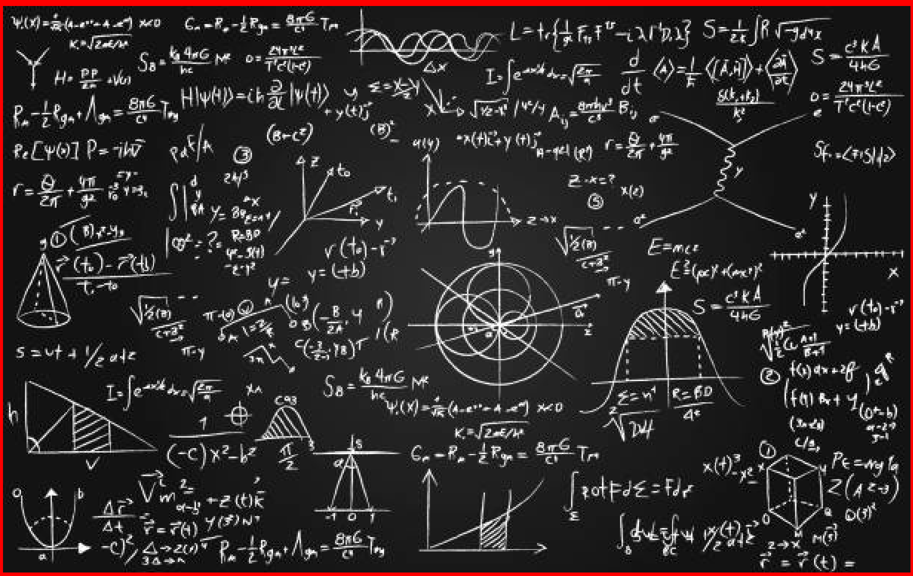
Discoveries due to the observation of mathematics

Author:
Riccardo Fantoni



December 13, 2022

Preface

In this book ...

...
Trieste, December 13, 2022

Foreword

In the scientific method usually we observe two kinds of processes going in opposite directions. The process where starting from the observation of nature one develops the mathematical model of the given phenomenon, which often stimulated the development of mathematics itself. And the opposite feedback process where starting from the mathematical constraints, evolution or solution of a given model one develops the experiment necessary to observe in a laboratory or in nature the predicted phenomenon, which often stimulated the development of new technologies. Most often this second kind of process have had simply the scope of an imitation of nature in a laboratory, that is the reproduction of natural phenomena using the techniques at our disposal. More rarely it allowed to uncover, "discover", new phenomena not previously observed in a laboratory or in nature. In this book we collect some examples of the successful realization of this kind of discoveries occurred in the history of physics. We give 7 notorious examples which can be read one each day. So that the first part of the book can be read in one week. The book is intended both for the lay reader and for the more educated one. We couldn't avoid to use some equations and give for granted some basic knowledge in mathematics and physics. Even if we tried to extract only the strictly necessary equations to understand the mathematical constraints leading to the discovery, we found nevertheless necessary to show them because of their beauty and profound scientific meaning. The book is written so that it can be fully understood by a good graduate student in physics. But the less educated reader should not be scared by the equations and should try to grasp the meaning from the various descriptive and historical information surrounding them.

The second part of the book, the last two chapters, deals with the complex relationship between mathematics, the arts and philosophy and about some ontological and theological problems, like the anthropic principle, raised by the existence of mathematics as an exact science and physics as a basic, fundamental, hard, and empirical science. Is the beauty of mathematics a fruit of God or just of the human beings? This part of the book has a more popularization character and unlike the first part contains very few equations.

As a physics student I had the chance to study the fundamentals of quantum mechanics in my third year at the University of Pisa, under the supervision of Prof. Pietro Menotti, and of relativistic quantum mechanics in my fourth and last year at the University of Pisa, under the supervision of Prof. Adriano Di Giacomo. Soon after my graduation at the University of Pisa I qualified to become a graduate student at the University of Illinois at Urbana-Champaign, where I had the chance to learn the path integral Monte Carlo method, under the supervision of Prof. David Ceperley, and general relativity, under the supervision of Prof. Stuart Shapiro. Other persons that had a great role in my physics knowledge has been Prof. Mario Tosi from the "Scuola Normale Superiore di Pisa" who taught me statistical mechanics, many body theory, and the theory of Coulomb liquids and Prof. Giorgio Pastore from the University of Trieste who deepened my knowledge on classical and quantum simple liquids. These has been only the main actors responsible for my growth as a physicist but I could mention many others.

As a physicist I learned the mathematics the other way around, starting from physics courses and studying the mathematics necessary for the understanding of such courses. I believe that in doing so even if I didn't acquire a deep vertical view of mathematics I nonetheless acquired a wide horizontal view of it. Also as a low energy (condensed and soft matter) physicist I don't believe that our entire physical reality is a mathematical structure, having no properties besides mathematical properties. For example how can we explain the difference between animate and inanimate nature, between a butterfly and a stone? How do we explain life? Giorgio Parisi, Nobel prize for physics 2021 writes in his book "In un volo di storni" [1] that "In physics and mathematics the disproportion between the effort to understand something new for the first time and the simplicity and naturalness of the result once the various steps have been completed is striking. In the finished product, in the sciences as in poetry, there is no trace of the fatigue of the creative process and of the doubts and hesitations that accompany it." and then "[...] on complexity, on the idea that in the presence of a large number of interacting agents (molecules, genes, cells, animals, species, depending on the level of discussion) there are new phenomena that emerge as an effect of collective interaction." A complex system like a cat differs profoundly by any collection of elementary particles. In fact whereas two of the latter always interact in the same way no matter how many times they interact, two cats after a first interaction will change forever their properties (they will learn something new) and any subsequent interaction between them will be necessarily influenced by their previous interaction.

A Grand Unified Theory (GUT) is a model in particle physics in which, at high energies, the three gauge interactions of the Standard Model comprising the electromagnetic, weak, and strong forces are merged into a single force. Although this unified force has not been directly observed, the many GUT models theorize its existence. If unification of these three interactions is possible, it raises the possibility that there was a grand unification epoch in the very early universe in which these three fundamental interactions were not yet distinct. Unifying gravity with the electronuclear interaction would provide a more comprehensive Theory Of Everything (TOE) rather than a Grand Unified Theory. Thus, GUTs are often seen as an intermediate step towards a TOE.

A TOE, final theory, ultimate theory, unified field theory or master theory is a hypothetical, singular, all-encompassing, coherent theoretical framework of physics that fully explains and links together all physical aspects of the universe. Finding a theory of everything is one of the major unsolved problems in physics. String theory and M-theory (a theory in physics that unifies all consistent versions of superstring theory) have been proposed as theories of everything (they propose an additional 6 or 7 dimensions of hyperspace + the 4 common dimensions = 10D or 11D spacetime).

Over the past few centuries, two theoretical frameworks have been developed that together, most closely resemble a theory of everything. These two theories upon which all modern physics rests are general relativity and quantum mechanics. General relativity is a theoretical framework that only focuses on gravity for understanding the universe in regions of both large scale and high mass: planets, stars, galaxies, clusters of galaxies etc. On the other hand, quantum mechanics is a theoretical framework that only focuses on three non-gravitational forces for understanding the universe in regions of both very small scale and low mass: subatomic particles, atoms, molecules, etc. Quantum mechanics successfully implemented the Standard Model that describes the three non-gravitational forces: strong nuclear, weak nuclear, and electromagnetic force – as well as all observed elementary particles.

General relativity and quantum mechanics have been repeatedly validated in their separate fields of relevance. Since the usual domains of applicability of general relativity and quantum mechanics are so different, most situations require that only one of the two theories be used. The two theories are considered incompatible in regions of extremely small scale – the Planck scale – such as those that exist within a black hole or during the beginning stages of the universe (i.e., the moment immediately following the Big Bang). To resolve the incompatibility, a theoretical framework revealing a deeper underlying reality, unifying gravity with the other three interactions, must be discovered to harmoniously integrate the realms of general relativity and quantum mechanics into a seamless whole: the theory of everything is a single theory that, in principle, is capable of describing all physical phenomena in this universe.

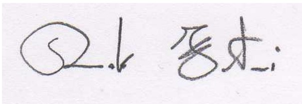
In pursuit of this goal, quantum gravity has become one area of active research. One example is string theory, which evolved into a candidate for the theory of everything, but not without drawbacks (most notably, its apparent lack of currently testable predictions) and controversy. String theory posits that at the beginning of the universe (up to 10^{-43} seconds after the Big Bang), the four fundamental forces were once a single fundamental force. According to string theory, every particle in the universe, at its most ultramicroscopic level (Planck length), consists of varying combinations of vibrating strings (or strands) with preferred patterns of vibration. String theory further claims that it is through these specific oscillatory patterns of strings that a particle of unique mass and force charge is created (that is to say, the electron is a type of string that vibrates one way, while the up quark is a type of string vibrating another way, and so forth).

The main criticism that I find to this field of research is that if a TOE really existed then it would be able to explain also our next thoughts as humans, since they would be predicted by the same theory, so at the very moment in which it would be found, at the moment in which we would write its last line, the entire man kind will 'die' since it would have no other reason to live for!

Certainly mathematics pervades many aspects of our life and not only its scientific characters. Of course it influences philosophy in logic but also the arts not only the visual ones. Many mathematicians, like for example G. H. Hardy in his 1940 essay "A Mathematician's Apology", believe that mathematics has value independent of possible applications and often locate this value in its beauty and elegance and not dullness and technicality. Giorgio Parisi writes [1] "But what exactly is mathematics? It is a science that operates on symbols purified of any concrete meaning; as Bertrand Russell says, "mathematics is that science that does not know what it is talking about." The reason is simple: if we say that 2 + 3 is 5 – it can be 2 phone calls + 3 phone calls making 5 phone calls or 2 cows + 3 cows making 5 cows –, we have no idea what the 5 'objects' in question are. This is true at an extremely low level of abstraction and becomes more and more relevant as we move towards more abstract concepts. Mathematical objects are purified of any sensible appearance and therefore mathematical propositions, like logical propositions, have a universal value."

I would like to thank my father Stefano Fantoni and my daughter Alice Fantoni for several discussions. I would like to thank my wife Laure Gouba for carefully proofreading the whole book. I would also like to thank the many colleagues from both the school system and the academic one, in particular the professor of religion Vitaliano Raimo in the school where I also teach as a mathematics professor, has been very important in the writing of the very last Section.

In the book ©Wikipedia articles have been used for background information, as a reference for correct terminology and search terms, and as a starting point for further research.



Trieste, December 13, 2022

Contents

Chapter 1

Introduction

In science we use mathematics as a tool to interpret the natural world.

In the scientific method put forward by Galileo Galilei [2] in the scientific revolution of the XVIII century, which rediscovers the Hellenistic revolution of Euclid [3], the experiment or the observation of nature is given priority over the development of a scientific advancement. One starts from the experiment and only later formulates the mathematical model that explains it. Any theory which is not substantiated by an experimental proof is void of any scientific value.

The scientific method has been so successful so far that we have been building ever more complex and expensive laboratories for our experiments with all the consequent technological advancement that contributed to make our lives better. And this has been justified by the predictive power of theoretical mathematical models. In his autobiography [4] Nikola Tesla writes about his magnifying Transmitter, the modern Wi-Fi, and the consequences of technological progress:

“My belief is firm in a law of compensation. The true rewards are ever in proportion to the labour and sacrifices made. This is one of the reasons why I feel certain that of all my inventions, the magnifying Transmitter will prove most important and valuable to future generations. I am prompted to this prediction, not so much by thoughts of the commercial and industrial revolution which it will surely bring about, but of the humanization consequences of the many achievements it makes possible. Considerations of mere utility weigh little in the balance against the higher benefits of civilisation. We are confronted with portentous problems which can not be solved just by providing for our material existence, however abundantly. On the contrary, progress in this direction is fraught with hazards and perils not less menacing than those born from want and suffering. If we were to release the energy of atoms or discover some other way of developing cheap and unlimited power at any point on the globe, this accomplishment, instead of being a blessing, might bring disaster to mankind in giving rise to dissension and anarchy, which would ultimately result in the enthronement of the hated regime of force. The greatest good will come from technical improvements tending to unification and harmony, and my wireless transmitter is preeminently such. By its means, the human voice and likeness will be reproduced everywhere and factories driven thousands of miles from waterfalls furnishing power. Aerial machines will be propelled around the earth without a stop and the sun’s energy controlled to create lakes and rivers for motive purposes and transformation of arid deserts into fertile land. Its introduction for telegraphic, telephonic and similar uses, will automatically cut out the statics and all other interferences which at present, impose narrow limits to the application of the wireless.”

