# Theories of Everything[1]

There is a story about a Professor of Physics who was thrown out of his office (it seems that his grant had been terminated) and forced to live in a cave on a mountain top. One day he was visited by a young graduate student who had an important question.

Student: Professor, I wonder could you tell me what holds up the world?

Professor: As you should have learned already, the world is supported on the backs of four giant white elephants.

Student: Yes, that's what they say. But I wonder, what holds up the elephants?

Professor: Well, that is a good question. I will tell you a deep truth; the elephants are standing on the back of a huge tortoise.

Student: That does set my mind at ease somewhat, but please, one more question. What holds up the tortoise?

Professor: I will tell you an even deeper truth. Even I don't know what holds up the tortoise.[2]

I like to tell this story to beginning physics students, because it illustrates the basic structure of physics and physics education. The undergraduate physics major learns, in effect, about the four white elephants; the graduate student learns about the tortoise; and the physics professor spends his entire career trying to figure out what holds up the tortoise. Or, to put it in a less apocryphal way, undergraduate physics cannot be fully understood without reference to more advanced theory, and the more advanced theory in it's turn raises questions for which no answers are known.

Let's examine this in the context of a specific problem. We know that the planets move around our sun in elliptical orbits that obey Kepler's laws of planetary motion. Furthermore, the planets all turn in the same direction, and the ellipses all lie in approximately the same plane. In a word, why?

Well, planets obey planetary laws because they are held in their orbits by the gravitational attraction of the sun. This attractive force acts between pairs of massive objects with a force proportional the the product of their masses and inversely proportional to the square of the distance between them. Kepler's laws follow from this if we assume that the total energy



[2]A variant of this story appears in Steven Hawking's *A Brief History of Time.* In his version the punchline is, "It's turtles all the way down."

(kinetic plus potential) of each sun-planet system is negative and furthermore that this energy is always constant. We will also need to assume that the angular momentum of the system is constant. Finally, the reason that the ellipses all lie in a plane and that they turn the same direction with the particular radii that they have is because somehow they *started out that way* when the solar system was formed.

Notice that there are three almost unrelated pieces to the answer. First you will recognize Newton's law of gravitation. Then there are statements about things that are constant. We call these conservation laws. Energy and angular momentum are *conserved*. Finally, there are statements about how the solar system started out that do not in any obvious way relate to the laws. These are the *initial conditions.*

These three ingredients, a force law, conservation laws, and initial conditions, answer the question Why?, but in a profound sense, they also define what the question means. This is a *paradigmatic* explanation. What do you want to know when you ask why? According to this paradigm the answer has three parts: a force law, some conserved quantities, and some initial conditions. When you have these things you can calculate the position of each planet as a function of time from the remote past to the distant future. You can then compare your calculations with astronomical records from the past and observations in the future. To the extent that these observations agree with your calculations we say that the theory is *right.*

The graduate student in the anecdote will not let the matter rest here, however. He would like to know *why* gravitational force is inversely proportional to distance and *why* energy and momentum are conserved and *why* the planets started out with the particular masses, radii, and velocities that they had. These are questions for graduate school, and each one seems to call for a new paradigm. Let's start with Newton's law of gravitation.

The inverse square law of gravitation is "explained" by Einstein's general theory of relativity, which holds that massive objects distort the structure of space-time in their vicinity in such a way that moving objects follow the trajectories predicted by Newtonian physics. This is a deep and formidably difficult theory that makes many predictions that go considerably beyond the realm of Newton's equations. To the extent that we have been able to test the theory using recent space-age technology it is "right," but there is an alternative explanation of the inverse square law based on quantum mechanics that seems equally valid. This holds that the gravitational force is mediated by massless quantum particles called "gravitons" that move with the speed of light between pairs of massive objects and carry the momentum and energy necessary to keep them in their proper Newtonian trajectories.

At this point our hypothetical graduate student has more than he bargined for. For one thing, he was looking for the explanation of a simple equation, $F = mMG/r^2$. He now has to deal with theories of enormous complexity and subtlety. It will take him another year or two of graduate mathematics just to understand the equations involved! Rather than having one explanation he now has two, and they seem completely unrelated. We know that both quantum gravity and the general theory of relativity as we currently understand them are approximate theories; general relativity applies to cosmic distances and quantum mechanics applies to atomic and sub-atomic distances. One day there will be a theory that incorporates both, but it is hard to see how gravitons and curved space-time could coexist in the same theory. Finally, both theories raise a host of why? questions. Why is space coupled to mass? Why are gravitons particles in view of that fact that they are not particles *of* anything? Why does space have three dimensions? Why are there not other kinds of long-range interactions between massive objects? It seems that the explanation (or explanations) raise many more questions than they answer.

The existence of conserved quantities like momentum and energy has a different kind of explanation. From a mathematical point of view, energy is conserved because the laws of physics must be invariant under transformations in time. The theory must have the same form whether the time $t = 0$ is taken to be now or any other time in the past or future. Similarly, momentum is conserved because the theory must be invariant under transformations in space. It must not make any real difference whether the origin of my coordinate system is here on my desk or any other point in space. Other conserved quantities, like electrical charge for example, arise because of the invariance of the theory under more abstract mathematical transformations. Invariance principles like these are called "symmetries." A symmetry is a specific instance in which nature is less complicated than she might be. The laws of physics *could* be completely different in your town and in mine. If that were the case, momentum would not be conserved, and more important, any real understanding of physical phenomena would be impossible. All this can be demonstrated mathematically, but it raises the inevitable question, why does nature posess the symmetries we observe and not some others? It may also present us with an opportunity. Perhaps nature has other symmetries that we don't know about.

Finally, what about the initial conditions? We believe that our sun and solar system condensed out of a cloud of gas and dust. Some of the gas was primordial, *i.e.* left over from the creation of the universe, but some of the matter including all the heavy elements was detritus from the explosions of

earlier stars. Because gravitational force is always attractive, the particles coalesced into tight spherical masses, which eventually became the sun and planets as we now know them. The original cloud must have had some net angular momentum, which appears today in the form of the rotation and revolution of the sun and planets. None of this exactly specifies the initial conditions, of course. There are two reasons why. First, the outcome of this condensation depends on still earlier "initial" conditions; and second, even if we knew the precise position and velocity of each particle in the cloud at some earlier time and had some super computer to compute their subsequent motion, we could still not come up with a unique result for the subsequent planetary motion. The reasons for this have to do with chaos and were discussed in an earlier chapter. Thus, even though the relevant laws of physics may be no more complicated than $F = mMG/r^2$, we still cannot predict the outcome because of the vast number of particles involved.

I have discussed this issue of planetary motion in some depth to illustrate how scientific explanation works and to make two overarching points: first, the notion of what constitutes an "explanation" evolves constantly as a science matures, and second, every current explanation raises further questions of the form Why? The question is easy to ask, even if one has no idea of what form the answer might take. In the case of general relativity and quantum mechanics it is hard to imagine even what a *hypothetical* answer might be like, let alone know the *right* answer. Great creative scientists like Newton and Einstein don't just answer questions, they create new paradigms that give the questions well-defined meanings so that answers can be sought.

