Naturalness in Theoretical Physics: Internal constraints on theories, especially the requirement of naturalness, play a pivotal role in physics
Author(s): Philip Nelson
Source: *American Scientist*, Vol. 73, No. 1 (January-February 1985), pp. 60-67
Published by: Sigma Xi, The Scientific Research Honor Society
Stable URL: https://www.jstor.org/stable/27853063

Naturalness in Theoretical Physics

Internal constraints on theories, especially the requirement of naturalness, play a pivotal role in physics

Philip Nelson

Theoretical physics is not what it used to be. In the past few decades, the theories used to describe and explain the small corner of human experience called "physics" have become less determined by experiment than before. Indeed, whole legions of rival theories now give plausible explanations of the same phenomena. Ultimately, each of these many theories makes testable predictions about the physical world which distinguish it from its competitors—in principle. In practice, today's fundamental theories cannot be fully tested, due both to the computational difficulty of discovering just what they do predict and to the practical difficulty (or impossibility) of performing the required experiments.

Although the difficulty of connecting theory to experiment is now more acute than in the past, it certainly is not a new problem. Copernicus had no means at his disposal of observing the earth's motion from an external fixed point, yet he argued that the earth must move. A modern scientist transported back to the sixteenth century would have found this proposition immediately persuasive, even though Copernicus lacked experimental authority to make it. Why? The scientist would probably have replied that the heliocentric system was more *natural* than its predecessor, which by comparison seems almost laughably unnatural.

On closer examination we can dissect this modern line of reasoning into two parts. First of all, the geocentric system was *structurally* unnatural. It had numerous bodies executing complicated motions for no apparent reason (Fig. 1). The new model had these bodies executing a different motion, whose origin was still unknown but whose nature was considerably simpler. The complicated behavior of the planets as observed from earth was then computed as a superposition of their simple, more fundamental motions around the sun.

Prior even to the structural issues, however, comes the question of why the earth should be stationary at all, an objection so seemingly obvious that we might easily take it for granted. The key point here is that once we accept (as we do today) that the velocity of the earth cannot be measured by its inhabitants, then it becomes *numerically* unnatural for that velocity to be zero. If no measurement rules out motion and no valid principle forces the velocity to be zero, then it seems highly improbable that the earth should be at rest "by accident." We can conclude that the earth probably is moving, even before we observe a single planet (or sunrise!).

The aim of this article is to explore the theme of numerical naturalness in theoretical physics. From a supporting-cast role opposite its famous cousin, structural naturalness, it has achieved star status in its own right. Today arguments of numerical naturalness occupy an important place in fundamental physics, helping us distinguish good theories from bad ones. They not only tell us that certain theories cannot be fundamental but also sometimes suggest just where such theories may fail and what modifications may be necessary. And yet, naturalness seems to be one of the best-kept secrets of physicists from the public, a secret weapon for evaluating and motivating theories of the world on its deepest levels.

For all that, naturalness sometimes gives poor counsel. I will conclude with a short critique of the idea.

Themes in modern physics

Before entering into our discussion of naturalness proper, we need to review some of the major themes in the development of modern theoretical physics (see also Holton 1973). Each of these themes suggests a principle for the formulation of "good" theories. These principles were not handed down on stone tablets, but rather were arrived at by dint of hard work and much trial and error. The fact that they have become dogma today rests not so much on their intrinsic "beauty" as on their pragmatic successes: each has led to new theories which later proved to be correct in more objective ways. Once accepted, each theme has taken on a driving character, and peripheral ideas have had to adjust to it.

The first main theme can be called reductive. It is the principle that classes of many complicated things should be reducible to fewer, simpler things. The success of this notion, for example in the reduction of planetary motion to simple orbits, led eventually to a more or less firm faith in structural naturalness as a property of the world. Theories of fundamental particles provide another illustration. Molecules were divided into smaller numbers of atoms, atoms into their constituents, with such success that when later the number of known constituents began to proliferate uncontrollably, the idea that they too must have smaller, simpler constituents became irresistible—this time even before the latter (the quarks) had been observed as isolated fragments.

Philip Nelson holds degrees in physics from Princeton and Harvard and is at present a member of the Harvard Society of Fellows. Lately his research has concentrated on geometrical and topological properties of classical and quantum field theories. This research was supported by an NSF grant and a graduate fellowship. Address: Lyman Laboratory of Physics, Harvard University, Cambridge, MA 02138.

The next theme of interest to us might be called Copernican. With the somewhat deflating news that the earth was not the center of the solar system came inevitably the expanded notion that the earth was not very special at all. Eventually this gave rise to the "cosmological principle," first enunciated in its modern form by Bondi in 1948. It says that our position in the cosmos is completely undistinguished. Apart from local irregularities like our own galaxy, what we see from earth is a good representative sample of the rest of the universe. In fact, when one stands far enough back, the entire universe is uniform, with just as much matter here as over there.