The trial and error process to interpret nature has been most often an imitation game, as for example in the development of the flight theory, but in some occasions we witnessed to *discoveries*, where simply by firmly trusting, *observing*, the mathematics of a theoretical model, but without any other clue from nature, we anticipated the experiment which uncovered the predicted phenomena in nature, in some cases by centuries.

In this book we want to describe some of the more striking examples for the predictive power of the scientific method. Cases in which our faith in mathematics has led to true discoveries of new phenomena, which couldn't be imitated simply because they could not be observed in nature at the time of the discovery.

It is important to realize that since ancient times in place of real experiments carried out in a laboratory one can produce *thought experiments* carried out in one imagination and only much more recently, from the advent of the new data science and the predictive algorithms of artificial intelligence, one can produce *computer experiments*, or numerical simulations which add, at all effects, a third dimension to the dualism theory-experiment.

Mathematics provides a compact and exact language used to describe the order in nature. This was noted and advocated by Pythagoras, Plato, Galileo, and Newton.

Physics uses mathematics to organize and formulate experimental results. From those results, precise or estimated solutions are obtained, or quantitative results, from which new predictions can be made and experimentally confirmed or negated. The results from physics experiments are numerical data, with their units of measure and estimates of the errors in the measurements. Technologies based on mathematics, like computation have made computational physics an active area of research.

The distinction between mathematics and physics is clear-cut, but not always obvious, especially in mathematical physics. Ontology is a prerequisite for physics, but not for mathematics. It means physics is ultimately concerned with descriptions of the real world, while mathematics is concerned with abstract patterns, even beyond the real world (or at least beyond the observable world). Thus physics statements are synthetic, while mathematical statements are analytic. Mathematics contains hypotheses, while physics contains theories. Mathematics statements have to be only logically true, while predictions of physics statements must match observed and experimental data. It is certainly true that we developed mathematics as a consequence of our need to interpret nature and most of the mathematics advancements are consequences of this need, but once created mathematics started to develop also by its own not always giving useful instruments for the physical interpretation of nature. As for any language also mathematics is able to describe a physical theory more or less faithfully. Some of its statements can be proven to be physically relevant (with some physical foundation) and some others to be physically absurd (because violating some generally accepted physical principle, like for the tachyons described in Section 2.3). Moreover with the development of technology and of instruments of measure we have been able to observe an ever increasing number of new phenomena not previously observable and their interpretation usually required the development of new mathematical tools. Even if this has been the more frequent scenario in the history of science, there has been notorious examples where the contrary happened. Namely, the mathematics observation anticipated the laboratory observation of new predicted phenomena. Whenever this has happened in history we witnessed to the most striking success of the scientific method. This is what the book is about. We may then, of course, ask the question whether mathematics is only a product of our brain or it is inherent in the natural world surrounding us as a whole. In other terms, does mathematics exists since the creation of the universe or does it exist only since the advent of the human species on our planet? This will be discussed in Section 10.3.

The distinction is clear-cut, but not always obvious. For example, mathematical physics is the application of mathematics in physics. Its methods are mathematical, but its subject is physical. The problems in this field start with a "mathematical model of a physical situation" (system) and a "mathematical description of a physical law" that will be applied to that system.

Pure physics is a branch of fundamental science (also called basic science). Physics is also called "*the* fundamental science" because all branches of natural science like chemistry, astronomy, geology, and biology are constrained by laws of physics (see Fig. 1.1). Similarly, chemistry is often called the central science because of its role in linking the physical sciences. For example, chemistry studies properties, structures, and reactions of matter (chemistry's focus on the molecular and atomic scale distinguishes it from physics). Structures are formed because particles exert electrical forces on each other, properties include physical characteristics of given substances, and reactions are bound by laws of physics, like conservation of energy, mass, and charge. Physics is applied in industries like engineering and medicine.

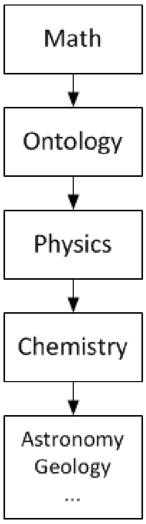


Figure 1.1: Mathematics and ontology are used in physics. Physics is used in chemistry and cosmology. Chemistry is used in biology. Biology is used in sociology ...

In the first part of the book we therefore concentrated in discoveries in physics that stemmed from mathematics (see Fig. 1.2) first than the observation of the natural world. Contemporary research in physics can be broadly divided into nuclear and particle physics; condensed matter physics; atomic, molecular, and optical physics; astrophysics; and applied physics. Most of our examples will be drawn from particle physics and astrophysics which can be thought as contained in high energy physics or more generally from those field of physics where the experiment has been for long time out of reach due to technological limitations. It is in fact in these context that mathematics, rather than observation through measure of the natural world, has been the real propulsive thrust towards new discoveries. The "faith" in the mathematical consequences drained from the mathematical model of the real world, like a third central "eye" on the nature in which we live has produced the conception and later realization of colossal experimental laboratories (like the particle accelerators of CERN for the detection of new particles or more recently the interferometers of VIRGO and LIGO for the detection of gravitational waves) which required the use of the most modern and sophisticated technologies and allowed their own push-forward, so important for the consequences and improvements in our everyday lives. Exactly like in the past the faith in God has been the driving force for the realization of magnificent cathedrals or

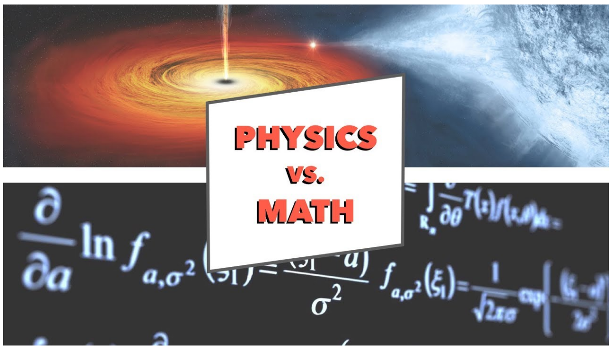


Figure 1.2: Mathematics vs. Physics.

tombs.

In physics we can say that we witnessed to three main scientific revolutions [5]: The first one has been the *Copernican revolution* [6] whose heliocentric theory of 1543 offered an alternative model of the universe to Ptolemy's geocentric system, which had been widely accepted since ancient times. This marked the beginning of the 'Scientific Renaissance', that was focused on the recovery of the knowledge of the ancients. This is generally considered to have ended in 1632 with publication of Galileo's Dialogue Concerning the Two Chief World Systems [2]. The completion of the Scientific Revolution is attributed to the "grand synthesis" of Isaac Newton's 1687 Principia [7]. The work formulated the laws of motion and universal gravitation, thereby completing the synthesis of a new cosmology. The second one has been the advent of the *quantum theory* at the beginning of the 20th century. And, at the same time, the third one has been the advent of the *theory of general relativity*. The philosopher Thomas S. Kuhn [5] argued for an episodic model describing such events in which periods of conceptual continuity where there is cumulative progress, which Kuhn referred to as periods of "normal science", were interrupted by periods of revolutionary science. The discovery of "anomalies" during revolutions in science leads to new paradigms. New paradigms then ask new questions of old data, move beyond the mere "puzzle-solving" of the previous paradigm, change the rules of the game and the "map" directing new research.

Usually in parallel to a scientific revolution we have mathematical progress. For example Newton was considered as one of the initiators of *infinitesimal calculus*. Quantum mechanics used the notion of the *Hilbert space*. And general relativity used the notion of *Riemannian geometry*.

Often we talk of the beauty of a mathematical equation. One notorious example is Euler's identity

$$e^{i\pi} + 1 = 0\,. \tag{1.1}$$

Three of the basic arithmetic operations occur exactly once each: addition, multiplication, and exponentiation. The identity also links five fundamental mathematical constants: (1) the number 0, the additive identity; (2) the number 1, the multiplicative identity; (3) the number π ($\pi = 3.14159265\ldots$), the fundamental circle constant; (4) the number e ($e = 2.71828182\ldots$), also known as Napier's number, which is the basis of the natural exponential function, the only function whose derivative is the function itself; (5) The number i, the imaginary unit of the complex numbers. The physicist Richard Feynman called such equation "our jewel" and "the most remarkable formula in mathematics". The aesthetic pleasure typically derived from the abstractness, purity, simplicity, elegance, depth or orderliness of mathematics was often recognized in *beauty* by many.

In this book we tried to use only the strictly necessary mathematical equations to illustrate the mathematical observation leading to the scientific discovery. Many of these equations are therefore not only beautiful but also with a profound meaning.

Often a mathematical equation cannot be solved by an expression involving only the coefficients of the equation, and the operations of addition, subtraction, multiplication, division, and nth root extraction, i.e. by a solution in radicals. A notorious example is given by polynomial equations of degree five or higher with arbitrary coefficients. the Abel-Ruffini theorem (also known as Abel's impossibility theorem) states that there is no such solution in this case. The theorem is named after Paolo Ruffini, who made an incomplete proof in 1799 [8], (which was refined and completed in 1813 and accepted by Cauchy) and Niels Henrik Abel, who provided a proof in 1824. Polynomial equations of degree two can be solved with the quadratic formula, which has been known since antiquity. Similarly the cubic formula for degree three, and the quartic formula for degree four, were found during the 16th century. At that time a fundamental problem was whether equations of higher degree could be solved in a similar way. The fact that every polynomial equation of positive degree has solutions, possibly non-real, was asserted during the 17th century, but completely proved only at the beginning of the 19th century. This is the fundamental theorem of algebra, which does not provide any tool for computing exactly the solutions, although Newton's method allows approximating the solutions to any desired accuracy. From the 16th century to beginning of the 19th century, the main problem of algebra was to search for a formula for the solutions of polynomial equations of degree five and higher, hence the name the "fundamental theorem of algebra". This meant a solution in radicals. The Abel-Ruffini theorem proves that this is impossible. Soon after Abel's publication of its proof, Évariste Galois introduced a theory, now called Galois theory that allows deciding, for any given equation, whether it is solvable in radicals.

This shows the necessity of *numerical analysis*. This is the study of algorithms that use numerical approximation (as opposed to symbolic manipulations) for the problems of mathematical analysis (as distinguished from discrete mathematics). Current growth in computing power has enabled the use of more complex numerical analysis, providing detailed and realistic mathematical models in science, engineering and even the arts. Before modern computers, numerical methods often relied on hand interpolation formulas, using data from large printed tables. Since the mid 20th century, computers calculate the required functions instead, but many of the same formulas continue to be used in software algorithms. The numerical point of view goes back to the earliest mathematical writings. A tablet from the Yale Babylonian Collection (YBC 7289 see Fig. 1.3), gives a sexagesimal numerical approximation of the square root of 2, the length of the diagonal in a unit square. Numerical analysis continues this long tradition: rather than giving exact symbolic answers translated into digits, approximate solutions within specified error bounds are used.