In view of this it is remarkable that some scientists are now claiming that we may be on the threshold of the ultimate theory, or as it is variously called, the theory of everything, the final theory, or the secret of the universe.[3] This theory would explain, in some sense, everything. The nature of the explanation would be such that the question Why? would no longer be appropriate. The purpose of this chapter is to examine the evidence in favor of such a theory and make some speculations regarding the form that it might take.

The hope for a final theory grows out of recent progress in particle physics and from a particular mindset called *reductionism.* Reductionism holds that a subject is "explained" if it can be reduced to the next simplest level of explanation. Biology is thus explained in terms of chemistry. Chemistry is

[3]Barrow, John D., *Theories of Everything; The Quest for Ultimate Understanding* (Fawcett Columbine, New York, 1991).
Weinberg, Steven, *Dreams of a Final Theory* (Vintage Books, New York, 1992).

explained in terms of atomic physics, and finally atomic physics is explained in terms of particle physics. Thus a theory that explains all of particle physics, in some sense, explains everything; but before we go any further it might be well to look closely at that phrase "in some sense."

Electrons and protons both fall within the domain of elementary particle physics. (Protons are composed of three quarks, but it is not possible to separate them into their constituent particles. Electrons, so far as we know, are genuinely elementary, *i.e.* not composed of anything else.) Put the two together and you have a hydrogen atom. Elementary particle physics predicts how electrons and protons will interact. This information can be put into the Schrodinger equation, and the solutions of this equation then tell us everything there is to know about hydrogen atoms. In this sense, elementary particle physics "explains" the hydrogen atom. Can it explain a water molecule in the same way? Yes, with two caveats. First, the oxygen nucleus is not an elementary particle. It is composed of eight protons and eight neutrons. *Presumably* it is possible to calculate all the properties of the nucleus from the properties of its constituent particles, but in fact such a calculation is way beyond our present capabilities. In the same way it should be possible to calculate all the properties of the water molecule from the properties of the ten electrons, the two protons and the oxygen nucleus. Again, this is far too complicated for us to do, except in a rough approximate way. The second caveat is that we must *assume* that the ensemble of ten protons, eight neutrons, and ten electrons does not introduce any new physics that is not contained in our fundamental theory. Until we are able to do the difficult calculations that these many-particle systems require this remains an article of faith.

With these two cautions in mind we can say that elementary particle physics explains atomic physics and most of chemistry. When it comes to complex organic molecules a new problem appears. Consider a molecule of DNA. In a sense it is "just" a complex combination of hydrogen, carbon, and oxygen atoms, but it also contains information. The information is so tightly packed that we have only recently decoded the DNA of a few simple life forms. Our fundamental theory cannot possibly tell us anything about this, for after all, the same DNA building blocks can encode the genetic information of a chimpanzee, a physicist, or a flatworm. There is a good analogy here between the DNA molecule and the modern digital computer. Of course the computer is an ensemble of atoms, but to say that it is "just" an ensemble of atoms or that it is "explained" by atomic physics is to ignore the circuit diagram and all the software.

In summary then, any putative theory of everything based on elementary

particles will always be vulnerable to these basic criticisms: first, one can never completely rule out the possibility that there are some new physical principles inherent in complex multi-particle systems, and second, the basic laws do not completely constrain complex systems, so these systems often carry information about which the Theory of Everything can tell us Nothing.

Suppose we adopt the more modest program of explaining everything about the existence and interactions of elementary particles. What are the prospects for such a theory, and how close are the current theories to qualifying? There are a number of questions that need to be examined.

First, what exactly is an ultimate explanation? What does it mean to explain "everything," even about elementary particles? This question has already come up in Chapter 1 in connection with Aristotle and his notion of first causes. Aristotle may not be entirely clear or consistent about this, but his idea of first cause seems to be inspired by the logical structure of (what we would now call) Euclidean geometry. An ultimate explanation is one that appeals to statements that, like the axioms of geometry, seem so self-evidently true that the question Why? is not appropriate. Why is it that parallel lines never meet? Well, they just DON'T, that's all. Whatever other shortcomings this idea might have, it is clearly inapplicable to the realm of elementary particles that are inconcievably small by human standards and obey the strange laws of quantum mechanics. What is self-evidentally true for us is irrelevant in the quantum world. We need another explanation of explanation. Several interesting suggestions have appeared in the literature recently. They are reviewed below. None of them seems to me to be entirely satisfactory.

John D. Barrow[4] has introduced the notion of algorithmic compressibility. Consider the following two sequences of integers:

$$01010101010101010\cdots$$

and

$$10001100111010010 1\cdots$$

Both sequences are infinite, only the first 20 1's or 0's are shown. The first sequence can be described completely by saying that it begins with 0 and then alternates 0 and 1. This sentence amounts to a short algorithm that describes every term of an infinite sequence. If you want to know the $n$-th term, simply decide whether $n$ is even or odd. If it is even the element is 1, otherwise it is 0. The second sequence, on the other hand, was obtained by

[4]Barrow, John D., *Op. Cit.*

a random process. There is no way to know the $n$-th element (for arbitrary $n$) without referring to the original sequence. We say that the first sequence is "algorithmically compressible," the second is not.

Theories, according to Barrow, are compression algorithms. The example of planetary motion illustrates this very well. The data consist of an endless table of the coordinates of each planet as a function of time. Kepler's laws (in addition to some initial conditions) predict all these coordinates using three simple statements. So the potentially infinite mass of data is compressed to a set of initial conditions and three laws. The three laws are further compressed to the succinct formula $F = mMG/r^2$ by Newton's law of universal gravitation.

Planetary motion also illustrates the shortcomings of this theory of theories as compression algorithms. Einstein's theory of general relativity is certainly more complicated that Newton's law of gravitation. If we think of it simply as an algorithm for computing things, it represents a catastrophic decompression! One could argue that it compresses a vast number of observations *that have not yet been made*, and that is exactly the point. The purpose of theories is not just to provide a succinct representation of the data but to suggest new observations, to unify areas of knowledge that had seemed unrelated, and in a word, to provide *understanding*.

Understanding is hard to quantify, however, and so is algorithmic compression. How would we know when we had the "final" theory? What is maximal compression, or for that matter, complete understanding? Several other ideas have been advanced that at least partly answer these questions. The most charming of these has been proposed by Gerard t'Hooft[5] and is based on Conway's Game of Life.

Conway's Game of Life was a pleasant diversion from the early days of home computers. Imagine a two-dimensionsal square lattice that is infinite in all directions. The individual cells are said to be "alive" or "dead." There is a clock that ticks like a metronome, and with each tick of the clock some cells die and some come to life. There are three simple rules that govern this:

- Each cell is adjacent to eight other cells, four of which touch its sides and four touch its corners. If exactly two of these neighbors are alive, the cell will stay as it is, either alive or dead.

- If exactly three neighbors are alive the cell in the center will live.

