected, in that everything interacts with everything else. There is no way we can have two theories of nature covering different phenomena, as if one had nothing to do with the other. Any claim for a final theory must be a complete theory of nature. It must encompass all we know.

Physics has survived a long time without that unified theory. The reason is that, as far as experiment is concerned, we have been able to divide the world into two realms. In the atomic realm, where quantum physics reigns, we can usually ignore gravity. We can treat space and time much as Newton did—as an unchanging background. The other realm is that of gravitation and cosmology. In that world, we can often ignore quantum phenomena.

But this cannot be anything other than a temporary, provisional solution. To go beyond it is the first great unsolved problem in theoretical physics:

Problem 1: Combine general relativity and quantum theory into a single theory that can claim to be the complete theory of nature.

This is called the *problem of quantum gravity.*

Besides the argument based on the unity of nature, there are problems specific to each theory that call for unification with the other. Each has a problem of infinities. In nature, we have yet to encounter anything measurable that has an infinite value. But in both quantum theory and general relativity, we encounter predictions of physically sensible quantities becoming infinite. This is likely the way that nature punishes impudent theorists who dare to break her unity.

General relativity has a problem with infinities because inside a black hole the density of matter and the strength of the gravitational field quickly become infinite. That appears to have also been the case very early in the history of the universe—at least, if we trust general relativity to describe its infancy. At the point at which the density becomes infinite, the equations of general relativity break down. Some people interpret this as time stopping, but a more sober view is that the theory is just inadequate. For a long time, wise people have speculated that it is inadequate because the effects of quantum physics have been neglected.

Quantum theory, in turn, has its own trouble with infinities. They appear whenever you attempt to use quantum mechanics to describe fields, like the electromagnetic field. The problem is that the electric and magnetic fields have values at every point in space. This means that there are an infinite number of variables (even in a finite volume there are an infinite number of points, hence an infinite number of variables). In quantum theory, there are uncontrollable fluctuations in the values of every quantum variable. An infinite number of variables, fluctuating uncontrollably, can lead to equations that get out of hand and predict infinite numbers when you ask questions about the probability of some event happening, or the strength of some force.

So this is another case where we can't help but feel that an essential part of physics has been left out. There has long been the hope that when gravity is taken into account, the fluctuations will be tamed and all will be finite. If infinities are signs of missing unification, a unified theory will have none. It will be what we call a *finite theory,* a theory that answers every question in terms of sensible, finite numbers.

Quantum mechanics has been extremely successful at explaining a vast realm of

phenomena. Its domain extends from radiation to the properties of transistors and from elementary-particle physics to the action of enzymes and other large molecules that are the building blocks of life. Its predictions have been borne out again and again over the course of the last century. But some physicists have always had misgivings about it, because the reality it describes is so bizarre. Quantum theory contains within it some apparent conceptual paradoxes that even after eighty years remain unresolved. An electron appears to be both a wave and a particle. So does light. Moreover, the theory gives only statistical predictions of subatomic behavior. Our ability to do any better than that is limited by the *uncertainty principle,* which tells us that we cannot measure a particle's positon and momentum at the same time. The theory yields only probabilities. A particle—an atomic electron, say—can be anywhere until we measure it; our observation in some sense determines its state. All of this suggests that quantum theory does not tell the whole story. As a result, in spite of its success, there are many experts who are convinced that quantum theory hides something essential about nature that we need to know.

One problem that has bedeviled the theory from the beginning is the question of the relationship between reality and the formalism. Physicists have traditionally expected that science should give an account of reality as it would be in our absence. Physics should be more than a set of formulas that predict what we will observe in an experiment; it should give a picture of what reality *is.* We are accidental descendants of an ancient primate, who appeared only very recently in the history of the world. It cannot be that reality depends on our existence. Nor can the problem of no observers be solved by raising the possibility of alien civilizations, for there was a time when the world existed but was far too hot and dense for organized intelligence to exist.

Philosophers call this view *realism.* It can be summarized by saying that the real world out there (or RWOT, as my first philosophy teacher used to put it) must exist independently of us. It follows that the terms by which science describes reality cannot involve in any essential way what we choose to measure or not measure.