The original cosmological principle was rather a lean mixture of observation and sheer expediency, since a large part of its motivation lay in the fact that cosmologists could not get on with the job of solving Einstein's equations without making assumptions as to what all parts of the universe were doing at once. More recent evidence, however, tends to support it strongly. Out to the limits of current observation, which well exceed the scales needed to smooth out local lumpiness, matter really is distributed uniformly (Peebles 1971). Even were this not so, though, the original Copernican notion would make sense. It would assert that if a theory of cosmology predicted that nearly all the universe had a local density of matter greater than ρ, say, but our observed local density were less than ρ, then that theory would probably not be correct, as it would require the earth to occupy a special place. But this is just a spatial version of the requirement of numerical naturalness discussed earlier. Indeed, the two notions are logically the same, the latter demanding that our world occupy an undistinguished location not in physical space but in some abstract space on a graph—a parameter space.

Just what we mean by "undistinguished location" is of course a subjective issue. How can some points be less typical than others? In the case of the earth's motion, most of us would agree intuitively that zero is a very special velocity. Most naturalness issues are based on such intuition: unnatural values are often very close to zero. We can sharpen this definition somewhat. Suppose that similar quantities in our theory (in our case the velocities of the planets) all have general magnitude roughly v, while the one in question, v_{earth}, is much smaller. Then v_{earth} is prima facie unnatural. If, in addition, changing v_{earth} to make it comparable to v would require drastic qualitative changes in the theory, then we will say that the small value is *strongly* unnatural and almost certainly in need of explanation.

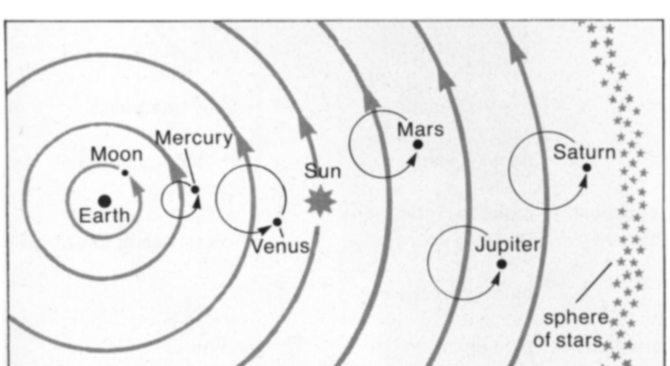

Figure 1. The Ptolemaic system, with the earth in the center, had the planets following complicated, epicyclical orbits within orbits, as shown here. Copernicus was able to explain the retrograde motion of the planets—their apparent change of direction at regular intervals—with the simpler system that we accept today.

As another example of subjective factors, our modern scientist transported back to the sixteenth century took it for granted that no "valid" constraint on a theory of the world could require a motionless earth. Obviously, Copernicus's contemporaries would have disagreed. Their intuitions told them just as strongly that the earth should *not* move. The meaning of the Copernican theme thus changes with time, along with changing notions about what might constitute a "valid" constraint on theories. In practice numerical naturalness is far less subjective than structural naturalness, which is notorious for the number of beautiful, wrong theories it generates each week.

To summarize, we have a strong naturalness problem whenever the set of theories which even remotely resemble our world is a tiny subset of all the acceptable theories. We must cure the problem by slicing the latter class down to size. This entails finding some new principle which renders most of its members unacceptable, leaving only a few—including of course at least one of the desired theories. In this way, theorists often permit the introduction of new structures into their theories, even when they are not strictly called for by observation. That is the point of this article.

Our third theme can be called hierarchical, and it is closely related to the first. Nature gratifies physicists by supplying a long chain of reasons-for-the-reasons no matter how many times we ask, "Why?" What is surprising is that the reasons all seem to have a roughly linear structure indexed by something we can call "fundamentalness." We say that the more fundamental statements "explain" their predecessors. Furthermore, fundamentalness seems always to be associated with size. Once we get to scales smaller than molecules, more fundamental constructions always seem to be associated with smaller sizes. On the other hand, we will see that on extremely long scales this relation is reversed: the larger domains become more fundamental on scales approaching the size of the universe.

There is an alternative to seeking explanations on ever more fundamental levels, namely, the possibility that the physical constants were set to special values by the agency of some kind of intelligence. This lies outside the scope of science. In any case it is of great interest to see just how far mechanistic explanations can be taken.

With the acceptance of the hierarchical theme came the notion of incomplete theories. Suppose our time-traveling modern scientist had suggested to a Newtonian like Laplace that while the law of universal gravitation worked almost perfectly for the planets, it nevertheless failed completely in different scale regimes, such as those inside neutron stars. Laplace probably would not have believed it. Had he believed it he might have discarded Newton's theory altogether; theories were either right or wrong (Merz 1904). Today we can look more kindly on such underachiever theories, associating with each a position in a sequence. Newton's law is perfectly correct within its range of validity, and we would no more discard it than we would hydrodynamics, even though we know that fluids are not really continuous.