[5]t'Hooft, Gerard, "Questioning the answers or stumbling upon good and bad Theories of Everything," in *Physics and Our View of the World* ed. by Jan Hilgevoord, (Cambridge University Press, Cambridge, 1994)

- In all other cases the cell dies.

A single isolated live cell will die at the next tick of the clock, but more complex patterns of live cells survive the tick and take on a life of their own. These clusters of living cells move and change shape. They can interact with other clusters of cells in complex ways.

Perhaps our universe is like this, t'Hooft speculates.

> Imagine that, given a large enough lattice and a patient enough computer, an 'experiment' is run for a sufficiently long time. Regular recognizable structures may appear again and again, satisfying their own 'laws of physics.' In principle these laws of physics should be derivable from the original rules mentioned above. We could call these structures 'atoms.' Atoms themselves will form 'molecules,' and so on. Eventually, one might find 'intelligent' creatures built out of these building blocks. We might call them 'humans.' They will investigate the world they are in, and perhaps ultimately discover the three fundatmental 'Laws of physics' on which their universe is based.

The 'laws of physics' that these 'humans' discover will have three properties:

- They will determine with infinite accuracy the outcome of any set of initial conditions.
- There exist no closely resembling alternative theory. No other set of rules will describe the observed universe.
- Evolution according to these laws will give rise to nearly infinite complexity – including the emergence of life and intelligence.

It is hard to take this example too seriously, but it does provide a complete (albeit tenative) answer to the question posed above, what constitutes an ultimate explanation? In doing so, however, it raises a number of profound questions:

- Conway's Game of Life is played on a computer with software written by a programer who decides what the rules will be. If this example is somehow applicable to the real world, does it not automatically guarantee the existence of God?

- Conway's Game of Life is absolutely deterministic. Everything that happens is an inescapable consequence of the initial conditions. Is rational intelligence possible under these circulstances? Is it relevant to our world, which seems (for several reasons) not to be deterministic?

- Conway's Game of Life is played in a universe consisting of discrete cells. So far as we know, space in our universe is continuous. Can Conway's Game of Life be generalized to continuous space? What is the game like in such a space?

When physicists find that a realistic calculation of some phenomenon would be impossibly difficult, they often replace the exact theory with some trivialized model that is simple enough to allow an exact calculation. Of course, the results will not be relevant to the actual phenomenon. The hope is rather that in doing the simpler calculation one will have learned *something.* Perhaps Conway's game of life is a similar exercise. It is a drastic oversimplification of the real world, and it is just this that enables us to derive a clear answer to the problem of what constitutes an ultimate explanation.

Steven Weinberg in *Dreams of a Final Theory*[6] suggests several ways that we might recognize the final theory when we have it. One is that the theory must be logically *isolated.* This point was made above in connection with Conway's Game of Life, but I will let Weinberg speak for himself.

> In a logically isolated theory every constant of nature could be calculated from first principles; a small change in the value of any constant would destroy the consistency of the theory. The final theory would be like a piece of fine porcelain that cannot be warped without shattering. In this case, although we may still not know why the final theory is true, we would know on the basis of pure mathematics and logic why the truth is not slightly different.

Weinberg makes another suggestion that is both puzzling and intriguing; the theory should be beautiful!

> It is when we study truly fundamental problems that we expect to find beautiful answers. We believe that, if we ask why the world is the way it is and then ask why that answer is the way it is, at the end of this chain of explanations we shall find a

[6]Weinberg, Steven, *Op. Cit.*

> few simple principles of compelling beauty. We think this in part because our historical experience teaches us that as we look beneath the surface of things, we find more and more beauty. Plato and the neo-Platonists taught that the beauty we see in nature is a reflection of the beauty of the ultimate, the *nous*. For us, too, the beauty of present theories is an anticipation, a premonition, of the beauty of the final theory. And in any case we sould not accept any theory as final unless it were beautiful.

It is interesting that Weinberg (who in the same book rejects the traditional religious claims for the existence of God) should fall back on Platonic metaphysics in discussing the final theory. The word "nous" ($\nu o\tilde{\nu}\zeta$) is usually translated wisdom or intuition, but in fact, it means much more. I quote the Plato scholar G. M. A. Grube:[7]

> It (nous) is the culmination of intellectual research, the flash of insight that comes to those, and only to those, who have made a thorough study of their subject, $\cdots$ the grasping of the mind of the universal above and beyond the particular and with it a knowledge of ultimate moral and aesthetic values, the power to think clearly and logically and to see universal relations in the phenomenal world, the faculty of leaping to a right conclusion based on a full knowledge of the facts available.

It seems that we have begun our quest for the ultimate theory with Aristotle and ended with Plato!

There is an element of elitism that runs through *Dreams*. It is the Nobel laureates like Weinberg himself who will recognize the final theory and appreciate its beauty.[8] I can easily overlook that. What is much more interesting is that the ultimate theory leads immediately to metaphysical ideas of an almost mystical nature, the "reflection of the beauty of the ultimate," for example. That seems to me to be exactly right.

Despite our uncertainty regarding exactly what constitutes ultimate explanation, it is easy to recognize things that *need* to be explained. The following is a list of questions that we at least know how to ask: What is the nature of space and time? In particular, why is it that there appear to be three space dimensions and one time dimension? Why is it that the world of elementary particles seems to be characterized by a set of universal

[7] Grube, G. M. A., *Plato's Thought* (Beacon Press, Boston, 1958)

[8] There is a gentle rebuttal of this attitude at the end of Stephen Hawking's *A Brief History of Time*.

constants, and why do these constants have their particular numerical values? Why is it that elementary particles obey certain symmetries? These questions will be discussed briefly in the sections below.

# 1 Space and Time

Movement, as Aristotle realized, involves space and time. In modern language the *velocity* of a rigid object, $\boldsymbol{v} = \Delta\boldsymbol{x}/\Delta t$. This means that the object moves a small distance $\Delta\boldsymbol{x}$ in a short time $\Delta t$. (The symbol delta $\Delta$ simply means "a small amount.") The ratio of these two quantities is $\boldsymbol{v}$, the velocity, which is basic to the understanding of motion. This equation appears in every introductory physics textbook, and yet it is fundamentally mysterious. Let's start with the concept of distance. We all have an intuitive concept of it, but it is probably impossible to define it in a way that is not circular. The dictionary, for example, defines distance in terms of length, and, of course, defines length in terms of distance! Distance is a number that we read off of a ruler. A ruler is a straight object that is marked off in equal units of distance. Ultimately the ruler compares the lenght of the object we are measuring with the length of some standard object that is kept as a reference. Everyone knows what these sentences mean, but I at least cannot think of any way of explaining them that does not involve further references to distance and/or length.

There is a further subtlety to consider. In order to completely specify the path of our hypothetical rigid object, we need to state its direction in addition to its distance. There are many ways of doing this, but they all have this in common: they require three numbers. We emphasize this with our notation by making $\boldsymbol{x}$ boldface. In fact, $\boldsymbol{x}$ is a *table* of three numbers. The significance of each number will depend on some set of conventions we choose to describe the motion, but there is no way to describe the motion completely with less than three, and any description with more than three numbers involves some redundancy. We say, therefore, that "space is three-dimensional."