A special role in mathematics is played by *linearity*. This is the property of a mathematical relationship (function) that can be graphically represented as a straight line. Generalized for functions in more than one dimension, linearity means the property of a function of being compatible with addition (additive) and scaling (homogeneous), also known as the superposition principle. The concept of linearity can be extended to linear operators. Important examples of linear operators include the derivative considered as a differential operator, and other operators constructed from it, such as the Laplacian. When a differential equation can be expressed in linear form, it can generally be solved by breaking the equation up into smaller pieces, solving each of those pieces, and summing the solutions. Linear algebra is the branch of mathematics concerned with the study of vectors, vector spaces (also called 'linear spaces'), Hilbert (and more

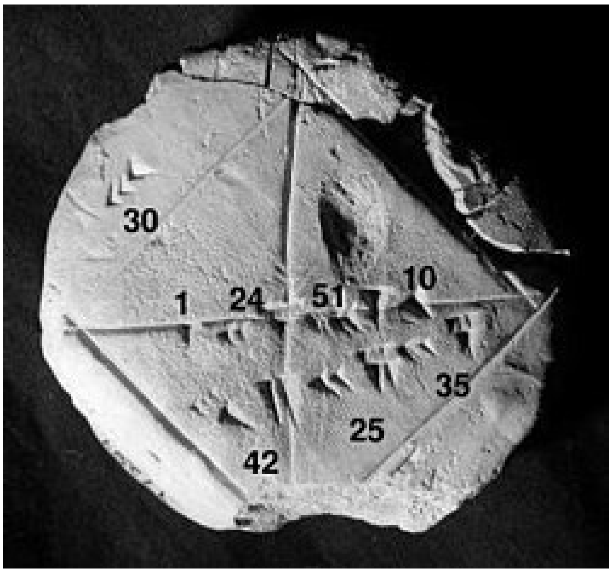


Figure 1.3: Babylonian clay tablet YBC 7289 (c. 1800–1600 BC) with annotations. The approximation of the square root of 2 in four sexagesimal figures, which is about six decimal figures: $1+24/60+51/602+10/603 = 1.41421296\ldots$.

generally Banach) spaces, linear transformations (also called 'linear maps'), and systems of linear equations.

Quite generally we can say that a problem that can be studied claiming linearity is usually much simpler to solve than one that can not. And most often non-linear problems require the use of numerical methods.

Mathematics (from Ancient Greek $\mu\alpha\theta\eta\mu\alpha$ 'knowledge, study, learning') is also made up of several sub-disciplines. One can clearly distinguish *geometry* from *analysis* where the former usually requires visual-intuitive skills whereas the latter calculating-pedantic ones. Geometry (from Ancient Greek $\gamma\epsilon\omega\mu\epsilon\tau\rho\iota\alpha$ 'land measurement'; from $\gamma\eta$ 'earth, land', and $\mu\epsilon\tau\rho o\nu$ 'a measure') is, with arithmetic, one of the oldest branches of mathematics. It is concerned with properties of space that are related with distance, shape, size, and relative position of figures. Until the 19th century, geometry was almost exclusively devoted to Euclidean geometry, which includes the notions of point, line, plane, distance, angle, surface, and curve, as fundamental concepts. During the 19th century several discoveries enlarged dramatically the scope of geometry.

> “Qual è il geometra che tutto s'affige
> per misurar lo cerchio, e non ritrova,
> pensando, quel principio ond'elli indige,
> tal era io a quella vista nova.”
>
> (Dante [9], Paradiso, XXXIII, 133-136)

One of the oldest such discoveries is Gauss' Theorema Egregium ("remarkable theorem") that asserts roughly that the Gaussian curvature of a surface is independent from any specific embedding in a Euclidean space. This implies that surfaces can be studied intrinsically, that is, as stand-alone spaces, and has been expanded into the theory of manifolds and Riemannian geometry. Later in the 19th century, it appeared that geometries without the parallel postulate (non-Euclidean geometries) can be developed without introducing any contradiction. The geometry that underlies general relativity is a famous application of non-Euclidean geometry.

Mathematics is essential in many fields of knowledge. Some areas of mathematics are developed in direct correlation with their applications, and are often grouped under the name of applied mathematics. Other mathematical areas are developed independently from any application and are therefore called pure mathematics, but practical applications are often discovered later [10].

Mathematics has been a human activity from as far back as written records exist. However, the concept of a "proof" and its associated "mathematical rigour" first appeared in Greek mathematics, most notably in Euclid's Elements. Mathematics developed at a relatively slow pace until the Renaissance, when algebra and infinitesimal calculus were added to arithmetic and geometry as main areas of mathematics. Since then the interaction between mathematical innovations and scientific discoveries have led to a rapid increase in the rate of mathematical discoveries. At the end of the 19th century, the foundational crisis of mathematics led to the systematization of the axiomatic method. This, in turn, gave rise to a dramatic increase in the number of mathematics areas and their fields of applications; a witness of this is the Mathematics Subject Classification, which lists more than sixty first-level areas of mathematics.

Of special importance for physics is the notion of *symmetry*. Most living being, included ourselves, are mirror symmetric. Moreover our left hand is chiral since it is not identical with its mirror image, the right hand, or, more precisely, it cannot be mapped to its mirror image by rotations and translations alone. In particle physics a free particle is described by a finite dimensional irreducible unitary representation of its group symmetries (the Poincaré group in the relativistic case, extended to the parity transformation). The invariants of the group are the mass and the spin of the particle. From this point of view a special role is played in mathematics and abstract algebra by *group theory* which studies the algebraic structures known as groups. The concept of a group is central to abstract algebra: other well-known algebraic structures, such as rings, fields, and vector spaces, can all be seen as groups endowed with additional operations and axioms. Groups recur throughout mathematics, and the methods of group theory have influenced many parts of algebra. Linear algebraic groups and Lie groups are two branches of group theory that have experienced advances and have become subject areas in their own right. Various physical systems, such as crystals, molecules and the hydrogen atom, and three of the four known fundamental forces in the universe, may be modeled by symmetry groups. Thus group theory and the closely related representation theory have many important applications in physics, chemistry, and materials science. The early history of group theory dates from the 19th century. One of the most important mathematical achievements of the 20th century was the collaborative effort, taking up more than 10,000 journal pages and mostly published between 1960 and 2004, that culminated in a complete classification of finite simple groups.

Often when we observe in nature a phase transition this is due to a *spontaneous symmetry breaking*. For example in the transition from the liquid to the solid state the translational symmetry is spontaneously broken. Spontaneous symmetry breaking requires the existence of physical laws (e.g. quantum mechanics) which are invariant under a symmetry transformation (such as translation or rotation), so that any pair of outcomes differing only by that transformation have the same probability distribution. For example if measurements of an observable at any two different positions have the same probability distribution, the observable has translational symmetry. Spontaneous symmetry breaking occurs when this relation breaks down, while the underlying physical laws remain symmetrical. Conversely, in explicit symmetry breaking, if two outcomes are considered, the probability distributions of a pair of outcomes can be different. For example in an electric field, the forces on a charged particle are different in different directions, so the rotational symmetry is explicitly broken by the electric field which does not have this symmetry. The world we live in is the way we see it because of the existence of surfaces which separate two different phases of matter!

When studying mechanical systems we often encounter the problem of *deterministic chaos*. The most common example for this is the three-body problem. This is the problem of taking the initial positions and velocities (or momenta) of three point masses and solving for their subsequent motion according to Newton's laws of motion and Newton's law of universal gravitation. The three-body problem is a special case of the n-body problem. Unlike two-body problems, no general closed-form solution exists (meaning there is no general solution that can be expressed in terms of a finite number of standard mathematical operations), as the resulting dynamical system is chaotic for most initial conditions, and numerical methods are generally required. Historically, the first specific three-body problem to receive extended study was the one involving the Moon, Earth, and the Sun. In an extended modern sense, a three-body problem is any problem in classical mechanics or quantum mechanics that models the motion of three particles. This kind of problems can be studied through chaos theory which is an interdisciplinary scientific theory and branch of mathematics focused on underlying patterns and deterministic laws highly sensitive to initial conditions. Chaos theory states that within the apparent randomness of chaotic complex systems, there are underlying patterns, interconnectedness, constant feedback loops, repetition, self-similarity, fractals, and self-organization. The butterfly effect, an underlying principle of chaos, describes how a small change in one state of a deterministic nonlinear system can result in large differences in a later state (meaning that there is sensitive dependence on initial conditions). Small differences in initial conditions, such as those due to errors in measurements or due to rounding errors in numerical computation, can yield widely diverging outcomes for such complex systems, rendering long-term prediction of their behavior impossible in general. This can happen even though these systems are deterministic, meaning that their future behavior follows a unique evolution and is fully determined by their initial conditions, with no random elements involved. In other words, the deterministic nature of these systems does not make them predictable. This behavior is known as deterministic chaos, or simply chaos. This behavior can be studied through the analysis of a chaotic mathematical model, or through analytical techniques such as recurrence plots and Poincaré maps (the simplest and archetypal one being the logistic map). Although most chaotic systems stems from non-linear maps it is also possible linear chaos.

Deterministic chaotic processes are brothers of *stochastic processes* described by a random probability distribution. Although stochasticity and randomness are distinct in that the former refers to a modeling approach and the latter refers to phenomena themselves, these two terms are often used synonymously. Furthermore, in probability theory, the formal concept of a stochastic process is also referred to as a random process. The best-known stochastic process to which stochastic calculus is applied is the Wiener process (named in honor of Norbert Wiener), which is used for modeling Brownian motion as described by Louis Bachelier in 1900 and by Albert Einstein in 1905 and other physical diffusion processes in space of particles subject to random forces. The main flavors of stochastic calculus are the Itô calculus and the related Stratonovich calculus. This field is created and started by the Japanese mathematician Kiyoshi Itô during World War 2. It rewrites differential equations satisfied by a probability distribution in terms of stochastic differential equations satisfied by the underlying stochastic or random variables. A random variable is a mathematical formalization of a quantity or object which depends on random events. Informally, randomness typically represents some fundamental element of chance, such as in the roll of a dice; it may also represent uncertainty, such as measurement error. However, the interpretation of probability is philosophically complicated, and even in specific cases is not always straightforward. The purely mathematical analysis of random variables is independent of such interpretational difficulties, and can be based upon a rigorous axiomatic setup. In the formal mathematical language of measure theory, a random variable is defined as a measurable function from a probability measure space (called the sample space) to a measurable space (a Borel space). This allows consideration of the measure, which is called the distribution of the random variable; the distribution is thus a probability measure on the set of all possible values of the random variable. It is common to consider the special cases of discrete random variables and absolutely continuous random variables, corresponding to whether a random variable is valued in a discrete set (such as a finite set) or in an interval of real numbers. A special role is played by the stochastic variable with a uniform (constant) probability distribution which is commonly called simply random number. This is a mathematical abstraction. Of course it is not possible to have an algorithm which generates ideal random numbers but only pseudo (or deterministic) random numbers. Any sequence of random numbers generated by an algorithm, a finite sequence of well-defined or unambiguous instructions (i.e. an instruction or expression whose definition assign it a unique interpretation or value), is not truly random, because it is completely determined by an initial value and therefore reproducible. These algorithms are central in applications such as computer experiments or simulations (e.g. for the Monte Carlo method [11]). Of course it is also impossible to find in the real world an ideal dice free of imperfections as we will discuss in Section 10.2.

Probably the most important mathematical function for a physicist is the Gaussian:

$$\psi(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right]. \tag{1.2}$$

It is named after the mathematician Carl Friedrich Gauss. It is used in statistics to describe the normal distribution of a normally distributed random variable with expected value μ and variance σ^2. In probability theory, the central limit theorem establishes that, in many situations, when independent random variables are summed up, their properly normalized sum tends toward a normal distribution even if the original variables themselves are not normally distributed. It is the Green's function for the (homogeneous and isotropic) diffusion equation (and to the heat equation, which is the same thing). It is the wave function of the ground state of the quantum harmonic oscillator (the excited states are the derivatives of the Gaussian function which can be represented using Hermite polynomials) and as such represents the fiducial vector for generating the coherent states. Consequently, Gaussian functions are also associated with the vacuum state in quantum field theory. The initial probability distribution and conditional probability distribution of the Wiener stochastic process (a particular Markov process) are Gaussians. In mathematics the Wiener process is used to represent the integral of a white noise Gaussian process. In physics it is used to study Brownian motion, the diffusion of minute particles suspended in fluid, and other types of diffusion via the Fokker-Planck deterministic equation (or the more general Smoluchowski equation) and Langevin stochastic equations. It also forms the basis for the rigorous path integral formulation of quantum mechanics (by the Feynman-Kac formula, a solution to the Schrödinger equation can be represented in terms of the Wiener process). Gaussian beams are used in optical systems, microwave systems and lasers.

