

Einstein and the search for unification

David Gross

Department of Physics, University of California, Santa Barbara, CA 93106, USA

Einstein spent the last thirty years of his life searching for a unified field theory. I discuss Einstein's attempts at unification. I examine his mistakes, ask why he went wrong, and wonder what might have happened if he had followed a slightly different route. I then discuss, very briefly, where we stand today in realizing Einstein's goals.

Keywords: Einstein, laws of nature, unification.

MY topic is at the heart of Einstein's scientific life, the search for a unified theory of nature. This was Einstein's main pursuit for more than half of his scientific career. Most contemporaries viewed his attempts as a waste of time, a total failure or, at best, premature. But today we look with some admiration at his foresight. Having understood by the middle 1970's, to a large extent, all the four forces of nature in the remarkable successful standard model, attention has returned to Einstein's dream of unifying all the forces with gravity. The goal of unification has been at the forefront of fundamental physics for the last three decades.

In this article I shall, fully aware of the ease of hindsight, discuss Einstein's goals, his attempts to unify general relativity and electromagnetism, and to include matter. I shall discuss his mistakes, ask why he went wrong, and wonder what might have happened if he had followed a slightly different route. As I am not a professional historian I can get away with murder. I shall then discuss, very briefly, where we stand today in realizing Einstein's goals.

For many physicists, certainly me, Einstein is both a hero and a model. He stated the goals of fundamental physics, that small part of physics that probes the frontiers of physics in a search for the underlying laws and principles of nature. Einstein was a superb epigramist, who could capture in a single sentence many deep thoughts.

Here is his definition of the goal of the physicist:

The supreme test of the physicist is to arrive at those universal laws of nature from which the cosmos can be built up by pure deduction.

I love this sentence. In one sentence Einstein asserts the strong reductionist view of nature: There exist universal, mathematical laws which can be deduced and from which

all the workings of the cosmos can (in principle) be deduced, starting from the elementary laws and building up.

Einstein, more than any other physicist, untroubled by either quantum uncertainty or classical complexity, believed in the possibility of a complete, perhaps final, theory of everything. He also believed that the fundamental laws and principles that would embody such a theory would be simple, powerful and beautiful. The 'old one', that Einstein often referred to, has exquisite taste.

This exciting goal, which I first learned of when I was thirteen by reading popular science books, seemed to me so exciting that I vowed to become a theoretical physicist. Although I certainly had no idea what that meant, I did know that I wanted to spend my life tackling the most fundamental questions of physics. This goal led me to elementary particle physics in the 1960's and to string theory in the 1980's. This goal motivated Einstein to spend the last thirty years of his life in a futile search for a unified theory of physics.

Physicists are an ambitious lot, but Einstein was the most ambitious of all. His demands of a fundamental theory were extremely strong. If a theory contained any arbitrary features or undetermined parameters then it was deficient, and the deficiency pointed the way to a deeper and more profound and more predictive theory. There should be no free parameters – no arbitrariness.

Nature, he stated with confidence, *is constituted so that it is possible to lay down such strong determined laws that within these laws only rationally, completely determined constants occur, not constants therefore that could be changed without completely destroying the theory.* This is a lofty goal, under threat nowadays from those who propose the Anthropic principle, whereby many of the fundamental constants of nature, even some of the laws, are environmental in nature and might be different in different parts of the universe. For me and for many others however, this remains the ultimate goal of physics, and a guiding principle. A theory that contains arbitrary parameters, or worst of all arbitrarily finely tuned parameters, is deficient.

After his enormous success at reconciling gravity with relativity, Einstein was troubled by the remaining arbitrariness of the theoretical scheme. First, the separate existence of gravitation and electromagnetism was unacceptable. According to his philosophy, electromagnetism must be unified with general relativity, so that one could not simply imagine that it did not exist. Furthermore, the existence of matter, the mass and the charge of the electron and the proton (the only elementary particles recognized back in the 1920s), were arbitrary features. One of the main goals

e-mail: gross@itp.ucsb.edu

of a unified theory should be to explain the existence and calculate the properties of matter.

Before passing to a discussion of Einstein's attempts at unification I wish to make a remark concerning his work on special relativity in 1905, whose centenary we celebrate this year. One of the most important aspects of this work was to revolutionize how we view symmetry. Principles of symmetry have dominated fundamental physics in the 20th century, starting with Einstein in 1905.