So why is space three-dimensional? It is easy in the context of pure mathematics to construct spaces of any non-zero integer dimensionality, yet the world of our experience has just three space-like dimensions. It is certainly true that life as we know it would be impossible in any dimensionality other than three. We exist, therefore space has three dimensions. This sort of reasoning is called an *anthropic* argument. As a matter of fact, our existence necessitates a great many things about the universe. We will have more to

say about this in a later chapter. From the standpoint of a hypothetical theory of everything, however, this sort of argument begs the question. We would like to have some formula or algorithm that would enable us to *compute* the dimensionality of space; perhaps an equation of the form $N = f(?)$, where $N$ is the dimensionality of space and $f$ is some function that arises naturally out of our theory. We do the calculation, and $N$ turns out to be three. What might go into the right side of this equation, however, is hard to imagine.

It is interesting to think about the question of dimensionality in the context of Conway's Game of Life. In this model the nature of space as well as the ultimate rules are "given." It would be pointless to try to calculate either. Perhaps our world is like this. We simply don't know.

The formula for velocity contains another mysterious term, the time that appears in the denomonator. We have an immediate intuitive understanding of time, but like space, it is difficult to define. Our prototypical moving object is at point $a$ and then at some later *time* it is at point $b$. The distance (in space) between these two points is $\Delta x$ and the interval in time between these two events is $\Delta t$. Everyone understands this sentence, but it is hard to explain it any further. Intervals of time are measured with "clocks," devices that do something (let's say they "tick") at regular intervals. We measure the time elapsed between two events by counting the clicks. The fact that different clocks will measure the same time interval between two specific events suggests, but certainly does not prove, that time has some independent existence apart from individual clocks. We might wonder if there would be time if there were nothing in the universe to change or move. This question has been debated for centuries, and in fact, it is not totally meaningless. The evidence from general relativity, as we will see, is that the existence of time and matter are closely related.

Time, unlike space, is one-dimensional. The time interval between two events is completely specified with a single number.[9] Perhaps a world in which time was two- or three-dimensional would make a good setting for a science fiction story, but so far as we know, it has no relevance to our universe.

Time is not only one-dimensional, it is also monotonic; it never runs backwards.[10] We usually take this asymmetry of time for granted, but

[9]The theory of relativity adds an important footnote to this statement. The time elapsed between two events depends on the coordinate system in which the time is measured. Two observers who are moving relative to one another will in general disagree about the time required for something to happen.

[10]Because of the effect mentioned in the previous footnote, there can be an ambiguity

it has some puzzling aspects.[11] To begin with, all the basic equations of motion in physics have a property called "time reversal invariance." This means simply that if one replaces the time variable $t$ everywhere it appears in the equation with $-t$, then all the extra minus signs introduced by this operation cancel and the equation remains unchanged.[12] Imagine making a movie of some particles interacting and then showing the movie backwards. That corresponds to replacing $t$ by $-t$. Time reversal invariance means that if the interaction that you photographed was consistent with the equations of motion, then so is the interaction you see when the film is played backwards. In this sense the equations make no distinction between past and future! And yet, everywhere around us we see evidence of the "one-way" sense of time; buildings decay, people age, and television gets worse with each passing year. We never see these things happening in reverse. The paradox is that equations that are symmetric in time describe a world that is so obviously not.

We can gain some insight into this problem by opening a new deck of playing cards and observing the sequence of cards. The first card (excluding the jokers) is always the ace of spades followed by the two of spades and so on to the king. This is followed by the ace of diamonds *etc.* with each card in a definite sequence. If you shuffle the deck you will of course obtain some different sequence. No matter how many subsequent shufflings you give the deck it will (almost) never return to its original sequence. The point is that there are 52! (roughly $8 \times 10^{67}$) possible sequences of 52 cards. Let's denote two specific sequences out of this vast number as $N_1$ and $N_2$. The probability that shuffling the deck when it is in configuration $N_1$ will put it in configuration $N_2$ is equal to the probability that shuffling the deck in $N_2$ will take it to $N_1$. In this sense shuffling the deck is time reversal invariant. The reason that shuffling seems asymmetric is that there are only a few sequences that seem to us orderly and inconceivably many that seem random. Thus shuffling a new deck takes it from a state of orderliness to which further shuffling never returns it. To put it another way, the appearance of asymmetry is due in part to our perception of what constitutes

in the temporal succession of certain events that occur in different places. Such events are said to be "space-like separated." Events that are "time-like separated" cary an unambiguous distinction between past and future.

[11] Davies, Paul, *The Physics os Time Asymmetry*

[12] There is a subtlety here regarding Schrodinger's equation in quantum mechanics. It is not time reversal invariant in the sense described above (because of the first time derivative), but the theory of which it is a part is time reversal invariant. This is a famous problem, which comes about at least partly because Schrodingers equation is not consistent with relativity.

order (the deck is always in *some* order) and partly due to the fact that the cards were put in a special order at the factory. This is an example of an initial condition as we discussed earlier in this chapter.

This example must be relevant to the real world, at least to some extent. Certainly the number of possible configurations of particles in the universe (possible in the sense that they do not violate any of the laws of physics) is much larger that the number of possible configurations of cards, and certainly our perceptions and prejudices play a role in what we see as the natural succession of events. This is the content of the second law of thermodynamics; things move from more orderly states to less orderly states. There are many examples of this in elementary textbooks. If we open up a bottle of compressed gas, the gas molecules fly out into the surrounding air. They will never, of their own accord, fly back into the bottle, because that is a much more orderly state than the condition in which the molecules are mixed randomly with air. In simple examples like this we can even quantify the notion of orderliness. That is the meaning of *entropy*.[13]

Does the second law of thermodynamics completely explain the one-way nature of temporal processes? There are at least two related difficulties. The first of these difficulties has to do with the initial conditions. The initial conditions, by definition, are not time reversal symmetric. They are *initial*; they specify the *past*. But if we leave aside for a moment the creation of the universe, then every set of initial conditions we can think of is in fact the end result of some earlier process, which *is* time reversal symmetric. Thus to say that the apparent lack of time reversal symmetry in the universe is just due to initial conditions is not to resolve the paradox at all. It is rather a systematic way of not thinking about it by predicating all apparent asymmetry to some earlier era.

The second difficulty is that the universe seems to be getting more orderly rather than less. The early universe as we understand it was an undifferentiated mass of particles. Out of this emerged galaxies and stars, planets and solar systems, and on at least one of these planets, living organisms of vast organized complexity. This is not necessarily inconsistent with the second law of thermodynamics, which only says that the entropy of a closed and isolated system must increase with every real process. So, for example, if we can approximate our galaxy as such a system, then the decrease in entropy brought about by the emergence of life on earth might be offset by a corresponding increase in the entropy of the galaxy as a whole. This argument,

[13]To be more precise, entropy is a measure of *dis*orderliness. Thus entropy always increases in any real process.

though not illogical, seems to me to be speculative and untestable. It raises further problems when we think about the entire universe. For one thing, it is difficult to make much sense of the "entropy of the universe." This is especially true if the universe is infinite in size. For another, the entire universe *that we can see* is more orderly than the early universe from which it evolved.