Quantum mechanics, at least in the form it was first proposed, did not fit easily with realism. This is because the theory presupposed a division of nature into two parts. On one side of the division is the system to be observed. We, the observers, are on the other side. With us are the instruments we use to prepare experiments and take measurements, and the clocks we use to record when things happen. Quantum theory can be described as a new kind of language to be used in a dialogue between us and the systems we study with our instruments. This quantum language contains verbs that refer to our preparations and measurements and nouns that refer to what is then seen. It tells us nothing about what the world would be like in our absence.

Since quantum theory was first proposed, a debate has raged between those who accept this way of doing science and those who reject it. Many of the founders of quantum mechanics, including Einstein, Erwin Schrödinger, and Louis de Broglie, found this approach to physics repugnant. They were realists. For them quantum theory, no matter how well it worked, was not a complete theory, because it did not provide a picture of reality absent our interaction with it. On the other side were Niels Bohr, Werner Heisenberg, and many others. Rather than being appalled, they embraced this new way of doing science.

Since then, the realists have scored some successes by pointing to inconsistencies in the

present formulation of quantum theory. Some of these apparent inconsistencies arise because, if it is universal, quantum theory should also describe *us.* Problems, then, come from the division of the world required to make sense of quantum theory. One difficulty is where you draw the dividing line, which depends on who is doing the observing. When you measure an atom, you and your instruments are on one side and the atom is on the other side. But suppose I watch you working through a videocam I have set up in your laboratory. I can consider your whole lab—including you and your instruments, as well as the atoms you play with—to constitute one system that I am observing. On the other side would be only me.

You and I hence describe two different "systems." Yours includes just the atom. Mine includes you, the atom, and everything you use to study it. What you see as a measurement, I see as two physical systems interacting with each other. Thus, even if you agree that it's fine to have the observers' actions as part of the theory, the theory as given is not sufficient. Quantum mechanics has to be expanded, to allow for many different descriptions, depending on who the observer is.

This whole issue goes under the name *the foundational problems of quantum mechanics.* It is the second great problem of contemporary physics.

Problem 2: Resolve the problems in the foundations of quantum mechanics, either by making sense of the theory as it stands or by inventing a new theory that does make sense.

There are several different ways one might do this.

1. Provide a sensible language for the theory, one that resolves all puzzles like the ones just mentioned and incorporates the division of the world into system and observer as an essential feature of the theory.
2. Find a new interpretation of the theory—a new way of reading the equations—that is realist, so that measurement and observation play no role in the description of fundamental reality.
3. Invent a new theory, one that gives a deeper understanding of nature than quantum mechanics does.

All three options are currently being pursued by a handful of smart people. There are unfortunately not many physicists who work on this problem. This is sometimes taken as an indication that the problem is either solved or unimportant. Neither is true. This is probably the most serious problem facing modern science. It is just so hard that progress is very slow. I deeply admire the physicists who work on it, both for the purity of their intentions and for their courage to ignore fashion and attack the hardest and most fundamental of problems.

But despite their best efforts, the problem remains unsolved. This suggests to me that it's not just a matter of finding a new way to think about quantum theory. Those who initially formulated the theory were not realists. They did not believe that human beings were capable of forming a true picture of the world as it exists independent of our actions and observations. They argued instead for a very different vision of science: In their view, science can be nothing but an extension of the ordinary language we use to describe our actions and observations to one another.

In more recent times, that view looks self-indulgent—the product of a time we hope we have

advanced beyond in many respects. Those who continue to defend quantum mechanics as formulated, and propose it as a theory of the world, do so mostly under the banner of realism. They argue for a reinterpretation of the theory along realist lines. However, while they have made some interesting proposals, none has been totally convincing.

It is possible that realism as a philosophy will simply die off, but this seems unlikely. After all, realism provides the motivation driving most scientists. For most of us, belief in the RWOT and the possibility of truly knowing it motivates us to do the hard work needed to become a scientist and contribute to the understanding of nature. Given the failure of realists to make sense of quantum theory as formulated, it appears more and more likely that the only option is the third one: the discovery of a new theory that will be more amenable to a realist interpretation.