Most often a mathematician is not interested in the explicit form of the solutions of a given problem but simply in answering to the existence and uniqueness questions that it poses. The contrary happens for a physicist. Moreover in mathematics one often finds conjectures that are seemingly true statements that has not yet been rigorously proven. Often a physicist is perfectly comfortable with a heuristic (from Ancient Greek $\epsilon\upsilon\rho\acute{\iota}\sigma\kappa\omega$ 'I find, discover') proof using for example trial and error, a rule of thumb or an educated guess. Therefore employing a practical method that is not guaranteed to be optimal, perfect, or rational, but is nevertheless sufficient for reaching an immediate, short-term goal or approximation. In mathematics, some common heuristics involve the use of visual representations, additional assumptions, forward/backward reasoning and simplification. Moreover a physicist, most often, before rigorous scientific evidence, looks for empirical evidence (the term empirical comes from Greek $\epsilon\mu\pi\epsilon\iota\rho\iota\alpha$ 'experience'), that is evidence resulting from sense experience or experimental procedure.

I will conclude this introduction by summarizing the content of the Chapter "The sense of science" of the book "In un volo di storni" [1] of Giorgio Parisi Nobel prize for physics 2021 where he compares the citation of Richard Feynman "Science is like sex, it also has practical consequences, but that's not why we do it" and Dante's imperative "Fatti non foste a viver come bruti ma per seguir virtute e canoscenza" (You were not made to live like brutes but to follow virtue and knowledge). Moreover he says "We can imagine a vivid metaphor of the scientific enterprise. Some sailors land at night on an unknown island and light a fire on the beach; they begin to see what surrounds them. The more wood they put on the fire, the more the visible area extends; but beyond this there is always a mysterious region, which is hardly perceived in the almost complete darkness, broken by the dim light of the distant fire, and which becomes bigger and bigger as the bonfire increases. The more we explore the universe, the more we discover new regions to explore: each discovery allows us to formulate many new questions that previously we were absolutely unable to conceive." and also "Technological development is unthinkable without a parallel advancement of pure science. As it was well highlighted in a 1977 book, "The Bee and the Architect", pure science not only provides applied science with the necessary knowledge to be able to develop (languages, metaphors, conceptual frameworks), but also has a more hidden role and not least. In fact, basic scientific activities function as a gigantic circuit of testing technological products and stimulating the consumption of advanced high-tech goods." and finally "Science must be defended not only for its practical aspects, but also for its cultural value. We should have the courage to take an example from Robert Wilson, who in 1969, faced with an American senator who insistently asked what were the applications of the construction of the accelerator at Fermilab, near Chicago, in particular if it was militarily useful to defend the country , he replied: "Its value lies in the love for culture: it is like painting, sculpture, poetry, like all those activities of which Americans are patriotically proud; it does not serve to defend our country but it makes it worthwhile to defend our country "."

The book is divided in two parts. In the first one it presents some notorious examples of discoveries in physics driven by predictions due to the observation of mathematical constraints. The second part is devoted to non-scientific consequences of the observation of mathematics like in the arts or in philosophy, ontology, and theology.

In the first part we give seven examples: *i.* the antiparticles (see Chapter 2), where Dirac firmly believed in the fact that in the energy-momentum relation of special relativity, $E^2 = |\mathbf{p}|^2 + m^2$, both solutions for the energy, $E = \pm\sqrt{\ldots}$, would have had physical meaning; *ii.* the black hole (see Chapter 3), where Einstein in solving his field equation for a stationary mass m in vacuum, firmly believed that the $1/0$ divergence in the resulting space-time metric at a distance $2m$ from the center of the mass would have had physical meaning; *iii.* the gravitational waves (see Chapter 4), where Einstein noticed that his field equation in the weak gravity limit would reduce to a wave equation for the space-time metric tensor, thereby predicting by one century the existence of these space-time deformation waves; *iv.* the Higgs boson (see Chapter 5), where Higgs firmly believed in the existence of a gauge field, the Higgs particle, that acquires a mass proportional to the vacuum expectation value of a scalar field, subject to a Mexican-hat potential density, due to the spontaneous symmetry breaking of a local gauge symmetry; *v.* the Casimir effect (see Chapter 6), where Casimir predicted the attractive force between two metal plates due to the zero-point energy of the vacuum of the electromagnetic field when quantized; *vi.* the Aharonov-Bohm effect (see Chapter 7), which gave a physical rather than just mathematical significance to the potential, proving that a charged quantum particle traveling on a closed path locally free of, or screened by, any magnetic field would acquire a phase, observable through interference experiments, proportional to the flux of the magnetic field through the surface bounded by the path; *vii.* the quasicrystal (see Chapter 8), where Roger Penrose developed the Penrose tiling, an aperiodic tiling. This led to Dan Shechtman's discovery in the laboratory of ordered but not periodic structures, a quasiperiodic crystal.

In the second part we talk about the two-ways influences between mathematics and the arts, as for example the one had by the golden ratio, and in philosophy or logic (see Chapter 9), as for example Gödel theorem, and in ontology, as for example the anthropic principle (both in mathematics in Section 10.3 and in physics in Section 10.4), and in theology (see Chapter 10): Is the beauty of mathematics a fruit of God or just of the human beings? Is God an external or an internal creator of the Universe?

Chapter 2

The antiparticle

In particle physics, every type of particle is associated with an *antiparticle* with the same mass but with opposite physical charges (such as electric charge). For example, the antiparticle of the electron is the antielectron (which is often referred to as positron). While the electron has a negative electric charge, the positron has a positive electric charge, and is produced naturally in certain types of radioactive decay. The opposite is also true: the antiparticle of the positron is the electron.

2.1 The mathematical observation

In special relativity one associates to a particle a four-momentum $p = (p^0, p^1, p^2, p^3) = (E/c, \mathbf{p})$ where c is the speed of light, $E/c = p^0$ is the particle energy, and $\mathbf{p} = (p_x, p_y, p_z) = \gamma m \mathbf{v}$ is the particle three-momentum, where $\mathbf{v} = (v_x, v_y, v_z)$ is the particle three-velocity, $\gamma = 1/\sqrt{1 - v^2/c^2}$ the Lorentz's factor, and m is the particle rest or proper mass.

The Minkowski norm squared of the four-momentum gives a Lorentz invariant quantity equal (up to factors of the speed of light c) to the square of the particle proper mass

$$p \cdot p = \frac{E^2}{c^2} - |\mathbf{p}|^2 = m^2 c^2 \,, \tag{2.1}$$

this energy-momentum relation was first established by Paul Dirac in 1928. Moreover from the relativistic dynamics of a massive particle follows that

$$\mathbf{p} = E\mathbf{v}/c^2 \,. \tag{2.2}$$

Then Dirac found, solving for the energy,

$$\frac{E}{c} = \pm\sqrt{m^2 c^2 + |\mathbf{p}|^2} \,, \tag{2.3}$$

and in 1930 Paul Dirac predicted that in parallel to a massive particle with positive energy there should exist a corresponding partner particle with negative energy, but same mass and spin, an *antiparticle*. Because charge is conserved, it is not possible to create an antiparticle without either destroying another particle of the same charge (as is for instance the case when antiparticles are produced naturally via beta decay or the collision of cosmic rays with Earth's atmosphere), or by the simultaneous creation of both a particle and its antiparticle, which can occur in particle accelerators such as the Large Hadron Collider at CERN (the European Council

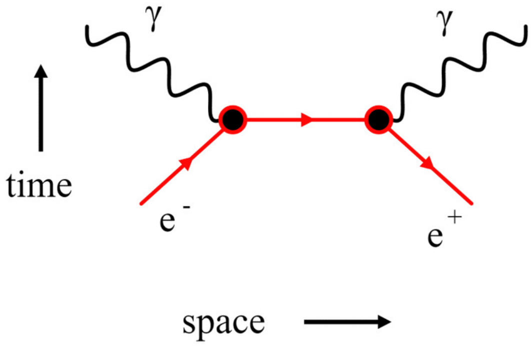


Figure 2.1: Feynman diagram for an electron positron annihilation into two photons, $e^- + e^+ \to \gamma + \gamma$.

for Nuclear Research). Particle-antiparticle pairs can annihilate each other, producing photons (see Fig. 2.1). Since the charges of the particle and antiparticle are opposite, total charge is conserved. A truly neutral particle is its own antiparticle. Electrically neutral particles need not be identical to their antiparticles. The neutron, for example, is made out of quarks, the antineutron from antiquarks, and they are distinguishable from one another because neutrons and antineutrons annihilate each other upon contact. The laws of nature are very nearly symmetrical with respect to particles and antiparticles. For example, an antiproton and a positron can form an antihydrogen atom, which is believed to have the same properties as a hydrogen atom. This leads to the question of why the formation of matter after the Big Bang resulted in a universe consisting almost entirely of matter, rather than being a half-and-half mixture of matter and antimatter. The discovery of charge parity violation helped to shed light on this problem by showing that this symmetry, originally thought to be perfect, was only approximate.

In relativistic quantum theory the four-momentum $p_\mu \to i\hbar\partial/\partial x^\mu$, $\mu = 0, 1, 2, 3$, acts as a differentiation Hermitian operator with respect to the four-position $x = (x^0, x^1, x^2, x^3) = (ct, \mathbf{x})$ which labels a space ($\mathbf{x}$)-time (t) point event. From the point of view of relativistic quantum theory the necessity to consider also the negative energy particles is due to the fact that in order for the description to be effectively linked to the point event it is necessary that the wave function $\psi(x)$ describing the particle *transforms locally* under its symmetries. A pointwise, structureless, elementary, free particle is described by a finite dimensional irreducible unitary representation of its group symmetries (the Poincaré group in the relativistic case, extended to the parity transformation). The invariants of the group are the mass and the spin. For example in the zero spin case one finds that the wave equation must satisfy the Klein-Gordon equation, which in natural units where $\hbar = c = 1$ reads,

$$(\Box + m^2)\psi(x) = 0\,, \tag{2.4}$$

where

$$\Box = \frac{\partial^2}{\partial {x^0}^2} - \sum_{i=1}^{3} \frac{\partial^2}{\partial {x^i}^2}, \tag{2.5}$$

is the d'Alambert operator. Eq. (2.4) is invariant under transformations of the Poincaré group. As a matter of fact the wave function for the positive energy particles obeys to the following equation

$$\left(i\frac{\partial}{\partial x^0} - \sqrt{m^2 - \nabla^2}\right)\psi_+(x) = 0\,, \tag{2.6}$$

which is non-local. If the time evolution has to be local we will need that $\psi(x)$ obeys to a partial differential equation with derivatives of finite order. The requirement for a local equation imposes to have negative energy solutions as well.

2.1.1 The Dirac hole theory

In relativistic quantum theory, particles of spin 1/2 obey to the Dirac equation. For the same reason described above also solutions of the Dirac equation contain negative energy quantum states. As a result, an electron could always radiate energy and fall into a negative energy state. Even worse, it could keep radiating infinite amounts of energy because there were infinitely many negative energy states available. To prevent this unphysical situation from happening, Dirac proposed that a "sea" of negative-energy electrons fills the universe, already occupying all of the lower-energy states so that, due to the Pauli exclusion principle, no other electron could fall into them. Sometimes, however, one of these negative-energy particles could be lifted out of this Dirac sea to become a positive-energy particle. But, when lifted out, it would leave behind a hole in the sea that would act exactly like a positive-energy electron with a reversed charge. These holes were interpreted as "negative-energy electrons" by Paul Dirac and mistakenly identified with protons in his 1930 paper [12]. However, these "negative-energy electrons" turned out to be *positrons*, and not protons.