These problems have led some physicists to postulate two opposite "arrows of time," one arrow implied by the second law of thermodynamics that moves systems toward chaos and randomness and a second arrow that impels organized complexity out of chaos.[14] How and on what terms do these two arrows coexist? Does the second arrow require new laws of physics that do not come into play in simple systems, or is it already enshrined in the elementary laws as we understand them? These are questions for further speculation and research.

## 1.1 Space, Time, and General Relativity

About a hundred years ago the physicist-philosopher Ernst Mach proposed the following argument: suppose there were nothing in the universe except a single object, you for example. There would be nothing else in the universe with respect to which you could determine your position; therefore, it would not be possible to assign any meaning to your velocity $\boldsymbol{v} = \Delta\boldsymbol{x}/\Delta t$, even if you had a watch with which to measure time. So far there is nothing new or controversial about this statement, which is usually called "Galilean relativity."[15] But if there is no meaning to velocity, then *a fortiori* there is no such thing as acceleration, which is *change* in velocity. This *is* puzzling, because acceleration is not "relative"; for example, if you are sitting in a powerful automobile that suddenly accelerates, you will feel a sudden force that tends to snap your head back. Such forces are called "inertial forces," because they are consequence of the principle of inertia (objects at rest try to remain at rest) or "fictitious forces," because they seem to come out of nowhere. Whenever objects accelerate they experience these forces, which can be measured unambiguously. Another example is the famous "water pail" thought experiment originally proposed by Isaac Newton. Half fill a bucket with water, hang the bucket from a rope, and turn the rope so that the bucket rotates. At first the surface of the water is flat, but as the water itself begins to rotate its surface becomes concave. The point is that rotation itself is a kind of acceleration, and so the water is pushed toward the side of

[14] Paul Davies

[15] So far as I know, Galileo had nothing to do with it.

the bucket by the (fictitious) centrifugal force. The curvature of the surface can be measured unambiguously. If there were nothing else in the universe but the pail of water, it would not be possible to give any meaning to its rotation, but we could measure the curvature of the water and get definite numbers to describe it. We could calculate exactly how fast the water was rotating but yet be totally unable to answer the question, rotating with respect to what?

The conclusion that Mach drew from this argument is that inertial or fictitious forces arise because of some interaction between the accelerating matter and the rest of the matter in the universe. This is sometimes called Mach's principle. In the examples above there is no other matter in the universe, and so there would not be any fictitious forces. This resolves the paradox; unfortunately (so far as Mach was concerned), the only long range forces we know about are gravity and electromagnetism, and neither of these seemed to have the properties necessary to explain inertial force.

These ideas form the background out of which Einstein's theory of general relativity emerged starting around 1916. General relativity not only resolves the paradox, it combines in a profound way our understanding of gravity, the properties of space, and the structure of the universe. Although the theory is technically very difficult,[16] it is possible to understand the basic ideas in a completely non-mathematical way. We'll begin with the equivalence principle.

Near the surface of the earth gravity causes falling objects to accelerate with a constant acceleration (ignoring air resistance) of 32 feet per second per second, a famous number that we call $g$. Let us imagine with Einstein an elevator located in outer space where gravity is negligible and imagine that the elevator is accelerating (with respect to the rest of the universe) at a rate exactly equal to $g$. A person in the elevator would experience a force that would feel exactly like gravity on the surface of the earth, even though gravity is a "real" force (it is the result of an interaction with the earth) and the force experienced by the person in our hypothetical elevator is an inertial or fictitious force. Perhaps this seemed like an unimportant coincidence to everyone before Einstein. Einstein realized its significance and gave it the status of a law of physics that we call the *equivalence principle.*

Now turn the argument around; suppose the observer is inside an elevator that is falling freely near the surface of the earth. With respect to

[16]There is a story, perhaps apocryphal, that a reporter once asked the great English physicist, Sir Arthur Eddington, "Is it true that there are only three people in the world who understand Einstein's general relativity?" Eddington replied, "I'm trying to think who the third person might be."

the earth the elevator is accelerating at a rate equal to $g$, but inside the elevator the observer feels no gravitational force whatever. Physics is very simple in this world. Moving objects travel in straight lines with constant velocity; motionless objects hover unsupported in midair. We call such an environment an "inertial reference frame." Einstein realized that this perspective makes the study of space and gravity especially simple and defined the equivalence principle in these terms.

> In any gravitational field it is always possible to find a frame of reference in which (over a sufficiently small region of space and time) the laws of nature take the form they would have in an unaccelerated reference frame with no gravity.

It is possible to test this principle in many ways (that don't involve elevators in outer space). No deviation has ever been observed.[17]

One possible test of the equivalence principle would be to allow a beam of light originating from outside the elevator to traverse the space inside the elevator. Let's suppose the elevator is accelerating at the rate $g$ near the surface of the earth. According to the equivalence principle light travels is straight lines in this frame of reference. But from the point of view of someone outside the elevator, this means that the beam of light would appear to curve. Based on this argument Einstein predicted that light would also curve and by exactly the same amount under the influence of gravity as seen by an observer on the earth. This hypothesis has also been tested in numerous ways and has been corroborated to within the limits of experimental error. (The effect is very small in ordinary gravitational fields.)

A scientist unfamiliar with Einstein's work might assume that light is "pulled" into curved paths by the attractive force of gravity. Although this viewpoint is not really wrong, Einstein pursued a different idea that has turned out to be much more profound and fruitful. His viewpoint is that the presence of very massive objects changes the properties of space and time in their vicinity so that light takes a path that is not the familiar "straight line" that forms the basis of Euclidean geometry. The equivalent of a straight line, *i.e.* the path that light actually takes in these modified spaces, is called a geodesic. Just as the straight line is the shortest distance between two points in Euclidean geometry, so geodesics are the shortest distance between two points in these spaces that have been modified by gravitational fields. These geodesics do not always obey the axioms of Euclidean geometry, *e.g.* paths

[17] A brief review of the experimental evidence can be found in *Gravitation* by C. W. Misner, K. S. Thorne and J. A. Wheeler (W. H. Freeman and Co., San Francisco, 1973)

that are parallel in one region of space can cross somewhere else, rather they obey the rules of more complicated geometries that were developed by mathematicians in the nineteenth century. The details are quite technical; informally we say that "space is curved."