Underlying the success of the hierarchical scheme is an idea we can call the principle of insulation: succeeding scales are insulated from one another, making it possible for us to understand one scale without understanding all the deeper ones. Instead, each level of our understanding can be described by theories containing a finite number of parameters, for example the mass density and viscosity of a fluid in hydrodynamics. These parameters appear arbitrary and are determined by experiment. When the next level of understanding is uncovered, however, in this case perhaps the kinetic theory of gases, the previously given parameters become computable, usually in terms of a smaller set of new parameters describing a world of shorter distances (Avogadro's number, for instance). In other words, the details of the deeper theory are encapsulated in a small number of quantities; these become the arbitrary parameters of the "effective" theory, which is thus insulated from the many other details of the underlying theory. As with the cosmological principle, we can gain some confidence in insulation by experimental observations in specific cases.

Tables 1 and 2 summarize some of the points made so far in this section; a few of the examples given there will be discussed later. Table 1 shows some highlights of the decreasing length hierarchy. Each entry has a name referring to a whole cluster of developments at around the given date and scale. Question marks denote speculation. Typical qualitative and quantitative properties explained and computed by theories, which previously were missing or had to be determined experimentally, come next. The last column lists some unnatural adjustments in earlier theories addressed by the given one.

Similarly, Table 2 shows the ascending staircase. The developments around 1965 are especially significant. In that year Penzias and Wilson detected a faint background noise in their microwave observations, which proved to be radiation coming from outside our galaxy. This discovery confirmed the theory that the universe has been expanding ever since it went through a hot, dense era when the radiation was emitted, about 18 billion years ago. This era almost certainly began with a "Big Bang," a time when the size of the universe was practically zero.

The Big Bang establishes the link between large and small scales mentioned earlier. Events which occurred very early in the history of the universe involved very short distances, but their traces have subsequently expanded to very large sizes. Later events have expanded less. Accordingly, Table 2 lists the hierarchy of scales in units of time after the Big Bang, with shorter times representing larger distance scales in today's universe.

To return to our list of themes, the fourth can be called symmetrical. Originally a refinement of structural naturalness, it has all but taken over the field. To begin, note that Newton's third law of motion, $F = ma$, treats every region of space in the same way. We say that it is invariant with respect to a uniform displacement through space, which simply means that the physics of billiard balls is independent of where we choose to place our billiard table. Nor does the physics change if we rotate the table, and so rotations are also an invariance of Newton's law. It has become customary to refer to all such invariances as "symmetries" of nature, by analogy with those of a snowflake: while the latter is not invariant under all rotations, still it is invariant under the subset of rotations by 60°—the symmetry operations of a snowflake.

The displacement symmetry of Newton's law lets us prove in all generality the theorem of the conservation of momentum. It also puts at our disposal powerful methods from mathematics for the analysis of *any* candidate laws of motion having this symmetry. For example, displacement symmetry prohibits the mass m which appears in Newton's law from varying from place to place. In fact, so restrictive are the resulting constraints on any law of motion that we discover that Newton's choice is essentially the only one possible. The important lesson we have learned, then, is that the presence of a symmetry in nature can force the equations describing dynamics to take on special forms.

We can also turn the argument around. Instead of arguing from an a priori symmetry to the form of the equations, we can argue, given some mysterious special feature of the observed world, that there must exist a structure, a symmetry, which explains it. That is, the symmetry cuts down the set of acceptable theories, as mentioned earlier. For example, among the many parameters in a theory of elementary particles, some of the most important are the masses of the various kinds of particles it describes. We will see some examples of how

Table 1. The hierarchy of decreasing scales

Scale	*Date*	*Name*	*Explains*	*Computes*	*Naturalness*
10^{-8} cm		molecules, kinetic theory	Boyle's law, etc.	density, viscosity	constancy of physical properties
10^{-8}		atoms	some chemical reactions	stoichiometric ratios	constancy of ratios
10^{-8}	1915	electron/nucleus, old quantum theory	Rutherford experiment	Rydberg energy	discreteness of ionization
10^{-12}	1935	nucleons, Weizsacker mass formula	decays, fission	binding energy curve	
10^{-13}	1961	Nambu-Jona-Lasinio pion	form of pion interactions	pion emission	pion mass
10^{-13}	1973	quarks, strong interaction theory	deep scattering	masses of strongly interacting particles	Zweig rule
10^{-16}	1967	weak interaction theory	neutral currents	corrections to weak processes	neutrino mass, weak universality
10^{-17}	1980	composite Higgs?		misc. exotica	Higgs mass
10^{-28}	1974	unified theory?	nucleon decay	θ_W, some masses	ratio of nucleons to photons

symmetry arguments can explain many otherwise puzzling and unnatural facts about these masses.

Einstein maintained this emphasis on symmetry in his 1905 special theory of relativity, changing only the exact set of invariances in question. Yet a nagging naturalness problem remained. The known symmetries of space and time explained much of the form of the laws of mechanics by disallowing other ones, but they could not explain the fact that the inertial mass in Newton's third law, $F = ma$, was always exactly equal to the mass appearing in his law of gravitation. Could some new symmetry be imposed on physics which would guarantee this equivalence of inertial and gravitational mass? The answer was yes, and the resulting "general coordinate invariance" was the basis of the 1915 general theory of relativity.