This picture implied an infinite negative charge for the universe – a problem of which Dirac was aware. Dirac tried to argue that we would perceive this as the normal state of zero charge. Another difficulty was the difference in masses of the electron and the proton. Dirac tried to argue that this was due to the electromagnetic interactions with the sea, until Hermann Weyl proved that hole theory was completely symmetric between negative and positive charges. Dirac also predicted a reaction $\mathrm{e}^- + \mathrm{p}^+ \to \gamma + \gamma$, where an electron and a proton annihilate to give two photons. Robert Oppenheimer and Igor Tamm, however, proved that this would cause ordinary matter to disappear too fast. A year later, in 1931, Dirac modified his theory and postulated the positron, a new particle of the same mass as the electron. The discovery of this particle the next year removed the last two objections to his theory.

2.2 The experiment

In 1932, soon after the prediction of positrons by Paul Dirac, Carl D. Anderson found that cosmic-ray collisions produced these particles in a cloud chamber – a particle detector in which moving electrons (or positrons) leave behind trails as they move through the gas (see Fig. 2.2). The electric charge-to-mass ratio of a particle can be measured by observing the radius of curling of its cloud chamber track in a magnetic field. Positrons, because of the direction that their paths curled, were at first mistaken for electrons traveling in the opposite direction. Positron paths in a cloud chamber trace the same helical path as an electron but rotate in the opposite direction with respect to the magnetic field direction due to their having the same magnitude of charge-to-mass ratio but with opposite charge and, therefore, opposite signed charge-to-mass ratios.

The antiproton and antineutron were found by Emilio Segrè and Owen Chamberlain in 1955 at the University of California, Berkeley [13]. Since then, the antiparticles of many other subatomic particles have been created in particle accelerator experiments. In recent years, complete atoms

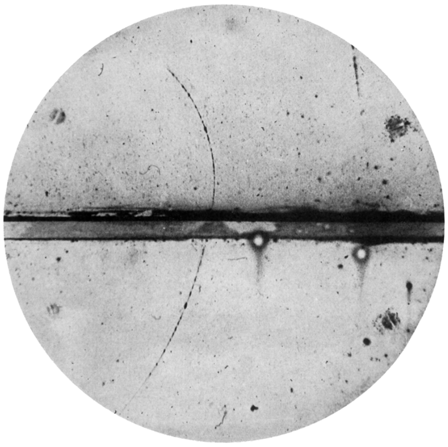


Figure 2.2: Cloud chamber photograph by C. D. Anderson of the first positron ever identified. A 6 mm lead plate separates the chamber. The deflection and direction of the particle's ion trail indicate that the particle is a positron.

of antimatter have been assembled out of antiprotons and positrons, collected in electromagnetic traps.

2.3 The tachyon

But mathematics alone often due to its own abstract nature allows for predictions that violate the physical common sense and therefore purely speculative and with no scientific foundation, that is without any experimental evidence. An example is the *tachyon* or a particle of imaginary mass that travels at speeds higher than the speed of light (superluminal) [1]. This particle was first coined by Gerald Feinberg in 1967 [14]. Of course such particle if existed would violate the causality principle of special relativity [15]. Their four-momentum being space-like they could not slow down to subluminal (meaning slower-than-light) speeds.

From Eqs. (2.1) and (2.2) follows that for a particle of rest mass m and speed v

$$E = \frac{mc^2}{\sqrt{1 - \frac{v^2}{c^2}}}\,, \tag{2.7}$$

this equation shows that when v (the particle velocity) is larger than c (the speed of light), the denominator the energy is imaginary, as the value under the square root is negative. Because the total energy of the particle must be real (and not a complex or imaginary number) in order to have any practical meaning as a measurement, the numerator must also be imaginary: i.e. the rest mass m must be imaginary, as a pure imaginary number divided by another pure imaginary number is a real number.

One effect that violates common sense is that, unlike ordinary particles, the speed of a tachyon increases as its energy decreases. In particular, E approaches zero when v approaches infinity. For ordinary bradyonic matter, E increases with increasing speed, becoming arbitrarily large

[1] The complementary particle types are called luxons (which always move at the speed of light) and bradyons (which always move slower than light); both of these particle types are known to exist.

as v approaches c, the speed of light. Therefore, just as bradyons are forbidden to break the light-speed barrier, so too are tachyons forbidden from slowing down to below c, because infinite energy is required to reach the barrier from either above or below.

Chapter 3

The black hole

A *black hole* is a region of spacetime where gravity is so strong that nothing – no particles or even electromagnetic radiation such as light – can escape from it. The theory of general relativity predicts that a sufficiently compact mass can deform spacetime to form a black hole. The boundary of no escape is called the event horizon. Although it has an enormous effect on the fate and circumstances of an object crossing it, it has no locally detectable features according to general relativity. In many ways, a black hole acts like an ideal black body, as it reflects no light [16]. Moreover, quantum field theory in curved spacetime predicts that event horizons emit Hawking radiation, with the same spectrum as a black body of a temperature inversely proportional to its mass. This temperature is of the order of billionths of a kelvin for stellar black holes, making it essentially impossible to observe directly.

Objects whose gravitational fields are too strong for light to escape were first considered in the 18th century by John Michell and Pierre-Simon Laplace [17]. In 1916, Karl Schwarzschild found the first modern solution of general relativity that would characterize a black hole. David Finkelstein, in 1958, first published the interpretation of "black hole" as a region of space from which nothing can escape. Black holes were long considered a mathematical curiosity; it was not until the 1960s that theoretical work showed they were a generic prediction of general relativity. The discovery of neutron stars by Jocelyn Bell Burnell in 1967 sparked interest in gravitationally collapsed compact objects as a possible astrophysical reality. The first black hole known was Cygnus X-1, identified by several researchers independently in 1971 [18, 19].

Black holes of stellar mass form when massive stars collapse at the end of their life cycle. After a black hole has formed, it can grow by absorbing mass from its surroundings. Supermassive black holes of millions of solar masses ($M_\odot$) may form by absorbing other stars and merging with other black holes. There is consensus that supermassive black holes exist in the centres of most galaxies.

The presence of a black hole can be inferred through its interaction with other matter and with electromagnetic radiation such as visible light. Any matter that falls onto a black hole can form an external accretion disk heated by friction, forming quasars, some of the brightest objects in the universe. Stars passing too close to a supermassive black hole can be shredded into streamers that shine very brightly before being "swallowed" [20]. If other stars are orbiting a black hole, their orbits can determine the black hole's mass and location. Such observations can be used to exclude possible alternatives such as neutron stars. In this way, astronomers have identified numerous stellar black hole candidates in binary systems and established that the radio source known as Sagittarius A*, at the core of the Milky Way galaxy, contains a supermassive black hole of about 4.3 million solar masses.

On 11 February 2016, the LIGO Scientific Collaboration and the Virgo collaboration announced the first direct detection of gravitational waves, representing the first observation of a black hole merger [21]. On 10 April 2019, the first direct image of a black hole and its vicinity was published, following observations made by the Event Horizon Telescope (EHT) in 2017 of the supermassive black hole in Messier 87's galactic centre [22, 23] (see Fig. 3.1). As of 2021, the nearest known body thought to be a black hole is around 1,500 light-years (460 parsecs) away. Though only a couple dozen black holes have been found so far in the Milky Way, there are thought to be hundreds of millions, most of which are solitary and do not cause emission of radiation. Therefore, they would only be detectable by gravitational lensing.

Figure 3.1: This is the first direct image of a supermassive black hole, located at the galactic core of Messier 87 (M87*) [24, 22]. It shows 1.3 mm radio-wave emission from a heated accretion ring orbiting the object at a mean separation of 350 AU, or ten times larger than the orbit of Neptune around the Sun. The dark center is the event horizon and its shadow. The image was released in 2019 by the Event Horizon Telescope Collaboration. On the right panel we show a view of the M87* supermassive black hole in polarised light, taken by Event Horizon Telescope and revealed on 24 March 2021. The direction of lines atop the total intensity mark the direction of the electromagnetic wave electric vector oscillations.

3.1 The mathematical observation

The Schwarzschild solution to the Einstein field equations

$$G_{\mu\nu} = \frac{8\pi G}{c^4} T_{\mu\nu} \,, \tag{3.1}$$

describes spacetime under the influence of a massive, non-rotating, spherically symmetric object. It is considered by some to be one of the simplest and most useful solutions to the Einstein field equations.

Working in a coordinate chart with coordinates (r, θ, ϕ, t) labeled 1 to 4 respectively, we begin with the metric in its most general form (10 independent components, each of which is a smooth function of 4 variables). The solution is assumed to be spherically symmetric, static and vacuum. That is:

- A spherically symmetric spacetime is one that is invariant under rotations and taking the mirror image.

- A static spacetime is one in which all metric components are independent of the time coordinate t (so that $\partial g_{\mu\nu}/\partial t = 0$) and the geometry of the spacetime is unchanged under a time-reversal $t \to -t$.

- A vacuum solution is one that satisfies the equation $T_{\mu\nu} = 0$ for the stress-energy tensor. From the Einstein field equations with zero cosmological constant (3.1), this implies that the Ricci curvature tensor $R_{\mu\nu} = 0$ since contracting the Einstein tensor $G_{\mu\nu} = R_{\mu\nu} - \frac{R}{2} g_{\mu\nu} = 0$ yields a scalar curvature $R = 0$.

- Metric signature used here is $(+,+,+,-)$.

The first simplification to be made is to diagonalise the metric. Under the coordinate transformation, $(r,\theta,\phi,t) \to (r,\theta,\phi,-t)$, all metric components should remain the same. The metric components $g_{\mu 4}$ $(\mu \neq 4)$ change under this transformation as:

$$g'_{\mu 4} = \frac{\partial x^{\alpha}}{\partial x'^{\mu}} \frac{\partial x^{\beta}}{\partial x'^{4}} g_{\alpha\beta} = -g_{\mu 4} \quad (\mu \neq 4)\,. \tag{3.2}$$

But, as we expect $g'_{\mu 4} = g_{\mu 4}$ (metric components remain the same), this means that $g_{\mu 4} = 0$ $(\mu \neq 4)$. Similarly, the coordinate transformations $(r,\theta,\phi,t) \to (r,\theta,-\phi,t)$ and $(r,\theta,\phi,t) \to (r,-\theta,\phi,t)$ respectively give $g_{\mu 3} = 0 \quad (\mu \neq 3)$ and $g_{\mu 2} = 0 \quad (\mu \neq 2)$ Putting all these together gives $g_{\mu\nu} = 0$ $(\mu \neq \nu)$ and hence the metric must be of the form:

$$ds^2 = g_{11}\, dr^2 + g_{22}\, d\theta^2 + g_{33}\, d\phi^2 + g_{44}\, dt^2\,, \tag{3.3}$$

where the four metric components are independent of the time coordinate t (by the static assumption).

On each hypersurface of constant t, constant θ and constant ϕ (i.e., on each radial line), g_{11} should only depend on r (by spherical symmetry). Hence $g_{11} = A(r)$. A similar argument applied to g_{44} shows that $g_{44} = B(r)$. g_{22} and g_{33} must be the same as for a flat spacetime since stretching or compressing an elastic material in a spherically symmetric manner (radially) will not change the angular distance between two points. Thus, the metric can be put in the form

$$ds^2 = A\,(r)\, dr^2 + r^2\, d\theta^2 + r^2 \sin^2\theta\, d\phi^2 + B\,(r)\, dt^2\,, \tag{3.4}$$

with A and B as yet undetermined functions of r. Note that if A or B is equal to zero at some point, the metric would be singular at that point.