The origin of space and time are no less mysterious in the context of general relativity, but the theory does offer some perspective on the relation between the two. To begin with, the equivalence principle holds that the laws of physics take their simplest form in inertial reference frames, *so long as they are viewed over a sufficiently small region of space and time.* This caveat is necessary because if the gravitational field were non-uniform, it would be possible to spot some non-uniformity in the laws of physics if the elevator were large enough and/or the observations were conducted over a sufficiently long time. Consider, for example, our hypothetical elevator in free fall near the surface of the earth. Suppose it could fall freely all the way to the center of the earth (through a long tunnel, perhaps). Since the earth is a sphere, its gravity pulls all things down toward a point at the center; so if there were two motionless objects side by side in the elevator, they would have to move toward one another and eventually touch when the elevator reached this center point. We would see this gradual motion and conclude that the elevator was moving. So how small is "sufficiently small"? This depends on how non-uniform the gravitational field is and how accurately we make our measurements. It also depends on two related matters: how far apart the particles are to start with and how long we watch them. Obviously, the closer they are, the longer it will take to spot any movement. One way to say this is that "sufficiently small" refers to some four-dimensional space-time volume. For many applications, in fact, it is convenient to measure time in units of $tc$, where $c$ is the velocity of light; *ie.* time has units of "light centimeters." We can then talk about a region of space-time as having a volume with units of $\mathrm{cm}^4$.

The key idea in the above paragraph is "space-time." Space and time have to be considered on an equal footing as coordinates of any point or event. It was this realization that enabled Einstein to analyze the dynamics of gravity entirely in terms of geometry, the geometry of four-dimensional space-time.[18] The equivalence principle really says that physics is most simple when it is described in an inertial reference frame *in terms of space-time*

[18]It turns out that general relativity was "almost" discovered forty years earlier by the mathematician Bernard Riemann. Riemann had the clear idea of formulating the laws of physics in terms of the geometry of space, and he had all the mathematical formalism at his disposal (he had developed most of it himself). He only lacked the concept of space-time and tried to formulate the entire theory in terms of three-dimensional space.

*variables that are defined locally within the reference frame.* The problem of relating one of these frames to another, however, is much more difficult. The ultimate solution to this problem is contained (along with a great many other things) in Einstein's field equation. Unfortunately, the reader will need a good graduate course in general relativity to understand the workings of this (aparently simple) equation.

One of the most important applications of the field equation has been the development of a model for the average behavior in terms of matter, space, and time, of the entire universe! This work will be described more fully in the chapter on cosmology. The important point for this section on space and time is that they are both created at the moment of the creation of the universe. They have no meaning or existence outside the universe.

## 1.2 Constants of Nature

The world of elementary particles, as we have come to understand it in the last twenty years, consists of the following: first there are six kinds of quarks (and their antiquarks) to which we give the odd names, up, down, charmed, strange, top and bottom. (These properties are called "flavors.") These are never observed in isolation but in groups of three, which are called generically nucleons, and in quark-antiquark pairs called mesons. Because they cannot be isolated it is hard to give an unambiguous number for their masses. (They also carry a quantum number fancifully called "color," which cannot be observed in isolation.) The quarks are held together by the exchange of massless particles called "gluons." Gluons (for reasons that are not entirely clear) are also never observed in isolation, and the evidence for them is indirect. Next, there are the six leptons consisting of the familiar electron and two other electron-like particles called the $\mu$ (mu) and $\tau$ (tau) and three massless or nearly massless neutrinos that are somehow matched up with them, so that there is an electron-like neutrino, a $\mu$-like neutrino, and a $\tau$-like neutrino. For each of these leptons there is a corresponding antilepton. Then there are three heavy particles called $W^+$, $W^-$, and $Z^\circ$, that only manifest themselves in very high energy weak interactions. There are the messenger particles, the photon and graviton, which are massless; and finally a very heavy particle called the Higgs boson whose existence is expected on theoretical grounds but which has never been observed. Thus, there are about eighteen different particles in all depending on how you count.[19] So far as we know there cannot be any other particles in addition to these;

[19]The $W^+$ and $W^-$ are antiparticles of one another. There could be several kinds of Higgs particles.

and furthermore, these particles are themselves not made up of any other "smaller" particles. We can in fact measure (indirectly) the size of the electron, the only one of the eighteen that is massive, stable, and accessible. Its size is zero.

So why are there eighteen and not some other number? Our hypothetical Theory of Everything will have to answer this question, and we already have a few clues that will be discussed in the section on symmetries. A deeper puzzle perhaps is the fact that every electron is identical with every other electron, and each of the other eighteen particles is identical with others of its kind to within some theoretical caveats.[20] Just to revel in the mystery, let us meditate on the electron. Each electron carries a mass of $9.11 \times 10^{-28}$ grams. This number, of course, depends on the units in which is is measured, in this case grams. This unit itself is the consequence of some historical accident; presumably an intellegent being from another planet would use a different set of units and have a different number for the mass. The mass itself, however, is a well defined quantity that measures the particle's resistance to acceleration. So far as we know, every electron in the universe has exactly this same mass. The same remarks apply to the electron's charge, which is $1.602 \times 10^{-19}$ Coulombs. The charge measures the strength with which electron's interact with one another and with other charged particles. The units are culturally determined (even in this country, five different systems of units are used), but once we decide on the units, then every electron in the universe has the same charge.[21]

The origin of charge and mass is somewhat mysterious. According to the special theory of relativity there is a simple relation between mass and energy given by the famous formula $E = mc^2$, where $m$ is some mass, $c$ is the speed of light, and $E$ is the equivalent energy. We might plausibly argue that in the case of the electron the "equivalent energy" represents the work that would be required "squeeze" all the electron's charge into the small volume of the electron. Calculating this work is a simple problem in classical electromagnetic theory. The electron's mass is then simply $m = E/c^2$. Alas! If the electron really has no size, then $m$ is infinite. If we assume that it has some small size perhaps on the order of the current experimental error in

[20] Unstable particles, for example, cannot, by the laws of quantum mechanics, have a unique mass.

[21] The results of science, unlike the theorems of mathematics, cannot be stated unequivocally. We should say that in the region of the universe that we have been able to explore with our telescopes and to within experimental error, all electrons have the same charge. The reader should add this as a silent footnote to all the categorical statements in this section.

the determination of its radius, the mass still comes out too large by many orders of magnitude. Obviously the origin of mass cannot be understood in the context of classical physics.

The mass should be calculable in a much more sophisticated way using modern relativistic quantum field theory. This theory presents us with a set of equations that could, in principle, be solved to find the mass. This has not been possible due to the intractable nature of the equations. We do have a systematic way of calculating approximate solutions, however, called perturbation theory. In this approach one starts with a zero-order approximation called the bare mass, and then proceeds to calculate corrections to it. These corrections form an infinite series called the perturbation series. Each term in the series contains complicated integrals; and unfortunately, all of these integrals suffer from a strange pathology that causes them to be infinite! Despite this it is still possible to extract very precise numbers regarding some of the properties of the electron. Only the charge and mass cannot be calculated. Nature is evidentally trying to tell us something about these two quantities, but we have not understood the message.

The charge of an electron is an example of genre of theoretical numbers called coupling constants. They determine how strongly two particles interact with one another. Charge, in general, determines the strength of the coupling between any charged particles, electrons or protons for example, and photons. Since it is the photons that mediate the electromagnetic force, the charge (indirectly) fixes the force with which charged particles attract or repel one another. The complete theory of elementary particles contains many other coupling constants, which, among other things, determine the strength with which gluons interact with quarks and the rates at which $W^{\pm}$ and $Z^{\circ}$ decay into $e$, $\mu$, and $\tau$ particles and their associated neutrinos. There are various constraints on these constants, and there are various ways of expressing them, but like the electric charge they cannot be calculated. We have to take their numerical value from experiment.