Until the twentieth century principles of symmetry played little conscious role in theoretical physics. The Greeks and others were fascinated by the symmetries of physical objects and believed that these would be mirrored in the structure of nature. Kepler attempted to impose his notions of symmetry on the motion of the planets. The laws of mechanics embodied symmetry principles, notably the principle of equivalence of inertial frames, or Galilean invariance.

The symmetries implied conservation laws. Although these conservation laws, especially those of momentum and energy, were regarded to be of fundamental importance, they were regarded as consequences of the dynamical laws of nature rather than as consequences of the symmetries that underlay these laws. Maxwell's equations, formulated in 1865, embodied both Lorentz invariance and gauge invariance. But these symmetries of electrodynamics were not fully appreciated for over forty years or more.

This situation changed dramatically in the twentieth century beginning with Einstein. Einstein's great advance in 1905 was to put symmetry first, to regard the symmetry principle as the primary feature of nature that constrains the allowable dynamical laws. Thus the transformation properties of the electromagnetic field were not to be derived from Maxwell's equations, as Lorentz did, but rather were consequences of relativistic invariance, and indeed largely dictate the form of Maxwell's equations. This is a profound change of attitude. Lorentz, who had derived the relativistic transformation laws from Maxwell's equations, must have felt that Einstein cheated. Einstein recognized the symmetry implicit in Maxwell's equations and elevated it to symmetry of space-time itself. This was the first instance of the *geometrization* of symmetry, and the beginning of the realization that symmetry is a primary feature of nature that constrains the allowed dynamical laws.

The traditional symmetries discovered in nature were global symmetries, transformations of a physical system in a way that is the same everywhere in space. Global symmetries are regularities of the laws of motion but are formulated in terms of physical events; the application of the symmetry transformation yields a different physical situation, but all observations are invariant under the transformation. Thus global rotations rotate the laboratory, including the observer and the physical apparatus, and all observations remain unchanged.

Gauge or local symmetry is of a totally different nature. Gauge symmetries are formulated only in terms of the laws of nature; the application of the symmetry transformation

merely changes our description of the same physical situation, does not lead to a different physical situation. Today we realize that local symmetry principles are very powerful – they dictate the form of the laws of nature.

In 1912–17 this point of view scored a spectacular success with Einstein's construction of general relativity. The principle of equivalence, a principle of local symmetry – the invariance of the laws of nature under local changes of the space-time coordinates – dictated the dynamics of gravity, of space-time itself. Fifty years later gauge theories, invariant under local symmetry transformations, not of space-time but of an internal space of particle labels, assumed a central position in the fundamental theories of nature. They provide the basis for the extremely successful standard model, a theory of the fundamental, non-gravitational forces of nature – the electromagnetic, weak and strong interactions.

Surprisingly Einstein did not follow the symmetry route. He did not, in his attempts to unify physics, search for extensions of the symmetries that he had promulgated. If he had he might very well have discovered non-Abelian gauge theory or perhaps even supersymmetry. Why not follow this route that has dominated theoretical speculation in the latter half of the 20th century? I think the reason was that Einstein was unaware of the phenomenon of symmetry breaking. All of the new symmetries discovered in the latter half of the 20th century, that are at the heart of the standard model of particle physics and attempts at unification, are approximate, or are broken spontaneously, or hidden by confinement. It was only in the 1960s, and early 1970s, that these mechanisms of symmetry breaking were elucidated and the possibility of imagining new symmetries, not directly manifest in the world, but still dictating the dynamics, was possible.

For Einstein the existence, the mass, the charge of the electron and the proton, the only elementary particles recognized back in the 1920s, were arbitrary features. One of the main goals of a unified theory should be to explain the existence and calculate the properties of matter. When he contemplated his equation he distinguished between the left-hand side of the equation, which was a beautiful consequence of the profound symmetry of general coordinate transformations, and captures the curvature of space-time; and the right-hand side, which was the source of curvature–mass, but had to be arbitrarily put in, with no principle to determine the properties of mass. As in politics Einstein greatly preferred the left to the right. To quote Einstein: *What appears certain to me, however, is that, in the foundations of any consistent field theory the particle concept must not appear in addition to the field concept. The whole theory must be based solely on partial differential equations and their singularity-free solutions.*

So Einstein's goals were to (i) Generalize general relativity to include electromagnetism. (ii) Eliminate the right hand side of his equations and deduce the existence of matter by constructing singularity free solutions that would describe stable lumps of energy. (iii) And finally, since he abhorred

the arbitrary nature of the quantum rules and their probabilistic interpretation, he hoped to deduce them from these non-singular solutions.