I should admit that I am a realist. I side with Einstein and the others who believe that quantum mechanics is an incomplete description of reality. Where, then, should we look for what is missing in quantum mechanics? It has always seemed to me that the solution will require more than a deeper understanding of quantum physics itself. I believe that if the problem has not been solved after all this time, it is because there is something missing, some link to other problems in physics. The problem of quantum mechanics is unlikely to be solved in isolation; instead, the solution will probably emerge as we make progress on the greater effort to unify physics.

But if this is true, it works both ways: We will not be able to solve the other big problems unless we also find a sensible replacement for quantum mechanics.

The idea that physics should be unified has probably motivated more work in physics than any other problem. But there are different ways that physics can be unified, and we should be careful to distinguish them. So far we have been discussing *unification through a single law.* It is hard to see how anyone could disagree that this is a necessary goal.

But there are other ways to unify the world. Einstein, who certainly thought as much about this as anyone, emphasized that we must distinguish two kinds of theories. There are *theories of principle* and *constructive theories.* A theory of principle is one that sets up the framework that makes a description of nature possible. By definition, a theory of principle must be universal: It must apply to everything because it sets out the basic language we use to talk about nature. There cannot be two different theories of principle, applying to different domains. Because the world is a unity, everything interacts ultimately with everything else, and there can be only one language used to describe those interactions. Quantum theory and general relativity are both theories of principle. As such, logic requires their unification.

The other kind of theories, constructive theories, describe some particular phenomenon in terms of specific models or equations.[1] The theory of the electromagnetic field and the theory of the electron are constructive theories. Such a theory cannot stand alone; it must be set within the context of a theory of principle. But as long as the theory of principle allows, there can be phenomena that obey different laws. For example, the electromagnetic field obeys laws different from those governing the postulated cosmological dark matter (thought to vastly outnumber the amount of ordinary atomic matter in our universe). One thing we know about the dark matter is that, whatever it is, it is dark. This means it gives off no light, so it likely doesn't interact with the electromagnetic field. Thus two different theories can coexist side by side.

The point is that the laws of electromagnetism do not dictate what else exists in the world. There can be quarks or not, neutrinos or not, dark matter or not. Similarly, the laws that describe the two forces—strong and weak—that act within the atomic nucleus do not necessarily require that there be an electromagnetic force. We can easily imagine a world with electromagnetism but no strong nuclear force, or the reverse. As far as we know, either possibility would be consistent.

But it is still possible to ask whether all the forces we observe in nature might be manifestations of a single, fundamental force. There seems, as far as I can tell, no logical argument that this *should* be true, but it is still something that *might* be true.

The desire to unify the various forces has led to several significant advances in the history of physics. James Clerk Maxwell, in 1867, unified electricity and magnetism into one theory, and a century later, physicists realized that the electromagnetic field and the field that propagates the weak nuclear force (the force responsible for radioactive decay) could be unified. This became the *electroweak* theory, whose predictions have been repeatedly confirmed in experiments over the last thirty years.

There are two fundamental forces in nature (that we know of) that remain outside the unification of the electromagnetic and weak fields. These are gravity and the strong nuclear force, the force responsible for binding the particles called quarks together to form the protons and neutrons making up the atomic nucleus. Can all four fundamental forces be unified?

This is our third great problem.

Problem 3: Determine whether or not the various particles and forces can be unified in a theory that explains them all as manifestations of a single, fundamental entity.

Let us call this problem *the unification of the particles and forces,* to distinguish it from the unification of laws, the unification we discussed earlier.

At first, this problem appears easy. The first proposal for how to unify gravity with electricity and magnetism was made in 1914, and many more have been offered since. They all work, as long as you forget one thing, which is that nature is quantum mechanical. If you leave quantum physics out of the picture, unified theories are easy to invent. But if you include quantum theory, the problem gets much, much harder. Since gravity is one of the four fundamental forces of nature, we must solve the problem of quantum gravity (that is, problem no. 1: how to reconcile general relativity and quantum theory) along with the problem of unification.