In addition to masses and coupling constants there are two other numbers that describe the universe in a fundamental way and presumably have the same values everywhere, the speed of light and Planck's constant. The speed of light is self-explanatory. According to the special theory of relativity light travels with the same speed, $c = 2.998 \times 10^8$ meters per second, regardless of the speed of the source or the observer. This very counterintuitive statement has been verified in numerous ways and is one of the cornerstones of modern physics.

Planck's constant is more difficult to explain. It appears in Schrodinger's equation and in all of its relativistic generalizations. Its value, $h = 6.626 \times$

$10^{-27}$ erg-seconds, has the effect of setting the scale of distance, time, momentum, and energy over which quantum phenomena are important. For example, Heisenberg's uncertainty principle in its simplest form, $\Delta x\ \Delta p \geq h$, states that the maximum precision $\Delta x$ with which we can measure the position of a particle times the maximum precision $\Delta p$ with which we can determine its momentum must be larger than $h$. In some idealized quantum measurements the product might be equal to $h$. It can never be smaller. In the macroscopic world, of course, the precision with which we can measure something is limited only by the quality of our apparatus, and under no circumstances would our measurement of momentum be limited by a measurement we had already made on position (or vice versa). But when the object being measured is very small and the measurements correspondingly accurate, there appears an uncertainty that has nothing to do with our apparatus but is somehow woven into the fabric of reality. Imagine plotting elementary phenomena on a graph in which the ordinate is labeled "position" and the abscissa is labeled "momentum." Planck's constant has the units of area on this graph. Roughly speaking it is the area of the region in which quantum phenomena are important. We could plot a similar graph with the ordinate labeled "energy" and the abscissa labeled "time." Planck's constant also has units of area on this graph reflecting Heisenberg's other uncertainty relation, $\Delta E\ \Delta t \geq h$.

The speed of light and Planck's constant tell us something important about the structure of reality, but unlike masses and coupling constants they are not specific to a particular kind of particle. *All* massless particles move with the speed of light, always, and all massive particles can approach the speed of light but never equal it. All particles and ensembles of particles obey Heisenberg's uncertainty relations and many other laws and formulas in which $h$ appears.

Presumably our Theory of Everything would tell us why $c$, $h$, and all the masses and coupling constants mentioned above have the values that they have and not some others. Part of the answer, of course, resides in the units with which the constants are expressed. That is a complication, because the units are man-made conventions, even though the constants tell us *something* about reality that would be true even if humans had never appeared on the scene. For this reason it may be useful to find combinations of these constants in which the units cancel leaving a pure number. Such combinations would have the same numerical value regardless of the units used. The best known of these combinations is called the fine sturcture constant $\alpha$, which appears naturally in theories in which electrons interact

with photons.

$$\alpha \equiv \frac{e^2}{\hbar c} \approx \frac{1}{137}$$

In this formula $e$ is the charge on an electron and $\hbar = h/2\pi$. In the early days of quantum mechanics before the constants $e$, $\hbar$ and $c$ were known with their current precision, it was hoped that $\alpha$ was *exactly* equal to 1/137. That 137 was an integer (and in fact a prime number) suggested that the value of $\alpha$ could be "derived" with some simple combinatorial argument. Now, alas, we know that it's really 1/137.04, and no one has been able to come up with a convincing argument for deriving it.

Finally, in thinking about numbers we should remember that mathematics is replete with constants like $\pi = 3.141599265\cdots$ and $e = 2.71828182\cdots$ (the base of the natural logarithms) that are dimensionless and (in some sense) have the same values everywhere in the universe. These, however, *seem* to be in a different category from the other constants mentioned above. We can at least imagine a universe in which the electron mass has a different value. Perhaps there are other universes in which electrons don't even exist. We cannot imagine a universe, however, where the ratio of the circumference of a circle to its radius is anything other than $\pi$. I emphasize the word "seem." The existence of mathematics and its success in describing the real world are profoundly mysterious. (We will return to this subject in a later chapter.) Perhaps one day we will understand that the electron mass is just as inevitable as $\pi$. Then we will know that we are in posession of the Final Theory.

## 2 Symmetries and Conservation Laws

As explained in the introduction to this chapter, physics is possible because nature is less complicated and arbitrary than she might be. There are two equivalent ways to describe the order and simplicity that are apparent in nature, symmetries (or invariance principles) and conservation laws. Let us start with few examples of each.

Human beings have a familiar symmetry that is apparent when you look into a mirror; the person you see in the mirror looks like you. We say that the human form is invariant under reflection. Your right hand is not invariant, of course, because its image is a left hand; but on the whole, your left and right sides are matched up so that your image looks at least approximately like you. This symmetry can be generalized so that it also applies to elementary particles. To do this this we need to describe reflection as a

mathematical transformation. Set up a coordinate system so that the $x$-axis is perpendicular to the mirror. Every point on your body with coordinates $(x, y, z)$ has its counterpart in the mirror with coordinates $(-x, y, z)$. Reflection is thus described by the transformation $(x, y, z) \rightarrow (-x, y, z)$. In the world of elementary particle physics, this is called a parity transformation.[22] Most elementary particles are something like your body in the sense that each one is *exactly* identical to its mirror image; but because the particle is a quantum mechanical entity, there is an extra subtlety. The mathematical function that describes the particle sometimes changes sign when the parity transformation is performed. The sign change depends on the kind of particle involved. If the sign changes we say that the particle has an intrinsic parity equal to $-1$, otherwise its parity is $+1$. (There are some particles that, like your right hand, are not identical to their reflections. These particles do not have an intrinsic parity.) The point of all this is the following: most of the interactions that elementary particles undergo "conserve" parity. This means that the parity of an ensemble of particles simply obtained by multiplying together the intrinsic parities of the individual particles,[23] *doesn't change* as the particles interact with one another. To summarize: the symmetry that particles possess (like the symmetry of your body) causes them to be invariant under a particular transformation, the parity transformation or reflection. This invariance in turn gives rise to a quantity, parity, that is conserved in most reactions. This conservation law limits in a non-trivial way the possibilities that are open to interacting particles.

There is no obvious reason why this should be true; nature in this respect is just less complicated than she could be. There is more to the story than this, however. There are some particles, neutrinos in particular, that are not invariant under the parity transformation; but even here there is an element of symmetry. Neutrinos are non-invariant in the same way that our left and right hands are non-invariant. Neutrinos are all left-handed, antineutrinos are right-handed.[24] The parity transformation turns neutrinos into antineutrinos and vice versa. In this case it seems that nature has opted not for the simplest but the next simplest alternative. All the elementary

[22]Strictly speaking, a parity transformation is $(x, y, z) \rightarrow (-x, -y, -z)$, but the point $(-x, y, z)$ can be changed into the point $(-x, -y, -z)$ with a trivial rotation.