He imagined that the demand of lack of singularities in the solutions that would describe matter would lead to overdetermined equations, whose solutions would only exist for some, quantized values of physical parameters, say the radii of electron orbits. Thus he could imagine reproducing the Bohr model of the atom. The core of this program was to include electromagnetism and derive the existence of matter in the form of, what we call today, solitons. As Einstein understood, nonlinear equations can possess regular solutions that describe lumps of energy that do not dissipate. Thus one could start with the non-linear field equations of general relativity and find localized particles. This was his hope:

'If one had the field equation of the total field, one would be compelled to demand that the particles themselves would everywhere be describable as singularity free solutions of the completed field equations. Only then would the general theory of relativity be a complete theory.'

As far as I can tell, Einstein knew of no example of solitons or any toy model that exhibited his hopes. Nonetheless, flushed with the success of general relativity, with the faith that electromagnetism had to be unified, that matter needed a reason for its existence, he studied the equations and tried to modify them as well, with the hope of finding such solutions and with the dream that quantization of mass and charge, and even the quantum rules would emerge from overdetermination.

Among all of the extensions of general relativity considered and pursued by Einstein, the idea that the other forces of nature could be reflections of gravity in higher dimensions was the most innovative and enduring. It was not Einstein's idea, but rather that of Kaluza in 1922, significantly developed by Oscar Klein in 1926. Kaluza and Klein showed that if one assumed general relativity in five dimensions, where one dimension was curled up, the resulting theory would look like a four-dimensional theory of electromagnetism and gravity. Electromagnetism emerged as a consequence of gravity in five dimensions.

Einstein was immediately attracted to this idea and wrote to Kaluza – *'The idea of achieving (a unified field theory) by means of a five-dimensional cylinder world never dawned on me. At first glance I like your idea enormously.'* He held this paper for two years before submitting it to be published, probably because he was confused, as was Kaluza, as to whether the fifth dimension was real or not. Einstein returned again and again to this idea for over thirty years.

Einstein and Bergman in 1938 finally gave the best reasoning for taking the fifth dimension seriously, arguing that it is consistent with observation if it is sufficiently small. Klein had identified the momentum of particles moving around the fifth dimension as electric charge, which is quantized if one assumes the quantum mechanical rules

of momentum quantization on circle. In modern versions of Kaluza–Klein, as they appear in string theory, this scenario is greatly amplified. In string theory there are six or seven extra-spatial dimensions. One can imagine that these are curled up to form a small manifold, and remarkably such six dimension compactifications (achieved by solving the generalization of Einstein's equations in ten dimensions) can produce a world remarkably like our own, in which the shape of the extra dimensions determines the complete matter content and all the forces of nature, as seen by a four-dimensional observer.

Why did not Einstein consider higher dimensional spaces? Much later he did play, for a while, with an eight-dimensional universe, a kind of complexification of Minkowski space, an approach severely criticized by Pauli and rapidly dropped by Einstein. But why did he not search systematically for higher dimension theories? If he had done so he might have discovered non-Abelian gauge theories, much as Oscar Klein almost did in 1938. I do not know, but suspect that part of the reason was that Einstein by and large ignored the nuclear forces altogether. His goal was to incorporate electromagnetism together with gravity – for this one extra dimension sufficed.

Einstein never thought much of this quantization of electric charge. Perhaps he thought, as Klein tried, to turn this around and derive the quantum rules from the quantization of charge. But in any case Einstein's main goal was to find particles as non-singular solutions of his equations and thus turned immediately to trying to find non-singular solutions of Kaluza–Klein theory.

Over the years Einstein came back again and again to this problem and tried to find non-singular solutions of Kaluza–Klein theory. He published at least three papers in which he proved that such solutions do not exist, with ever increasing generality. The last of these was a paper published with Pauli, who spent some of the war years in Princeton. The remark made in this paper that: *When one tries to find a unified theory of the gravitational and electromagnetic fields, he cannot help feeling that there is some truth in Kaluza's five-dimensional theory*, expressed how much Einstein was attracted to this approach. He must have been incredibly disappointed that he could not find matter as solitons in this theory.