Over the last century, our physical description of the world has simplified quite a bit. As far as particles are concerned, there appear to be only two kinds, quarks and leptons. Quarks are the constituents of protons and neutrons and many particles we have discovered similar to them. The class of leptons encompasses all particles not made of quarks, including electrons and neutrinos. Altogether, the known world is explained by six kinds of quarks and six kinds of leptons, which interact with each other through the four forces (or interactions, as they are also known): gravity, electromagnetism, and the strong and weak nuclear forces.

Twelve particles and four forces are all we need to explain everything in the known world. We also understand very well the basic physics of these particles and forces. This understanding is expressed in terms of a theory that accounts for all of these particles and all of the forces except for gravity. It's called *the standard model of elementary-particle physics*—or the standard model, for short. This theory does not have the problem of infinities mentioned earlier. Anything we want to compute in this theory we can, and it results in a finite number. In

the more than thirty years since it was formulated, many predictions made by this theory have been checked experimentally. In each and every case, the theory has been confirmed.

The standard model was formulated in the early 1970s. Except for the discovery that neutrinos have mass, it has not required adjustment since. So why wasn't physics over by 1975? What remained to be done?

For all its usefulness, the standard model has a big problem: It has a long list of adjustable constants. When we state the laws of the theory, we must specify the values of these constants. As far as we know, any values will do, because the theory is mathematically consistent no matter which values we put in. These constants specify the properties of the particles. Some tell us the masses of the quarks and the leptons, while others tell us the strengths of the forces. We have no idea why these numbers have the values they do; we simply determine them by experiments and then plug in the numbers. If you think of the standard model as a calculator, then the constants will be dials that can be set to whatever positions you like each time the program is run.

There are about twenty such constants, and the fact that there are that many freely specifiable constants in what is supposed to be a fundamental theory is a tremendous embarrassment. Each one represents some basic fact of which we are ignorant: namely, the physical reason or mechanism responsible for setting the constant to its observed value.

This is our fourth big problem.

Problem 4: Explain how the values of the free constants in the standard model of particle physics are chosen in nature.

It is devoutly hoped that a true unified theory of the particles and forces will give a unique answer to this question.

In 1900, William Thomson (Lord Kelvin), an influential British physicist, famously proclaimed that physics was over, except for two small clouds on the horizon. These "clouds" turned out to be the clues that led us to quantum theory and relativity theory. Now, even as we celebrate the encompassing of all known phenomena in the standard model plus general relativity, we, too, are aware of two clouds. These are the dark matter and the dark energy.

Apart from the issue of its relationship with the quantum, we think we understand gravity very well. The predictions of general relativity have been found to be in agreement with observation to a very precise degree. The observations in question extend from falling bodies and light on Earth, to the detailed motion of the planets and their moons, to the scales of galaxies and clusters of galaxies. Formerly exotic phenomena—such as gravitational lensing, an effect of the curvature of space by matter—are now so well understood that they are used to measure the distributions of mass in galactic clusters.

In many cases—those in which velocities are small compared with that of light, and masses are not too compact—Newton's laws of gravity and motion provide an excellent approximation to the predictions of general relativity. Certainly they should help us predict how the motion of a particular star is influenced by the masses of stars and other matter in its galaxy. But they don't. Newton's law of gravity says that the acceleration of any object as it orbits another is proportional to the mass of the body it is orbiting. The heavier the star, the faster the orbital motion of the planet. That is, if two stars are each orbited by a planet, and the planets are the same distances from their stars, the planet orbiting the more massive star will move faster. Thus if you know the speed of a body in orbit around a star and its distance from the star, you

can measure the mass of that star. The same holds for stars in orbit around the center of their galaxy; by measuring the orbital speeds of the stars, you can measure the distribution of mass in that galaxy.

Over the last decades, astronomers have done a very simple experiment in which they measure the distribution of mass in a galaxy in two different ways and compare the results. First, they measure the mass by observing the orbital speeds of the stars; second, they make a more direct measurement of the mass by counting all the stars, gas, and dust they can see in the galaxy. The idea is to compare the two measurements: Each should tell them both the total mass in the galaxy and how it is distributed. Given that we understand gravity well, and that all known forms of matter give off light, the two methods should agree.

They don't. Astronomers have compared the two methods of measuring mass in more than a hundred galaxies. In almost all cases, the two measurements don't agree, and not by just a small amount but by factors of up to 10. Moreover, the error always goes in one direction: There is always more mass needed to explain the observed motions of the stars than is seen by directly counting up all the stars, gas, and dust.