With the advent of quantum mechanics, symmetry took on even greater importance, particularly after the work of Wigner in the thirties. Shortly thereafter a crucial new idea emerged: perhaps not all symmetries had to do with space. Formally, there is no problem with expanding our notion of space to something larger with "internal" degrees of freedom and having arbitrary specified symmetries. Indeed, fruitful results emerged immediately when the observed equivalence between the nuclear forces felt by protons and neutrons was attributed to such an internal symmetry. Just as an electron has two states described by the direction of its angular momentum, so the proton and neutron can be thought of as two internal states of a single entity, the "nucleon." Positing an internal *symmetry* of the nuclear force now explains why protons and neutrons interact in the same ways. Symmetry soon became routinely accepted as a valid principle for reducing problems of numerical naturalness to questions of structure.

The next step came after World War II with the development of quantum electrodynamics. In quantum theory, electromagnetic effects are caused by the interactions of a particle called the photon, whose mass is experimentally known to be zero to very great accuracy, less than 10^{-20} times the mass of the electron (Jackson 1975). What principle could force this parameter of the theory to take on such a special value? The answer is that a "gauge symmetry" does the job. A gauge symmetry is a stronger, more restrictive version of some ordinary internal symmetry, and a theory possessing such a symmetry is called a "gauge theory." In this case the symmetry gives conservation of electric charge (just as displacement symmetry gives conservation of momentum), and any mass term for the photon spoils gauge invariance. Viewed differently, the imposition of gauge symmetry *forbids* us to give the photon a mass. We say that the symmetry "protects" the photon.

Meanwhile, in 1964 another internal symmetry was discovered, the so-called SU(3) of Gell-Mann, which described the arrangement of all particles subject to the strong interactions. By now there was no stopping the stampede. Symmetry was firmly entrenched in physics, and the validity of imposing it to eliminate some unwanted dynamical effect was never to be questioned again.

The fifth and last theme we will need is the one of hidden symmetry, often referred to as "spontaneously broken symmetry." While the laws of nature are symmetrical with respect to displacements and rotations, it would be difficult to convince a small physicist whose laboratory was only 10^{-8} cm long of this fact, if he were embedded in a large crystal. The scientist's measurements would all be affected by the intense electric fields in his neighborhood, leading him to infer a bias in one direction. In other words, the true symmetry of the world would be hidden from him. Were the crystal to melt, the scientist would discover the full set of true symmetries. Alternately, were he to shrink still further, he would find nuclear physics to be quite unaffected by the crystalline structure.

This example (from Coleman 1975) illustrates three important features of hidden symmetries. First, the apparent symmetry of the world can be very different on different scales. That is, in the hierarchy of scales some symmetries can become manifest only on deep levels while new, effective symmetries appear only on the more superficial ones. Second, even on one fixed scale the apparent symmetry of a system can depend on its state—which can change. Finally, a system with a hidden symmetry usually supports wave motions (in this case the sound waves in the crystal) whose energies can be arbitrarily small. In quantum theory, these waves correspond to particles. Since the smallest energy an object can have is proportional to its mass by Einstein's relation $E = mc^2$, the presence of a hidden symmetry in physics thus leads to the definite prediction that massless

Table 2. The hierarchy of increasing scales

Scale	*Date*	*Name*	*Explains*	*Computes*	*Naturalness*
	1543	Copernicus's solar system	apparent epicycles of planets		motion of Earth
	1916	general relativity	deflection of light	orbit of Mercury	equivalence of inertial and gravitational mass
10^{10} yr	1929	expansion of universe (Hubble's law)	recession of galaxies	distances to galaxies	isotropy of recession
	1932	Big Bang solution to Einstein equations	expansion	time of Beginning	
10^{12} sec	1965	microwave background, Standard Model of cosmology		isotope abundances	
	1970	singularity theorems			Big Bang is generic
10^{-35} sec	1980	inflation?	many domains	fluctuations in matter distribution	homogeneity of microwaves; inflation is generic
10^{-43} sec	1983	primordial inflation?			

particles should be present. These are called "Goldstone particles" associated to the original symmetry.

From their simple application to crystals, hidden symmetries went on to prove their worth in describing superfluids. And by analogy with superfluids, Nambu and Jona-Lasinio argued in 1961 that a hidden symmetry (which is named "chiral" symmetry) also existed in theories describing the strong interactions among elementary particles. In particular, it explained the otherwise unnatural existence of the pion, whose mass is nearly zero. The pion, they argued, was to be regarded as the Goldstone particle associated to the hidden chiral symmetry.

Given the presence of a massless particle in nature, symmetry thus provides at least two possible explanations. The particle may be protected by a manifest gauge symmetry, like the photon, or it may be massless due to a hidden symmetry, like the pion. Particles of spin $1/2$, like the neutrino, have a third option: it turns out that here a chiral symmetry which is not hidden can again force the mass to vanish. In the late 1960s Glashow, Weinberg, and Salam incorporated this mechanism into a theory of the weak interactions, which are responsible for radioactive nuclear decay. By using a *chiral* gauge symmetry, their theory could guarantee the masslessness of the neutrino. By using a chiral *gauge* symmetry, it could simultaneously account for the exact structure of the weak interactions, a mystery in the previous theory of Fermi. Both of these symmetries resolved naturalness problems. Finally, by using a *hidden* symmetry, their theory could incorporate electrodynamics without making it as weak as the weak interactions (Table 3). Hidden symmetry found a permanent place in physics.