But Einstein was wrong. There do exist solitons, non-singular solutions of his equations in Kaluza–Klein theory, which behave as particles – magnetic monopoles, with quantized magnetic charge. These were discovered in the early 1980s, by Perry and me, and independently by Sorkin, when Kaluza–Klein theory was revived. In our paper we added a footnote pointing out that these solutions contradicted Einstein. The referees suggested that we remove the footnote since it was disrespectful. We, of course, refused, how could we resist.

I have wondered what would have happened if these solutions had been discovered back in the 1920s; they could have. It would have given an enormous boost to

Einstein's program, even though the solitons were magnetic and not electric, and very massive. But this did not happen and Einstein's attempts to find non-singular solutions failed, as did his attempts to construct satisfactory unified theories.

After sometime in the late 1920s Einstein became more and more isolated from the mainstream of fundamental physics. To a large extent this was due to his attitude towards quantum mechanics, the field to which he had made so many revolutionary contributions. Einstein, who understood better than most the implications of the emerging interpretations of quantum mechanics, could never accept it as a final theory of physics. He had no doubt that it worked, that it was a successful interim theory of physics, but he was convinced that it would be eventually replaced by a deeper, deterministic theory. His main hope in this regard seems to have been the hope that by demanding singularity free solutions of the nonlinear equations of general relativity one would get an overdetermined system of equations that would lead to quantization conditions.

Because of his opposition to quantum mechanics he allowed himself to ignore most of the important developments in fundamental physics for over twenty five years, as he himself admitted in 1954, '*I must seem like an ostrich who buries its head in the relativistic sand in order not to face the evil quanta*'. If there is one thing that I fault Einstein for, it is his lack of interest in the development of quantum field theory. To be sure many of the inventors of quantum field theory were soon to abandon it when faced with ultraviolet divergences, but it is hard to understand how Einstein, could not have been impressed with the successes of the marriage of his children quantum mechanics and special relativity. The Dirac equation and quantum electrodynamics had remarkable successes, especially the prediction of anti-particles. How could Einstein not have been impressed?

The only way to understand this is that general relativity was so important to him as to eclipse everything else. As Pauli remarked: '*If we would have presented Einstein with a synthesis of his general relativity and the quantum theory – then the discussion with him would have been considerably easier*'. But since general relativity and quantum mechanics seemed so incompatible, a situation that continued until quite recently, he felt free to ignore the exciting advances that were made in special relativistic quantum mechanics.

I turn now to the situation today, or more precisely thirty years ago, after the completion of the standard model of elementary particle physics, where we now have direct evidence for the unification of all forces dreamed by Einstein.

One of the most important implications of asymptotic freedom is the insight it gave into the unification of all the forces of nature. Almost immediately after the discovery of asymptotic freedom and the proposal of quantum chromodynamics, the first attempts were made to unify all the forces. This was natural, given that one was using very similar theories to describe all the known interactions.

The apparently insurmountable barrier to unification, namely the large difference in the strength of the strong and the electro-weak force, was seen to be a low energy phenomenon. Since the strong force decreases with increasing energy, all forces could have a common origin at very high energy. Indeed the couplings run in such a way as to merge about 10^{14} to 10^{16} GeV, close to the point where gravity becomes equally strong. This is our most direct clue as to where the next threshold of fundamental physics lies, and hints that at this immense energy all the forces of nature, including gravity, are unified.

In more recent times this extrapolation has greatly improved, due to the beautiful measurements of many experimenters and the hard work done by many theorists. Now the forces all meet only if we hypothesize a new space–time symmetry-supersymmetry – and if this new symmetry is broken at reasonably low energy; increasing hopes that a new super-world will be revealed at the Large Hadron Collider, soon to be completed at CERN. Supersymmetry is a beautiful, natural and unique extension of relativistic and general relativistic symmetries of nature. Einstein would, if he had studied it, have loved it. It can be thought of as the space–time symmetries of super-space, a space–time with extra dimensions. But the extra dimensions, here denoted collectively by \mathbf{q} , are measured with anti-commuting numbers. These are generalizations of ordinary real numbers, much as imaginary or complex numbers are; numbers that anti-commute, so that multiplication depends on the order, thus $\mathbf{q}_1\mathbf{q}_2 = -\mathbf{q}_2\mathbf{q}_1$. If it is hard to imagine a space of four or more dimensions, super-space is even weirder, but totally mathematically consistent. A theory formulated in super-space, and invariant under transformations or rotations of super-space, has many beautiful and appealing features. Supersymmetric extensions of the standard model can solve many important problems, such as why is there this enormous disparity, at low energy, between the strength of the gravitational force and the other forces of nature. The discovery of supersymmetry, which we all hope and some expect in a few years from now at the Large Hadron Collider, would be tantamount to the discovery of quantum dimensions of space–time.