There are only two explanations for this. Either the second method fails because there is much more mass in a galaxy than is visible, or Newton's laws fail to correctly predict the motions of stars in the gravitational field of their galaxy.

All the forms of matter we know about give off light, either directly as in starlight or reflected from planets or interstellar rocks, gas, and dust. So if there is matter we don't see, it must be in some novel form that neither emits nor reflects light. And because the discrepancy is so large, the majority of the matter in galaxies must be in this new form.

Today most astronomers and physicists believe that this is the right answer to the puzzle. There is missing matter, which is actually there but which we don't see. This mysterious missing matter is referred to as the *dark matter.* The dark-matter hypothesis is preferred mostly because the only other possibility—that we are wrong about Newton's laws, and by extension general relativity—is too scary to contemplate.

Things have become even more mysterious. We have recently discovered that when we make observations at still larger scales, corresponding to billions of light-years, the equations of general relativity are not satisfied even when the dark matter is added in. The expansion of the universe, set in motion by the Big Bang some 13.7 billion years ago, appears to be accelerating, whereas, given the observed matter plus the calculated amount of dark matter, it should be doing the opposite—decelerating.

Again, there are two possible explanations. General relativity could simply be wrong. It has been verified precisely only within our solar system and nearby systems in our own galaxy. Perhaps when one gets to a scale comparable to the size of the whole universe, general relativity is simply no longer applicable.

Or there is a new form of matter—or energy (recall Einstein's famous equation $E = mc^2$, showing the equivalence of energy and mass)—that becomes relevant on these very large scales: That is, this new form of energy affects only the expansion of the universe. To do this, it cannot clump around galaxies or even clusters of galaxies. This strange new energy, which we have postulated to fit the data, is called the *dark energy.*

Most kinds of matter are under pressure, but the dark energy is under tension—that is, it pulls things together rather than pushes them apart. For this reason, tension is sometimes

called negative pressure. In spite of the fact that the dark energy is under tension, it causes the universe to expand faster. If you are confused by this, I sympathize. One would think that a gas with negative pressure would act like a rubber band connecting the galaxies and slow the expansion down. *But it turns out that when the negative pressure is negative enough, in general relativity it has the opposite effect.* It causes the expansion of the universe to accelerate.

Recent measurements reveal a universe consisting mostly of the unknown. Fully 70 percent of the matter density appears to be in the form of dark energy. Twenty-six percent is dark matter. Only 4 percent is ordinary matter. So less than 1 part in 20 is made out of matter we have observed experimentally or described in the standard model of particle physics. Of the other 96 percent, apart from the properties just mentioned, we know absolutely nothing.

In the last ten years, cosmological measurements have gotten much more precise. This is partly a side effect of Moore's law, which states that every eighteen months or so, the processing speeds of computer chips will double. All the new experiments use microchips in either satellites or ground-based telescopes, so as the chips have gotten better, so have the observations. Today we know a lot about the basic characteristics of the universe, such as the overall matter density and the rate of expansion. There is now a standard model of cosmology, just as there is a standard model of elementary-particle physics. Just like its counterpart, the standard model of cosmology has a list of freely specifiable constants—in this case, about fifteen. These denote, among other things, the density of different kinds of matter and energy and the expansion rate. No one knows anything about why these constants have the values they do. As in particle physics, the values of the constants are taken from observations but are not yet explained by any theory.

These cosmological mysteries make up the fifth great problem.

Problem 5: Explain dark matter and dark energy. Or, if they don't exist, determine how and why gravity is modified on large scales. More generally, explain why the constants of the standard model of cosmology, including the dark energy, have the values they do.

These five problems represent the boundaries to present knowledge. They are what keep theoretical physicists up at night. Together they drive most current work on the frontiers of theoretical physics.

Any theory that claims to be a fundamental theory of nature must answer each one of them. One of the aims of this book is to evaluate just how well recent physical theories, such as string theory, have done in achieving this goal. But before we do that, we need to examine some earlier attempts at unification. We have a great deal to learn from the successes—and also from the failures.