Examples of naturalness in recent theories

Now that we have some idea of how naturalness arguments have worked in the past, we can proceed to more recent examples. By their nature, some of these examples will have to be rather technical. The nonspecialist reader can at any point skip to the last section.

In addition to explaining the masslessness of the photon and the structure of the weak interactions, gauge symmetry also proved to be the key to understanding the strong interactions. The resulting theory is called quantum chromodynamics, or QCD. Together with the Glashow-Weinberg-Salam theory it constitutes today's enormously successful Standard Model of the weak, electromagnetic, and strong interactions. For all its success, though, the Standard Model has a glaring naturalness problem. While it rigidly fixes all the strong charges on the various constituent particles relative to one another, it leaves their electric charges completely arbitrary. In nature, on the other hand, all electric charges are multiples of a fundamental unit to extraordinary accuracy, better than one part in 10^{20} (Jackson 1975). It seems clear that the electromagnetic part of the model has to be embedded in some theory with a larger set of symmetries even more hidden than that of the weak interactions and having the same desirable properties as that of QCD.

Georgi and Glashow took this point still further in 1974 with a remarkable observation: it proves possible and extremely attractive on structural grounds to make the larger theory *include* QCD. Since such a unified scheme would require that all three types of interactions be of the same strength, this proposal at first seems ridiculous; these interactions differ in strength by many orders of magnitude on the scales probed so far in the laboratory. But along with the realization that QCD was the correct theory of the strong interactions came the development in 1973 of techniques (Wilson's "renormalization group") to compute its effects on many different scales. These calculations showed that the strong interactions become effectively less strong at distances much smaller than the size of the proton. Perhaps at some very short scale, L_G, all interactions really are of the same strength. If they all were related by a symmetry hidden for distances greater than L_G, then the apparent conflict with experiment would be resolved.

The unified idea sounds fine until we compute that L_G is about 10^{-28} cm, a long plunge indeed from all other scales (Table 1). And yet perhaps this is not so farfetched. In the ascent from the unified underworld to the length scales characteristic of mortal physics, the unified theory becomes effectively the Standard Model. In so doing, it loses some of its original symmetries, which become hidden. At the same time, however, the unified theory picks up one new symmetry. This additional symmetry prohibits interactions which change the total number of nucleons (protons and neutrons) in the world. Such transactions take place routinely at the unified scale, but our world is insulated from these effects of the true theory by many intervening orders of magnitude. Indeed, no violation of nucleon number has ever been observed. If we see any violations, they will occur at an extremely, unnaturally small rate, which will be well explained by the concept of a unified theory. To put it another way, our insulation from the unified scale may be large, but it is not perfect. There should be a very small probability of our being able to observe a nucleon number violation. Given a large enough number of

Table 3. Elementary particle interactions in the Standard Model

Force	Physical phenomena	Relative strength[a]	Scale	Radiation quanta	Matter quanta	Quantum field theory
Strong	nuclear bonds, fission, fusion	1	10^{-13} cm	gluons	quarks	quantum chromodynamics
Electromagnetic	electricity, magnetism, light	10^{-2}	infinite	photon	quarks, charged leptons	Glashow-Weinberg-Salam theory
Weak	radioactive decay	10^{-5}	10^{-16} cm	*W*, *Z*, Higgs	quarks, leptons	Glashow-Weinberg-Salam theory

[a] $E \simeq 1$ GeV, or $r \simeq 10^{-14}$ cm

protons, for example, we ought to be able to see one decay. But how large a number do we need, and how long would we have to wait? Recent attempts to detect the decay of protons have yet to resolve this question (see, for example, Sulak 1982).

Of course anyone can invent elaborate theories without observable consequences. But recall that the distance regime we now observe was not always appropriate to describe the world. Shortly after the Big Bang, the size of the entire universe was L_G. Is there any evidence that nucleon number violations took place then? Indeed there is (Schramm 1983). In a world without nucleon violation, we can imagine two natural initial conditions for the Big Bang. One has a net nucleon number (nucleons minus their antiparticles) equal to zero. Essentially all nucleons meet their antiparticles and annihilate each other in bursts of radiation, leaving a world consisting only of radiation, that is, photons. This is not our world. The other, "generic" scenario has a net nucleon number of the same order as the total number of nucleons plus antinucleons. Then the number surviving annihilation is comparable to the number of annihilations, so that there are about as many nucleons as photons. In our world, however, the ratio of nucleons to photons is more like 10^{-10}. The universe is almost, but not quite, pure radiation. This is unnatural in itself. Moreover, things would be drastically different were this ratio closer to unity; so we have a strong naturalness problem.

Unified theories in principle make unambiguous predictions about the ratio of nucleons to photons: starting with no net nucleons they predict a small net production. Just how small a production depends on L_G. Although calculations based on estimates of photons and nucleons in the universe are too crude to be anything but suggestive, it seems that L_G must indeed be at least as small as the value obtained by the wholly independent considerations of particle interactions discussed above (Dolgov and Zeldovich 1981). Unified theories thus may provide the only solution to this naturalness problem.