Perhaps the most important feature of the extrapolation of the standard models forces is that the energy at which they appear to unify is very close, if not identical, to the point at which gravity becomes equally strong. This indicates that the next stage of unification should include, as Einstein expected, unification of the non-gravitational forces and gravity. It is an important clue to that unification since it is not easy to quantize general relativity. A straightforward quantization of Einstein's theory does not work; the quantum fluctuations of the metric, at the characteristic distance scale of gravity, where the force becomes strong are too violent and uncontrollable. It seems inescapable that Einstein's theory is only an effective theory, adequate at long distances, but to be replaced by a more fundamental theory at the Planck scale of 10^{-33} cm.

Luckily such an extension of general relativity is available – string theory. String theory was not invented to describe gravity; instead it originated in an attempt to describe the strong interactions, wherein mesons can be thought of as open strings with quarks at their ends. The fact that the theory automatically described closed strings as well, and that closed strings invariably produced gravitons and gravity, and that the resulting quantum theory of gravity was finite and consistent is one of the most appealing aspects of this theory. String theory is a theory in development. We have learned much about this theory in the last decades, but much more remains. What has been achieved so far?

First, string theory is a consistent logical extension of the conceptual framework of fundamental physics. Such an extension is not easy and it is rare.

Second, string theory provides us for the first time with a consistent and finite quantum theory of gravity. This not only proves that quantum mechanics and general relativity are mutually compatible, it also provides us with the tools to explore many of the paradoxical issues that arise when the metric of space–time is quantized. Already string theory has clarified many of the mysteries of black holes. Thus the suspicion raised by Hawking as to whether black holes indicate the loss of information in fundamental physics has been dispelled, even to the point where Hawking himself has agreed that information is not lost in the process of formation and evaporation of black holes.

Finally, string theory has a rich structure that could yield a theory that unifies all of the forces of nature and explain all the constituents of matter. It automatically contains gravity as well as the gauge theories of the standard model. Certain of its four-dimensional compactifications give rise to low energy dynamics that is remarkably close to the standard model.

But string theory is still in the process of development, and although it has produced many surprises and lessons it still has not broken dramatically with the conceptual framework of relativistic quantum field theory. Many of us believe that ultimately string theory will give rise to a revolution in physics, as important as the two revolutions that took place in the 20th century, relativity and quantum mechanics. These revolutions are associated with two of the three fundamental dimensionful parameters of nature, the velocity of light and Planck’s constant. The revolution in string theory presumably has to do with Newton’s constant, that defines a length, the Planck length of 10^{-33} cm. String theory, I believe, will ultimately modify in a fundamental way our concepts at distances of order this length.

Where will the revolution take place? I believe that it will involve our understanding of the nature of space–time, a subject dear to Einstein’s heart. To quote some leading string theorists:

Space and time may be doomed. – E. Witten

I am almost certain that space and time are illusions.
– N. Seiberg

The notion of space–time is clearly something we’re going to have to give up. – A. Strominger

The real change that’s around the corner is in the way we think about space and time. We haven’t come to grips with what Einstein taught us. But that’s coming. And that will make the world around us stranger than any of us can imagine. – D. Gross

Why is space–time doomed? There are many reasons, among which: In string theory we can change the dimension of space–time by changing the strength of the string force. Thus, the so-called II-A string theory, which semi-classically describes closed strings moving in ten-dimensional flat space for very weak coupling is dual for strong coupling to a theory, called M-theory, that at low energies is described by eleven dimensional supergravity. By increasing the string coupling we can grow an extra dimension. How can the spatial continuum be fundamental if the number of spatial dimensions can be so changed?

We can continuously tear the fabric of space. Thus a string theory solution that describes strings moving on a background wherein some of the spatial dimensions are compactified on a manifold M_1 can be continuously deformed, by varying some of the parameters of the solution, to one that describes the strings moving on a background M_2 of different topology. In between there is no such simple description of the solution as strings moving on a geometric background, but the deformation is continuous and the strings do not mind at all that the fabric of space has been torn so as to modify the topology. Again this suggests that the spatial continuum cannot be fundamental if its topology can be changed in this smooth fashion.