Using the metric above, we find the Christoffel symbols, $\Gamma^{i}_{kl} = \frac{1}{2} g^{im}(g_{mk,l} + g_{ml,k} - g_{kl,m})$, where the indices are $(1,2,3,4) = (r,\theta,\phi,t)$ and a comma stands for a partial derivative with respect to the given space-time component. The sign $'$ denotes a total derivative of a function.

$$\Gamma^1_{ik} = \begin{bmatrix} A'/(2A) & 0 & 0 & 0 \\ 0 & -r/A & 0 & 0 \\ 0 & 0 & -r\sin^2\theta/A & 0 \\ 0 & 0 & 0 & -B'/(2A) \end{bmatrix}, \tag{3.5}$$

$$\Gamma^2_{ik} = \begin{bmatrix} 0 & 1/r & 0 & 0 \\ 1/r & 0 & 0 & 0 \\ 0 & 0 & -\sin\theta\cos\theta & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \tag{3.6}$$

$$\Gamma^3_{ik} = \begin{bmatrix} 0 & 0 & 1/r & 0 \\ 0 & 0 & \cot\theta & 0 \\ 1/r & \cot\theta & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \tag{3.7}$$

$$\Gamma^4_{ik} = \begin{bmatrix} 0 & 0 & 0 & B'/(2B) \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ B'/(2B) & 0 & 0 & 0 \end{bmatrix}. \tag{3.8}$$

To determine A and B, the vacuum field equations $R_{\alpha\beta} = 0$ are employed. Hence:

$$R_{\alpha\beta} = \Gamma^\rho_{\beta\alpha,\rho} - \Gamma^\rho_{\rho\alpha,\beta} + \Gamma^\rho_{\rho\lambda}\Gamma^\lambda_{\beta\alpha} - \Gamma^\rho_{\beta\lambda}\Gamma^\lambda_{\rho\alpha} = 0\,, \tag{3.9}$$

where a comma is used to set off the index that is being used for the derivative. The Ricci curvature is diagonal in the given coordinates:

$$R_{tt} = -\frac{1}{4}\frac{B'}{A}\left(\frac{A'}{A} - \frac{B'}{B} + \frac{4}{r}\right) - \frac{1}{2}\left(\frac{B'}{A}\right)'\,, \tag{3.10}$$

$$R_{rr} = -\frac{1}{2}\left(\frac{B'}{B}\right)' - \frac{1}{4}\left(\frac{B'}{B}\right)^2 + \frac{1}{4}\frac{A'}{A}\left(\frac{B'}{B} + \frac{4}{r}\right)\,, \tag{3.11}$$

$$R_{\theta\theta} = 1 - \left(\frac{r}{A}\right)' - \frac{r}{2A}\left(\frac{A'}{A} + \frac{B'}{B}\right)\,, \tag{3.12}$$

$$R_{\phi\phi} = \sin^2(\theta) R_{\theta\theta}\,, \tag{3.13}$$

where the prime means the r derivative of the functions. Only three of the field equations are nontrivial and upon simplification become:

$$4A'B^2 - 2rB''AB + rA'B'B + rB'^2A = 0\,, \tag{3.14}$$

$$rA'B + 2A^2B - 2AB - rB'A = 0\,, \tag{3.15}$$

$$-2rB''AB + rA'B'B + rB'^2A - 4B'AB = 0\,, \tag{3.16}$$

(the fourth equation is just $\sin^2\theta$ times the second equation). Subtracting the first and third equations produces $A'B + AB' = 0 \Rightarrow A(r)B(r) = K$ where K is a non-zero real constant. Substituting $A(r)B(r) = K$ into the second equation and tidying up gives $rA' = A(1 - A)$ which has general solution:

$$A(r) = \left(1 + \frac{1}{Sr}\right)^{-1}\,, \tag{3.17}$$

for some non-zero real constant S. Hence, the metric for a static, spherically symmetric vacuum solution is now of the form:

$$ds^2 = \left(1 + \frac{1}{Sr}\right)^{-1} dr^2 + r^2(d\theta^2 + \sin^2\theta d\phi^2) + K\left(1 + \frac{1}{Sr}\right) dt^2 \,. \tag{3.18}$$

Note that the spacetime represented by the above metric is asymptotically flat, i.e. as $r \to \infty$, the metric approaches that of the Minkowski metric and the spacetime manifold resembles that of Minkowski space.

When the mass m of the object is small we can use the weak-field approximation in which the geodesic equation $d^2x^\mu/dt^2 = \Gamma^\mu_{44} = \frac{1}{2}dg_{44}/dr$ must reduce to the Newton equation $d^2x^\mu/dt^2 = -Gm/r^2$. Therefore

$$g_{44} = K\left(1 + \frac{1}{Sr}\right) \approx -c^2 + \frac{2Gm}{r} \,, \tag{3.19}$$

where we used the fact that the desired solution degenerates to Minkowski metric when the motion happens far away from the object (r approaches to positive infinity). We then find $K = -c^2$ and $1/S = -2Gm/c^2$. Introducing the Schwarzschild radius $r_s = 2Gm/c^2$ the Schwarzschild metric may be rewritten in the alternative form

$$ds^2 = \left(1 - \frac{r_s}{r}\right)^{-1} dr^2 + r^2(d\theta^2 + \sin^2\theta d\phi^2) - c^2\left(1 - \frac{r_s}{r}\right) dt^2 \,, \tag{3.20}$$

which shows that the metric becomes singular approaching the event horizon (that is, $r \to r_s$. The metric singularity is not a physical one (although there is a real physical singularity at $r = 0$), as can be shown by using a suitable coordinate transformation (e.g. the Kruskal-Szekeres coordinate system) [25, 26, 27]. In Fig. 3.2 we show a Flamm's paraboloid whose local geometry is fixed by the metric

$$ds^2 = \left(1 - \frac{r_s}{r}\right)^{-1} dr^2 + r^2 d\theta^2 \,. \tag{3.21}$$

It is composed by two identical "universes": The one at $z > 0$ and the one at $z < 0$. These are both multiply connected surfaces with the "Schwarzschild wormhole" providing the path from one to the other.

Conventional black holes are formed by gravitational collapse of heavy objects such as stars, but they can also in theory be formed by other processes. Whenever a star shrinks inside its Schwarzschild radius the black hole forms and the star become causally hidden behind it.

Gravitational collapse occurs when an object's internal pressure is insufficient to resist the object's own gravity. For stars this usually occurs either because a star has too little "fuel" left to maintain its temperature through stellar nucleosynthesis, or because a star that would have been stable receives extra matter in a way that does not raise its core temperature. In either case the star's temperature is no longer high enough to prevent it from collapsing under its own weight. The collapse may be stopped by the degeneracy pressure of the star's constituents, allowing the condensation of matter into an exotic denser state. The result is one of the various types of compact star. Which type forms depends on the mass of the remnant of the original star left if the outer layers have been blown away (for example, in a Type II supernova). The mass of the remnant, the collapsed object that survives the explosion, can be substantially less than that of the original star. Remnants exceeding 5 $M_\odot$ are produced by stars that were over 20 $M_\odot$ before the collapse.

If the mass of the remnant exceeds about $3 - 4\ M_\odot$ (the Tolman–Oppenheimer–Volkoff limit [28]), either because the original star was very heavy or because the remnant collected additional

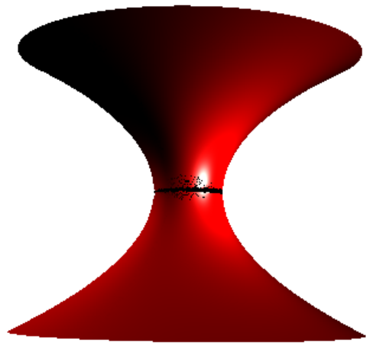


Figure 3.2: Flamm's paraboloid: The Riemannian surface with the metric of Eq. (3.21) which is embeddable in the three dimensional euclidean space with cylindrical coordinates (r, θ, z) with $ds^2 = dz^2 + dr^2 + r^2 d\theta^2$ and $z(r) = \pm 2\sqrt{r_s(r - r_s)}$. It has a hole of radius r_s. The $r = r_s$ region of the surface its usually called its "horizon".

mass through accretion of matter, even the degeneracy pressure of neutrons is insufficient to stop the collapse. No known mechanism (except possibly quark degeneracy pressure, see quark star) is powerful enough to stop the implosion and the object will inevitably collapse to form a black hole.

The gravitational collapse of heavy stars is assumed to be responsible for the formation of stellar mass black holes. Star formation in the early universe may have resulted in very massive stars, which upon their collapse would have produced black holes of up to $10^3\ M_\odot$. These black holes could be the seeds of the supermassive black holes found in the centres of most galaxies. It has further been suggested that massive black holes with typical masses of $\sim 10^5\ M_\odot$ could have formed from the direct collapse of gas clouds in the young universe. These massive objects have been proposed as the seeds that eventually formed the earliest quasars observed already at redshift. Some candidates for such objects have been found in observations of the young universe.

While most of the energy released during gravitational collapse is emitted very quickly, an outside observer does not actually see the end of this process. Even though the collapse takes a finite amount of time from the reference frame of infalling matter, a distant observer would see the infalling material slow and halt just above the event horizon, due to gravitational time dilation. Light from the collapsing material takes longer and longer to reach the observer, with the light emitted just before the event horizon forms delayed an infinite amount of time. Thus the external observer never sees the formation of the event horizon; instead, the collapsing material seems to become dimmer and increasingly red-shifted, eventually fading away [29].

Gravitational collapse requires great density. In the current epoch of the universe these high densities are found only in stars, but in the early universe shortly after the Big Bang densities were much greater, possibly allowing for the creation of black holes. High density alone is not enough to allow black hole formation since a uniform mass distribution will not allow the mass to bunch up. In order for primordial black holes to have formed in such a dense medium, there must have been initial density perturbations that could then grow under their own gravity. Different models for the early universe vary widely in their predictions of the scale of these fluctuations. Various models predict the creation of primordial black holes ranging in size from a Planck mass

($m_P = \sqrt{\hbar c/G} \approx 1.2 \times 10^{19}\mathrm{GeV}/c^2 \approx 2.2 \times 10^{-8}\mathrm{kg}$) to hundreds of thousands of solar masses [30].

Despite the early universe being extremely dense – far denser than is usually required to form a black hole – it did not re-collapse into a black hole during the Big Bang. Models for the gravitational collapse of objects of relatively constant size, such as stars, do not necessarily apply in the same way to rapidly expanding space such as the Big Bang.

Gravitational collapse is not the only process that could create black holes. In principle, black holes could be formed in high-energy collisions that achieve sufficient density. As of 2002, no such events have been detected, either directly or indirectly as a deficiency of the mass balance in particle accelerator experiments [31]. This suggests that there must be a lower limit for the mass of black holes. Theoretically, this boundary is expected to lie around the Planck mass, where quantum effects are expected to invalidate the predictions of general relativity [32]. This would put the creation of black holes firmly out of reach of any high-energy process occurring on or near the Earth. However, certain developments in quantum gravity suggest that the minimum black hole mass could be much lower: some braneworld scenarios for example put the boundary as low as 1 TeV/c^2 [33]. This would make it conceivable for micro black holes to be created in the high-energy collisions that occur when cosmic rays hit the Earth's atmosphere, or possibly in the Large Hadron Collider at CERN. These theories are very speculative, and the creation of black holes in these processes is deemed unlikely by many specialists. Even if micro black holes could be formed, it is expected that they would evaporate in about $10 - 25$ seconds, posing no threat to the Earth.

3.2 The experiment

By nature, black holes do not themselves emit any electromagnetic radiation other than the hypothetical Hawking radiation, so astrophysicists searching for black holes must generally rely on indirect observations. For example, a black hole's existence can sometimes be inferred by observing its gravitational influence on its surroundings.

On 10 April 2019, an image was released of a black hole (see Fig. 3.1), which is seen magnified because the light paths near the event horizon are highly bent. The dark shadow in the middle results from light paths absorbed by the black hole [34]. The image is in false color, as the detected light halo in this image is not in the visible spectrum, but radio waves.