Nor have we exhausted the implications of unified theories for cosmology. Not only do these models with hidden symmetry undergo transitions in the very early universe from phases with a different manifest symmetry, but such transitions also have a latent "heat," much as does the melting of our little scientist's crystalline world. In 1980 Guth observed that this latent heat could change Einstein's equations for the evolution of the universe when its size was comparable to L_G, giving rise to a period of explosive expansion at a rate much faster than previously assumed (Guth and Steinhardt 1984). This feature is not tacked on or separately postulated; it follows inevitably in any unified model. It means that the entire observed universe comes from a region which, prior to the transition, was at least 10^{50} times smaller than previously thought.

Now, if we heat a large piece of metal at one end and then measure its temperature soon after at various points, we will find it to be nonuniform, since heat takes a while to move from one end to the other. If, however, we examine one cubic millimeter of the metal, we will find that all points are at the same temperature, since all have had plenty of time to exchange heat and arrive at equilibrium. Similarly, large chunks of the early universe were expected to have widely varying temperatures, giving rise to a very uneven spectrum in the microwave background radiation as we look out at the sky from various angles. Instead, we know that this background is uniform to better than one part in 10^4, which for some time posed a serious naturalness problem. Why should the universe have been so homogeneous? This is equivalent to occupying a very distinguished location. With no valid constraint requiring thermal homogeneity, the probability of such a universe existing "by accident" is extremely small.

Guth's "inflationary" mechanism does away with the problem in the same way that we lose sight of the nonuniformity in the piece of metal. The observed universe comes from a chunk so small as to have been in thermal equilibrium when it emitted the microwave radiation we see today. This is, in fact, a general property of the inflationary idea: it provides insulation. Initial conditions such as the temperature distribution tend to be forgotten, "inflated away," by the end of one or more symmetry-breaking transitions, and so they need not be assumed to be unnaturally uniform. It is a theorist's dream. As with many dreams, though, the euphoria does not last long, for on closer inspection the mechanism responsible for the very large amount of expansion requires still other, unnatural adjustments to the parameters of the theory. This may well be a resolvable technical problem.

Not only inflation but the unified theories themselves suffer from new naturalness problems even as they solve old ones. Renormalization, for example, can explain why the scale of the strong interactions is so much larger than L_G, but no such argument exists to explain why the scale of the weak interactions, L_W, should also be large. Were L_W comparable to L_G, the weak interactions too would be insulated and as rare as nucleon decay. Then the stars would not burn, since their burning depends on a reaction step involving the weak interactions, and things would be decidedly different around here. We are thus faced with a strong naturalness problem.

Wilson's criterion

To see what can be done about this last problem, we turn to some general remarks about naturalness in the specific context of field theories. Up to this point we have been rather cavalier about particles like the pion which are "almost" massless. In fact, imposing a symmetry can explain only why a mass should be exactly zero, as with the photon. What we have implicitly been doing has amounted to assuming that the theory in which we impose the symmetry is really just an effective theory, partly insulated from a deeper, less symmetrical one. Small symmetry-breaking effects will still seep through, as with nucleon number violation, so that the effective theory will have only an approximate symmetry, and hence its particles will be only approximately massless. In the case of the strong interactions, an exact chiral symmetry would require a massless pion, but some unknown level of structure instead spoils the symmetry slightly, giving the pion its observed small mass. The pion's slight deviation from masslessness will thus be explained by the deeper theory when it is found.

This is the sort of argument we would like to repeat for the unified models. It turns out that the naturalness

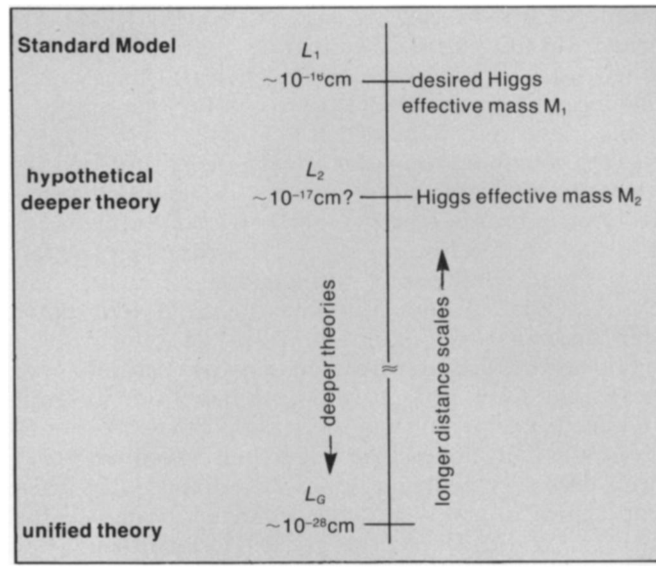

Figure 2. To explain why the Higgs effective mass on the scale L_1 has the right value to make the weak interactions work properly, we might invoke some new theory on a deeper scale L_2. Wilson's criterion implies that no such theory can explain M_1 naturally if L_2 is shorter than about $L_1/10$. If naturalness is correct, this means that qualitatively new physics will be within reach of the next generation of accelerators, which will be energetic enough to probe this domain.