On the other hand in string theory we cannot probe arbitrarily small distances. In string theory we can ask what is the smallest distance that can operationally be explored, analysing (as Heisenberg did in the case of quantum mechanics) how a microscope works. In string theory the light rays of a microscope are really strings. Consequently, as we increase the energy of the light, so as to overcome the quantum mechanical uncertainty in the measurement of distance, the strings expand and prevent us from resolving arbitrarily small distances. The minimum distance that we can explore is, not surprisingly, of order the Planck length.

We also cannot squeeze spatial volumes to zero size. If one of the spatial dimensions is compactified to form a circle of radius R , it turns out that string theory in this background is identical to string theory in a background where the radius of this circle is $1/R$ (in Planckian units). Thus if we try to squeeze this dimension and reduce R to zero, we find that the more natural description is in terms of the dual theory, and the minimal size of the compact circle is finite and of order the Planck length.

These phenomena suggest that there is no operational meaning to distances smaller than the Planck length, that the spatial continuum should be replaced by something else. I believe that space for sure, and presumably time as well,

will be emergent. We already have many hints and examples where space is an emergent concept. These include the famous AdS/CFT duality, wherein string theory in ten dimensions, with a background geometry of five dimensional Anti-DeSitter space times a five sphere, is dual to supersymmetric gauge in flat four-dimensional space-time. Six spatial dimensions emerge from the gauge theory description, together with gravity. We have no understanding, however, what it would mean that time itself would be an emergent concept.

I like to depict our confusion in poetic form. Democritus expressed 2500 years ago the atomic hypothesis in the following verse:

*By convention there is color,
by convention sweetness,
by convention bitterness,
But in reality there are atoms and space.*

I say: We are convinced that

*By convention there is space,
By convention there is time,
But in reality there is . . .*

The problem is that I do not know how to finish the verse.

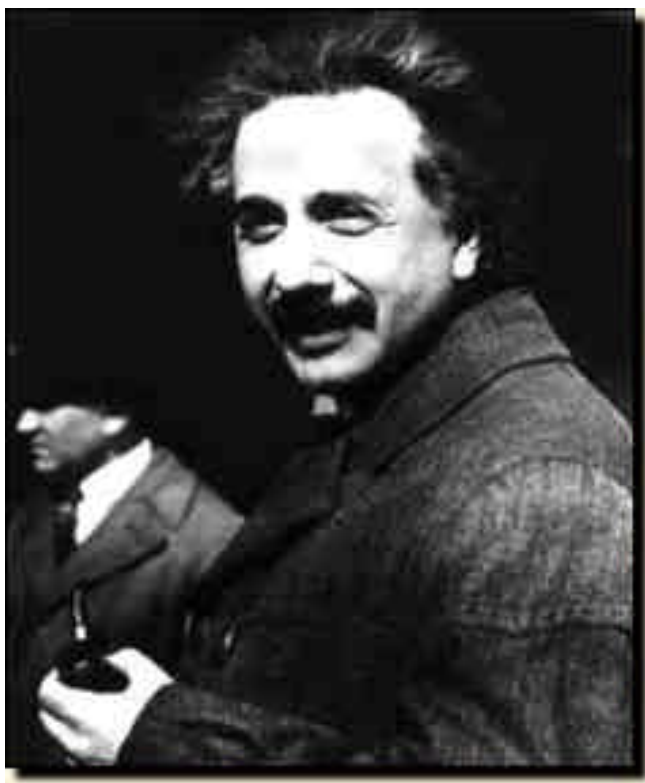
So did Einstein go wrong in the latter part of his life? The answer is both yes and no.

Yes, he refused to accept quantum mechanics. He ignored the developments in nuclear and particle physics. These mistakes ensured his failure, but they are quite understandable and forgivable.

No, he knew that gravity must be unified with the other forces. And this we too know today is the central issue in fundamental physics.

And for those of us faced with the fact that we cannot yet directly probe the Planck scale, he believed in the possibility of successful speculative theory. As Einstein stated: *'The successful attempt to derive delicate laws of nature, along a purely mental path, by following a belief in the formal unity of the structure of reality, encourages continuation in this speculative direction, the dangers of which everyone vividly must keep in sight who dares follow it'*.

To all physicists, but especially to those working in speculative areas, Einstein remains an inspiration for his foresight, and his unyielding determination and courage.



Einstein. Photo courtesy: AIP Emilio Segre Archives.