The Event Horizon Telescope (EHT) is an active program that directly observes the immediate environment of black holes' event horizons, such as the black hole at the centre of the Milky Way. In April 2017, EHT began observing the black hole at the centre of Messier 87. In all, eight radio observatories on six mountains and four continents observed the galaxy in Virgo on and off for 10 days in April 2017 to provide the data yielding the image in April 2019. After two years of data processing, EHT released the first direct image of a black hole; specifically, the supermassive black hole that lies in the centre of the aforementioned galaxy. What is visible is not the black hole – which shows as black because of the loss of all light within this dark region. Instead, it is the gases at the edge of the event horizon (displayed as orange or red) that define the black hole.

The brightening of this material in the 'bottom' half of the processed EHT image is thought to be caused by Doppler beaming, whereby material approaching the viewer at relativistic speeds is perceived as brighter than material moving away. In the case of a black hole, this phenomenon implies that the visible material is rotating at relativistic speeds ($> 1,000 km/s$), the only speeds at which it is possible to centrifugally balance the immense gravitational attraction of the singularity, and thereby remain in orbit above the event horizon. This configuration of bright material implies that the EHT observed M87* from a perspective catching the black hole's accretion disc nearly edge-on, as the whole system rotated clockwise [35]. However, the extreme gravitational lensing associated with black holes produces the illusion of a perspective that sees the accretion disc from above. In reality, most of the ring in the EHT image was created when the light emitted by the far side of the accretion disc bent around the black hole's gravity well and escaped, meaning that most of the possible perspectives on M87* can see the entire disc, even that directly behind the "shadow".

In 2015, the EHT detected magnetic fields just outside the event horizon of Sagittarius A* and even discerned some of their properties. The field lines that pass through the accretion disc were a complex mixture of ordered and tangled. Theoretical studies of black holes had predicted the existence of magnetic fields [36].

On 14 September 2015, the LIGO gravitational wave observatory made the first-ever successful direct observation of gravitational waves [21]. The signal was consistent with theoretical predictions for the gravitational waves produced by the merger of two black holes: one with about 36 solar masses, and the other around 29 solar masses [21, 37]. This observation provides the most concrete evidence for the existence of black holes to date. For instance, the gravitational wave signal suggests that the separation of the two objects before the merger was just $350km$ (or roughly four times the Schwarzschild radius corresponding to the inferred masses). The objects must therefore have been extremely compact, leaving black holes as the most plausible interpretation [21].

More importantly, the signal observed by LIGO also included the start of the post-merger ringdown, the signal produced as the newly formed compact object settles down to a stationary state. Arguably, the ringdown is the most direct way of observing a black hole [38]. From the LIGO signal, it is possible to extract the frequency and damping time of the dominant mode of the ringdown. From these, it is possible to infer the mass and angular momentum of the final object, which match independent predictions from numerical simulations of the merger [39]. The frequency and decay time of the dominant mode are determined by the geometry of the photon sphere. Hence, observation of this mode confirms the presence of a photon sphere; however, it cannot exclude possible exotic alternatives to black holes that are compact enough to have a photon sphere [38].

The observation also provides the first observational evidence for the existence of stellar-mass black hole binaries. Furthermore, it is the first observational evidence of stellar-mass black holes weighing 25 solar masses or more [40].

Since then, many more gravitational wave events have been observed.

Chapter 4

The gravitational waves

Gravitational waves are disturbances or ripples in the curvature of spacetime, generated by accelerated masses, that propagate as waves outward from their source at the speed of light (see Fig. 4.1). They were proposed by Henri Poincaré in 1905 and subsequently predicted in 1916 [41, 42] by Albert Einstein on the basis of his general theory of relativity. Later he refused to accept gravitational waves. Gravitational waves transport energy as gravitational radiation, a form of radiant energy similar to electromagnetic radiation [43]. Newton's law of universal gravitation, part of classical mechanics, does not provide for their existence, since that law is predicated on the assumption that physical interactions propagate instantaneously (at infinite speed) – showing one of the ways the methods of classical physics are unable to explain phenomena associated with relativity.

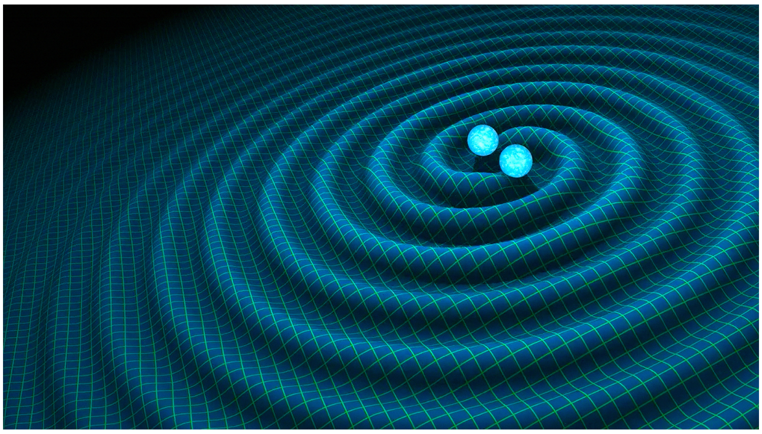


Figure 4.1: Pictorial representation of a gravitational wave.

These are one of several analogies between weak-field gravity and electromagnetism in that, they are analogous to electromagnetic waves. Since Einstein's equations are non-linear, arbitrarily strong gravitational waves do not obey linear superposition, making their description difficult. However, linear approximations of gravitational waves are sufficiently accurate to describe the exceedingly weak waves that are expected to arrive here on Earth from far-off cosmic events. For gravitational waves produced in astrophysically relevant situations, such as the merger of two black holes, numerical methods are presently the only way to construct appropriate models [44].

4.1 The mathematical observation

The Einstein field equation (EFE) describing the geometry of spacetime is given as (using natural units)

$$R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu} = 8\pi G T_{\mu\nu} \,, \tag{4.1}$$

where $R_{\mu\nu}$ is the Ricci tensor, R is the Ricci scalar, $T_{\mu\nu}$ is the energy–momentum tensor, and $g_{\mu\nu}$ is the spacetime metric tensor that represent the solutions of the equation.

Although succinct when written out using Einstein notation, hidden within the Ricci tensor and Ricci scalar are exceptionally nonlinear dependencies on the metric which render the prospect of finding exact solutions impractical in most systems. However, when describing particular systems for which the curvature of spacetime is small (meaning that terms in the EFE that are quadratic in $g_{\mu\nu}$ do not significantly contribute to the equations of motion), one can model the solution of the field equations as being the Minkowski metric $\eta_{\mu\nu}$ plus a small perturbation term $h_{\mu\nu}$. In other words:

$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}, \qquad |h_{\mu\nu}| \ll 1 \,. \tag{4.2}$$

In this regime, substituting the general metric $g_{\mu\nu}$ for this perturbative approximation results in a simplified expression for the Ricci tensor:

$$R_{\mu\nu} = \frac{1}{2}(\partial_\sigma \partial_\mu h^\sigma_\nu + \partial_\sigma \partial_\nu h^\sigma_\mu - \partial_\mu \partial_\nu h - \Box h_{\mu\nu}) \,, \tag{4.3}$$

where $h = \eta^{\mu\nu} h_{\mu\nu}$ is the trace of the perturbation, ∂_μ denotes the partial derivative with respect to the x^μ coordinate of spacetime, and $\Box = \eta^{\mu\nu}\partial_\mu\partial_\nu$ is the d'Alembert operator.

Together with the Ricci scalar,

$$R = \eta_{\mu\nu} R^{\mu\nu} = \partial_\mu \partial_\nu h^{\mu\nu} - \Box h \,, \tag{4.4}$$

the left side of the field equation reduces to

$$R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu} = \frac{1}{2}(\partial_\sigma \partial_\mu h^\sigma_\nu + \partial_\sigma \partial_\nu h^\sigma_\mu - \partial_\mu \partial_\nu h - \Box h_{\mu\nu} - \eta_{\mu\nu} \partial_\rho \partial_\lambda h^{\rho\lambda} + \eta_{\mu\nu} \Box h) \,, \tag{4.5}$$

and thus the EFE is reduced to a linear, second order partial differential equation in terms of $h_{\mu\nu}$.

The process of decomposing the general spacetime $g_{\mu\nu}$ into the Minkowski metric plus a perturbation term is not unique. This is due to the fact that different choices for coordinates may give different forms for $h_{\mu\nu}$. In order to capture this phenomenon, the application of gauge symmetry is introduced.

Gauge symmetries are a mathematical device for describing a system that does not change when the underlying coordinate system is "shifted" by an infinitesimal amount. So although the perturbation metric $h_{\mu\nu}$ is not consistently defined between different coordinate systems, the overall system which it describes is.

To capture this formally, the non-uniqueness of the perturbation $h_{\mu\nu}$ is represented as being a consequence of the diverse collection of diffeomorphisms on spacetime that leave $h_{\mu\nu}$ sufficiently small. Therefore to continue, it is required that $h_{\mu\nu}$ be defined in terms of a general set of diffeomorphisms then select the subset of these that preserve the small scale that is required by the weak-field approximation. One may thus define ϕ to denote an arbitrary diffeomorphism that maps the flat Minkowski spacetime to the more general spacetime represented by the metric $g_{\mu\nu}$. With this, the perturbation metric may be defined as the difference between the pullback of $g_{\mu\nu}$ and the Minkowski metric:

$$h_{\mu\nu} = (\phi^* g)_{\mu\nu} - \eta_{\mu\nu} \,. \tag{4.6}$$

The diffeomorphisms ϕ may thus be chosen such that $|h_{\mu\nu}| \ll 1$.

Given then a vector field ξ^μ defined on the flat, background spacetime, an additional family of diffeomorphisms ψ_ϵ may be defined as those generated by ξ^μ and parameterized by $\epsilon > 0$. These new diffeomorphisms will be used to represent the coordinate transformations for "infinitesimal shifts" as discussed above. Together with ϕ , a family of perturbations is given by

$$\begin{aligned} h^{(\epsilon)}_{\mu\nu} &= [(\phi \circ \psi_\epsilon)^* g]_{\mu\nu} - \eta_{\mu\nu} \\ &= [\psi^*_\epsilon(\phi^* g)]_{\mu\nu} - \eta_{\mu\nu} \\ &= \psi^*_\epsilon(h + \eta)_{\mu\nu} - \eta_{\mu\nu} \\ &= (\psi^*_\epsilon h)_{\mu\nu} + \epsilon \left[\frac{(\psi^*_\epsilon \eta)_{\mu\nu} - \eta_{\mu\nu}}{\epsilon} \right] . \end{aligned} \tag{4.7}$$

Therefore, in the limit $\epsilon \to 0$,

$$h^{(\epsilon)}_{\mu\nu} = h_{\mu\nu} + \epsilon \mathcal{L}_\xi \eta_{\mu\nu} \,, \tag{4.8}$$

where $\mathcal{L}_\xi$ is the Lie derivative along the vector field ξ_μ.

The Lie derivative works out to yield the final gauge transformation of the perturbation metric $h_{\mu\nu}$:

$$h^{(\epsilon)}_{\mu\nu} = h_{\mu\nu} + \epsilon(\partial_\mu \xi_\nu + \partial_\nu \xi_\mu) \,, \tag{4.9}$$

which precisely define the set of perturbation metrics that describe the same physical system. In other words, it characterizes the gauge symmetry of the linearized field equations.