problem involving the unexplained scale of the weak interactions amounts to getting a mass for a certain particle named the "Higgs" (after one of its inventors), whose dynamics determine whether and by how much the weak symmetry will be hidden. The Higgs mass has to be something like a thousand times that of the proton; while this is not small by our standards, it *is* "almost zero" compared with the typical mass appearing in unified models, and that is the relevant comparison. Can we invoke a new level of structure on some distant scale to resolve this naturalness problem, the way we do with the pion mass? In fact we cannot, as Wilson realized in a somewhat different context in 1971. (Here we will follow a more recent treatment due to 't Hooft 1980.) Wilson realized that the crucial feature of the pion mass which allowed it to be small was the way in which it is renormalized. Let us spend a few moments on this idea.

According to renormalization theory, not only the strengths of the various interactions but the masses of the participating particles appear to vary on differing length scales. To get a feel for this seemingly paradoxical statement, imagine firing a cannon underwater. Even neglecting friction, the trajectory will be very different from the corresponding one on land, since the cannonball must now drag with it a considerable amount of water, modifying its apparent, or "effective," mass. We can experimentally measure the cannonball's effective mass by shaking it to and fro at a rate ω, computing the mass from $F = ma$. (This is how astronauts "weigh" themselves in space.) Having found the effective mass, we can now replace the difficult problem of underwater ballistics by a simplified approximation: we ignore the water altogether, but in Newton's equations we simply replace the true cannonball mass by the effective mass. The complicated details of the interaction with the medium are thus reduced to determining one effective parameter.

A key feature of this approach is that the effective mass so computed depends on ω, since as ω approaches zero, for example, the water has no effect whatever. In other words, the presence of a medium can introduce a scale-dependent effective mass. We say that the effective mass is "renormalized" by the medium. In quantum physics, every particle moves through a "medium" consisting of the quantum fluctuations of all particles present in the theory. We again take into account this medium by ignoring it but changing the values of our parameters to scale-dependent "effective" values.

In order to have a particle of a given effective mass M_1 on our ordinary length scale L_1, we must therefore choose a particular value M_2, computed via renormalization, on the shorter scale L_2 where the next-deeper theory feeds into this one. In fact, many different parameters at L_2 can all feed in to M_1, and so if M_1 is a special value they will all have to be finely tuned in order to get the desired result. Thus, to get a very small M_1, it does not in general suffice to find an underlying theory which gives a small M_2. The farther away the deeper scale is (and as we have seen for unified theories, it is far indeed), the worse the problem; so in general a mass which looks normal on our scale will begin to look more and more unnatural at shorter scales until the theory breaks down and a deeper one takes its place.

In the particular case of the pion, however, no such problem arises: M_1 simply is not renormalized (at least not very much). This is because M_1 is protected by a symmetry. If it were exactly zero it would remain so on all scales, regardless of any other parameters; likewise a small nonzero value for M_1 gives an M_2 which is also small. We can therefore imagine solving our naturalness problem of the pion's mass with some theory at the deep scale L_2 which supplies such a small M_2. More generally, Wilson's criterion states that a small parameter in an effective theory is acceptable only if setting it to zero yields a more symmetrical theory.

Now we can return to the Higgs particle. Its mass M_1 must also be small. But when we set M_1 to zero, the Standard Model becomes no more symmetrical than before. Accordingly we expect that M_1 will be renormalized by a large amount on any scale L_2 which is too different from L_1. Actual calculations bear out this expectation and set 10^{-17} cm as the point where L_2 is "too different" from ordinary scales (Fig. 2). Thus *no* deeper theory can explain the Higgs mass naturally if its scale is shorter than about one-tenth the weak scale L_W.

Arguments of naturalness applied to the Higgs have thus made a remarkable prediction: there must be a specific intermediate scale, $L_I \cong 10^{-17}$ cm, where something new must happen. But what? Susskind (1979) offered a scheme in which L_I arises the same way the QCD scale does, but via the interaction of new particles. In this model the Higgs is actually a composite of the new fields, and its mass at L_I is just right to make the weak interactions work properly. More recent elaborations of this scheme go by names like "hypercolor" or "technicolor"; still other theories, in which naturalness demands that quarks, too, be composite structures, are called "rishon" or "preon" models. In the next few years there is a good chance that at least some of these ideas will be tested, and nothing will be more important to the future of naturalness as a physical criterion than the nature of the Higgs, if and when it is found.

Should we believe in naturalness?

It should be clear by now that naturalness has been voted in on the basis of a record of solid achievement. Nevertheless, a cautionary tale is in order.

Gell-Mann's SU(3) symmetry, which came up in our discussion of internal symmetries, described the tendency of the known particles subject to the strong interactions to assemble into sets of eight or ten with similar properties and masses. The theory made moderately successful quantitative statements relating various masses and reaction rates. Such relations seemed unnatural in the absence of some deeper symmetry, and so physicists immediately concluded that SU(3) had some fundamental significance. A symmetrical effective theory of the strong interactions was to give way to a deeper, less symmetrical "medium-strong" theory. We now know this to be completely wrong. The invariance operations described by SU(3) simply express the equivalence in the eyes of QCD of any quarks which happen to weigh less than the proton; the "breaking" of this apparent symmetry means only that, of the three lightest quarks, one is not as light as the others.