[23]I am oversimplifying a bit here for the benefit of readers who have not studied quantum mechanics. One needs to include the parity of the spatial wave function.

[24]A particle is said to be left-handed if its angular momentum is is always aligned in the opposite direction to its momentum. This can only be true unambiguously if the particle is massless. I should mention a persistent speculation that neutrinos in fact have an unmeasurably small mass. If this turns out to be true it would complicate to some extent the symmetries I have described.

particles we know about are either identical with their mirror image or they come in "matched pairs" like the left- and right-handed neutrinos.

I have explained parity in some detail, because it is paradigmatic: a symmetry is expressed in terms of an invariance principle, and the invariance principle in turn allows us to define a quantity that is conserved. It is somewhat atypical, however, because it is a *discrete* transformation. For example, if we perform two successive reflections, $(x, y, z) \rightarrow (-x, y, z) \rightarrow (x, y, z)$, and the system has returned to its original state. There are many other symmetries leading to conserved quantities that are based on *continuous* transformations. I mentioned in the introduction that the laws of physics are invariant under transformations in space and time. If we simply displace a system by a distance $a$ along the $x$-axis, $(x, y, z) \rightarrow (x + a, y, z)$, and $a$ can be any number whatsoever. Thus there are an infinite number of possible configurations $(x + a, y, z)$, and in all of them the laws of physics are exactly the same. This allows us to define momentum, which is also a continuous quantity that can have any numerical value (unlike parity, which can only be $\pm 1$). There is another example called the Gauge transformation, which has no counterpart in the classical world. Our theory of electromagnetic interaction is unchanged by this transformation. This has, as a consequence, the conservation of electric charge. It also limits and nearly fixes the way that photons interact with charged particles.

There is a body of mathematical lore surrounding symmetries based on transformations such as those we have just discussed called "group theory." Group theory itself is divided into two sub-fields, one dealing with discrete transformations, like parity, the other with continuous transformations as in the examples immediately above. This latter branch is called Lie group theory.[25] Lie groups are further subdivided into those transformations, like translations, in which the transforming parameter or parameters (like $a$ in the above example) can take on any value and those, like rotations, in which the parameters can take on only a limited range of values. In a simple rotation, for example $\theta \rightarrow \theta + \theta_\circ$, the transforming angle $\theta_\circ$ has to be less that 360°. Groups like this in which the range of transformations is limited by some periodicity are called "compact groups"; others, like displacement in which the parameters are not bounded, are said to be "non-compact." The compact groups are particularly interesting, because the quantities that are conserved (by virtue of symmetry under the transformations) take on only discrete values, but these values make complicated, symmetric, multi-

[25]The theory is named after the Swedish mathematician Sophus Lie (pronounced "Lee") who pioneered it during the late nineteenth century.

dimensional patterns. All possible groups of this sort and all possible patterns associated with them were cataloged by mathematicians in the early twentieth century. This work is part of the basic toolbox of every modern theorist in particle physics.

The catalog of all possible compact Lie groups is an infinite list, but there are only two of them that appear to be relevant to the physics of the real world. One, of course, is the group of rotations in ordinary three-dimensional space. Because physics is invariant under these rotations, angular momentum is not only conserved, but also it can take on only certain discrete values as we observe in the atomic world. The other group, which we call $SU(3)$, is the group of all rotations in some very abstract mathematical space in which there are eight rotation angles. Somehow this group is woven into the fabric of reality so that some mysterious quantum mechanical quantities called color and flavor are conserved. This is about all I can say without developing the complete theory of quantum chromodynamics, a subject for an advanced graduate course in particle physics. It is not necessary to understand the mathematics, however, to appreciate the mystery. Flavor and color are conserved because of a symmetry that is not really manifested in the real world, so far as we know, but only in some abstract mathematical space. Despite this, the quantities that are conserved, *are* conserved in the real world! Furthermore, it is only $SU(3)$ that works this way. The other members of the infinite list of possible symmetries are apparently not represented.

# 3 Conclusion

In the early 1960's there was a fashionable theory of elementary particles called the nuclear democracy. At this time many heavy unstable strongly-interacting particles were known, and more were discovered with each passing year. There seemed no way to argue that any of them were more "fundamental" or "elementary" than others, and there was no theory that could predict the existence of any the the particles before it was discovered. Out of this milieu grew the untimate theory of desperation, which is sometimes called the "bootstrap hypothesis." The basic idea is that there is no such thing as an elementary particle. Every particle is somehow composed of every other particle (thus the democracy), and since there is presumably only one way that this can come about, you can construct the ultimate theory without making any other assumptions, (*i.e.* you can pick yourself up by pulling on your own bootstraps). If this quest had been successful, it would

presumably have demonstrated that all particles have the properties that they have because only these properties are consistent with the bootstrap hypothesis. This would have been the final theory; everything would have been explained assuming nothing more than logical consistency. Unfortunately, the bootstrap theorists never achieved their goal, and the pursuit has fallen into disrepute even though its claims may ultimately be true.

During the heyday of the nuclear democracy, another group of theorists was pursuing an idea that seemed at first quite far-fetched; strongly interacting particles are composed of genuinely elementary particles called quarks. Quarks can never be observed in isolation, they carry fractional charge, and have a number of other bizarre properties. Eventually the interactions of quarks were quantified with a formula called the "color SU(3) Hamiltonian," a theory so formidably complicated and non-linear that exact solutions are out of the question. Despite its unpromising aspects, this theory has been much more fruitful than the bootstrap hypothesis and is now the cornerstone of most research in the field of strongly interacting particles.

Both the nuclear democracy and the quark theory apply to strongly interacting particles. They can be made consistent with the electro-weak theory, but it is a forced marriage. Their connection with gravitation and general relativity is obscure indeed. Let us overlook that, however, a consider both theories as prototypes of an ultimate theory of everything. The quark theory consists of a complicated equation into which one must insert the numerical values of various masses and other constants that cannot be calculated from the theory itself. The theory has built into it certain kinds of symmetry but without any ultimate explanation of why these symmetries are operative and not some others. Given the constants and the equation, however, one can, *in principle* calculate everything about the interactions of elementary particles. We might call this the Theory of Everything But. The bootstrap hypothesis, on the other hand, has no such limitations. It explains everything simply on the basis of logical consistency, but with our present state of knowledge, we cannot calculate anything or take any fruitful next step toward deepening our understanding. Let us call this the Theory of Everything and Nothing.

Perhaps this is a false dichotomy, but the quark theory has been enormously productive. It has given rise to a number of useful approximations and numerical calculations that we can apply to simple systems of strongly interacting particles. It is also the starting point for various attempts to unify the strong interactions with gravity. The nuclear democracy seems like a dead end, although it did stimulate some important work in clarifying the mathematical structure of scattering theory. If there is a lesson to be

learned from this comparison it is surely this: the most valuable theories in terms of deepening our understanding and advancing research are the incomplete and imperfect ones. At least until we are much wiser than we are at present, the theory of everything will turn out to be the theory of nothing at all.