SU(3) is no more fundamental than the "symmetry" interchanging the three lightest planets, and the naturalness issue of why there are regularities among the strongly interacting particles is no more fundamental than the question of why the three lightest planets are lighter than the rest. They just are.

Nonetheless, the world seems to be a pretty natural place so far. As we have seen, though, a number of challenges loom at the next levels on the ladder. Why is the Higgs particle just right for stellar evolution? How about Newton's constant, the proton mass, the binding energy of deuterium? These quantities all seem to have in common a tender sensibility for the human race, since the slightest change in any would render the universe unfit for habitation. And yet no known principle can constrain them to such life-supporting values.

Some physicists see this last problem as fatal for naturalness. In 1961 Dicke coined the term "anthropic principle" to denote the idea that the ability to support life was itself an a priori valid constraint on theories (Gale 1981). Since we would not exist to observe a hostile universe, Dicke reasoned, no explanation is needed for the adjustments described above.

The anthropic principle's greatest liability is that of running against a successful incumbent. If nothing *needs* to be explained, then why *can* so much be explained? For instance, the details of organic chemistry are just as crucial to life as those of stellar evolution, and yet it would have been a regrettable error if, at the turn of the century, scientists had concluded that the relative bond strengths and so on could consistently have taken any value, and that we just inhabit a world conveniently arranged for us. Instead, they found a deeper theory, quantum mechanics, which made the bond strengths computable, not arbitrary, and so eliminated the naturalness problem. Thus any rejection of naturalness on deep scales must also explain its successes on less deep scales. Can we do that?

Perhaps we can. Dicke realized that for his idea to work there would have to be many universes, each with varying values of the physical constants. The fact that we find ourselves in a friendly universe is then as tautological as the fact that we live on Earth and not Pluto. Many physicists reject this many-worlds assumption as metaphysical. But consider one last time the little scientist embedded in a crystal. Suppose that his crystal was formed quickly, so that it has various domains, each pointing in different directions, or even having different crystal structures altogether. A second, distant scientist might well find her world a very different place.

We have already seen how our universe has many regions which were out of touch at the time of symmetry-breaking (or, in our analogy, freezing of the crystal). If some form of the inflationary theory of the early universe is correct, then our own domain is probably bigger than what we can see; being little scientists ourselves, we might erroneously conclude that we live in a perfect crystal, not one with many domains. What is important is that there is now nothing metaphysical about the notion of many worlds. Experiments in our laboratories can in principle determine which unified theory is correct, if any, fixing the amount of inflation to be expected. If we find a billion domains with varying values for the Higgs mass, then we can probably conclude that no explanation of its value in our particular world is needed.

This is Linde's "smorgasbord" picture (unpubl.). It rejects naturalness, but only on scales deeper than the Standard Model. It neatly explains why things look homogeneous in regions smaller than a domain, and why on scales less deep than the Standard Model physics does look natural; for inflation smooths out our region and tends to make it forget its particular initial conditions. Insulation begets naturalness on scales less fundamental than a domain.

This may be the end of the road for naturalness. One day the elaborate theories mentioned in the preceding section may look like the search for the significance of SU(3). Time will tell.

References

Coleman, S. 1975. Secret symmetry. In *Laws of Hadronic Matter*, ed. A. Zichichi, p. 141. Academic.

Dolgov, A., and Ya. Zeldovich. 1981. Cosmology and elementary particles. *Rev. Mod. Phys.* 53:1.

Gale, G. 1981. The anthropic principle. *Sci. Am.* 245(Dec.): 154.

Guth, A., and P. Steinhardt. 1984. The inflationary universe. *Sci. Am.* 250(May): 116.

Holton, G. 1973. *Thematic Origins of Scientific Thought.* Harvard Univ. Press.

Hooft, G. 't. 1980. Naturalness, chiral symmetry, and spontaneous chiral symmetry breaking. In *Recent Developments in Gauge Theory*, ed. G. 't Hooft et al., pp. 101–16. Plenum.

Jackson, J. 1975. *Classical Electrodynamics*, 2nd ed. Wiley.

Linde, A. Unpubl. Talk given at Harvard University, June 1983.

Merz, J. 1904. *A History of European Thought in the Nineteenth Century*, p. 350. London: Blackwood.

Peebles, P. 1971. *Physical Cosmology.* Princeton Univ. Press.

Schramm, D. 1983. The early universe and high-energy physics. *Phys. Today* 36:27.

Sulak, L. 1982. Waiting for the proton to decay. *Am. Sci.* 70:616.

Susskind, L. 1979. Dynamics of spontaneous symmetry breaking in Weinberg-Salam Theory. *Phys. Rev.* D20:2619.

Wilson, K. 1971. The renormalization group and strong interactions. *Phys. Rev.* D3:1840.