The harmonic gauge (also referred to as the Lorenz gauge) is selected whenever it is necessary to reduce the linearized field equations as much as possible. This can be done if the condition

$$\partial_\mu h^\mu_\nu = \frac{1}{2} \partial_\nu h \,, \tag{4.10}$$

is true. To achieve this, ξ_μ is required to satisfy the relation

$$\Box \xi_\mu = -\partial_\nu h^\nu_\mu + \frac{1}{2} \partial_\mu h \,. \tag{4.11}$$

Consequently, by using the harmonic gauge, the Einstein tensor $G_{\mu\nu} = R_{\mu\nu} - \dfrac{1}{2} R g_{\mu\nu}$ reduces to

$$G_{\mu\nu} = -\frac{1}{2} \Box \left(h^{(\epsilon)}_{\mu\nu} - \frac{1}{2} h^{(\epsilon)} \eta_{\mu\nu} \right) \,. \tag{4.12}$$

Therefore, by writing it in terms of a "trace-reversed" metric, $\bar{h}^{(\epsilon)}_{\mu\nu} = h^{(\epsilon)}_{\mu\nu} - \dfrac{1}{2} h^{(\epsilon)} \eta_{\mu\nu}$, the linearized field equations reduce to

$$\Box \bar{h}^{(\epsilon)}_{\mu\nu} = -16\pi G T_{\mu\nu} \,. \tag{4.13}$$

Which can be solved exactly using the wave solutions that define gravitational radiation.

4.2 The experiment

The first indirect evidence for the existence of gravitational waves came from the observed orbital decay of the Hulse-Taylor binary pulsar, which matched the decay predicted by general relativity as energy is lost to gravitational radiation. In 1993, Russell A. Hulse and Joseph Hooton Taylor Jr. received the Nobel Prize in Physics for this discovery. The first direct observation of gravitational waves was not made until 14 September 2015, when a signal generated by the merger of two black holes was received by the LIGO gravitational wave detectors in Livingston, Louisiana, and in Hanford, Washington (announced by the LIGO and Virgo collaborations on 11 February 2016 [21]). The waveform, detected by both LIGO observatories, matched the predictions of general relativity [45, 46, 47] for a gravitational wave emanating from the inward spiral and merger of a pair of black holes of around 36 and 29 solar masses (merging about 1.3 billion light-years away) and the subsequent "ringdown" of the single resulting black hole (The ringdown phase is the settling down of the merged black hole into a sphere). The signal was named GW150914 (from "Gravitational Wave" and the date of observation 2015-09-14;see Fig. 4.2) [21]. During the final fraction of a second of the merger (the binary system as it spiralled

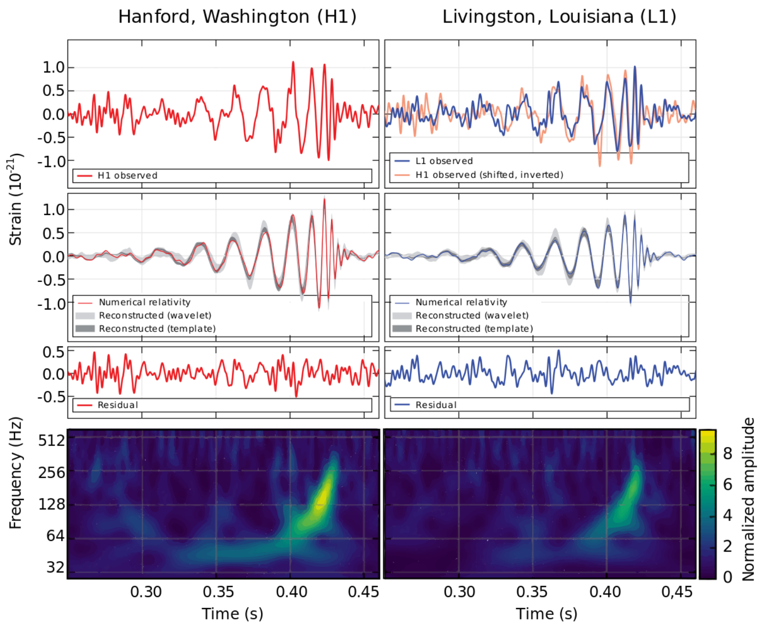


Figure 4.2: Fingerprint of GW150914.

together and merged), it released more than 50 times the power of all the stars in the observable universe combined. The signal increased in frequency from 35 to 250 Hz over 10 cycles (5 orbits) as it rose in strength for a period of 0.2 second [21]. The mass of the new merged black hole was 62 solar masses. Energy equivalent to three solar masses was emitted as gravitational waves. The signal was seen by both LIGO detectors in Livingston and Hanford, with a time difference of 7 milliseconds due to the angle between the two detectors and the source. The signal came from the Southern Celestial Hemisphere, in the rough direction of (but much farther away than) the Magellanic Clouds. The gravitational waves were observed in the region more than 5 sigma (in other words, 99.99997% chances of showing/getting the same result), the probability of finding enough to have been assessed/considered as the evidence/proof in a experiment of statistical physics. It was also the first observation of a binary black hole merger, demonstrating both the existence of binary stellar-mass black hole systems and the fact that such mergers could occur within the current age of the universe.

The waves given off by the cataclysmic merger of GW150914 reached Earth as a ripple in spacetime that changed the length of a 4 km LIGO arm by a thousandth of the width of a proton, proportionally equivalent to changing the distance to the nearest star outside the Solar System by one hair's width.

The 2017 Nobel Prize in Physics was subsequently awarded to Rainer Weiss, Kip Thorne and Barry Barish "for decisive contributions to the LIGO detector and the observation of gravitational waves."

The observation confirms the last remaining directly undetected prediction of general relativity and corroborates its predictions of space-time distortion in the context of large scale cosmic events (known as strong field tests). It was also heralded as inaugurating a new era of gravitational-wave astronomy, which will enable observations of violent astrophysical events that were not previously possible and potentially allow the direct observation of the very earliest history of the universe. In gravitational-wave astronomy, observations of gravitational waves are used to infer data about the sources of gravitational waves. Sources that can be studied this way include binary star systems composed of white dwarfs, neutron stars [48], and black holes; events such as supernovae; and the formation of the early universe shortly after the Big Bang.

Chapter 5

The Higgs boson

The *Higgs boson*, sometimes called the Higgs particle, is an elementary particle in the Standard Model of particle physics produced by the quantum excitation of the Higgs field, one of the fields in particle physics theory. In the Standard Model, the Higgs particle is a massive scalar boson with zero spin, even (positive) parity, no electric charge, and no colour charge, that couples to (interacts with) mass. It is also very unstable, decaying into other particles almost immediately.

The Higgs field is a scalar field, with two neutral and two electrically charged components that form a complex doublet of the weak isospin SU(2) symmetry. Its "Mexican hat-shaped" potential has a nonzero value everywhere (including otherwise empty space), which breaks the weak isospin symmetry of the electroweak interaction, and via the Higgs mechanism gives some particles mass.

Both the field and the boson are named after physicist Peter Higgs, who in 1964 along with five other scientists in three teams, proposed the Higgs mechanism, a way that some particles can acquire mass. (All fundamental particles known at the time should be massless at very high energies, but fully explaining how some particles gain mass at lower energies, had been extremely difficult.) If these ideas were correct, a particle known as a scalar boson should also exist, with certain properties. This particle was called the Higgs boson, and could be used to test whether the Higgs field was the correct explanation.

After a 40 year search, a subatomic particle with the expected properties was discovered in 2012 by the ATLAS and CMS experiments at the Large Hadron Collider (LHC) at CERN near Geneva, Switzerland. The new particle was subsequently confirmed to match the expected properties of a Higgs boson. Physicists from two of the three teams, Peter Higgs and François Englert, were awarded the Nobel Prize in Physics in 2013 for their theoretical predictions. Although Higgs's name has come to be associated with this theory, several researchers between about 1960 and 1972 independently developed different parts of it.

The following has been extracted with minor variations from chapter 11 of the book "A Modern Introduction to Quantum Field Theory" [49] by Michele Maggiore.

5.1 The mathematical observation and the Spontaneous Symmetry Breaking (SSB) mechanism

In this Section we present the phenomenon of spontaneous symmetry breaking (SSB). This is a mechanism of great importance both in particle physics and in condensed matter physics. Its generality and importance stem from the fact that it deals with how a symmetry of the action in Quantum Field Theory (QFT) or of the Hamiltonian in a statistical system is reflected on the ground state of the system.

As we will see in Section 5.1.1, SSB strictly speaking can only take place in a system with an infinite number of degrees of freedom. It is therefore a genuinely field-theoretical phenomenon, which does not appear in Quantum Mechanical (QM) systems with a finite number of variables.

We will examine the effect of SSB on different types of symmetries. In Section 5.1.2 we will discuss the SSB of global symmetries, and the emergence of Goldstone bosons. In Section 5.1.3 we will examine the SSB of local abelian symmetries, and we will see that it is a crucial element in the BCS theory of superconductivity, when the latter is formulated in field theoretical language.

5.1.1 Degenerate vacua in QM and QFT

Spontaneous symmetry breaking is a very general phenomenon characterized by the fact that the action has a symmetry (global or local) but the quantum theory, instead of having a unique vacuum state which respects this symmetry, has a family of degenerate vacua that transform into each other under the action of the symmetry group. A simple example is given by Jellium [50], an electron gas, in any spatial dimension at low temperatures, in the quantum regime. Above a critical electron density Jellium is translationally invariant. Below a critical electron density it becomes thermodynamically favorable to develop the Wigner crystal which is a vaccum where the translational invariance is broken into a body centered cubic lattice in 3D, a triangular lattice in 2D, and an evenly spaced lattice in 1D

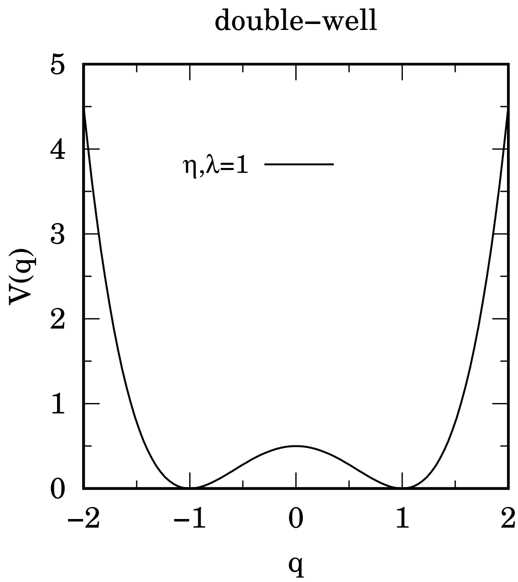


Figure 5.1: A double-well potential of Eq. (5.1).

The original invariance of the Lagrangian is now reflected in the fact that, instead of a single vacuum state, there is a whole family of vacua related to each other by rotations, since the magnetization can in principle develop in any direction. However, the system will choose one of these states as its vacuum state. The symmetry is then said to be *spontaneously broken* by the choice of a vacuum. SSB is a phenomenon that cannot take place in a quantum mechanical system with a finite number of degrees of freedom, since in this case, if we have a family of "vacua", the true vacuum state is a superposition of them which respects the original symmetry. To illustrate this point, we consider for instance the quantum mechanics of a particle, described by a coordinate $q(t)$, in a potential

$$V(q) = \frac{\lambda^2}{2}[q^2(t) - \eta^2]^2 \,, \tag{5.1}$$

with λ, η parameters. This potential is shown in Fig. 5.1, and is called a double-well potential. The Lagrangian is

$$\mathcal{L} = \frac{1}{2}m\dot{q}^2 - V(q) \,, \tag{5.2}$$

and is symmetric under the parity transformation $q(t) \to -q(t)$ (this is also called a Z_2 symmetry, where Z_2 is the finite group formed by 1 and -1 under multiplication). The potential has two minima, at $q = \pm\eta$. We can solve the Schrödinger equation expanding the potential around the minimum at $q = +\eta$, retaining only the quadratic term in the Taylor expansion of the potential around η (so that we have a harmonic oscillator), and treating in perturbation theory all higher powers of the expansion of the potential. We call $|+\rangle$ the ground state obtained in this way; more precisely, this is a perturbative vacuum. We can do the same expanding around $-\eta$, and we call $|-\rangle$ the corresponding perturbative vacuum. However, the true ground state of the theory is neither $|+\rangle$ nor $|-\rangle$. At the non-perturbative level there is a nonvanishing amplitude for the transition between these two states, due to the possibility of tunneling under the barrier which separates the two minima, and which can be computed in a semiclassical WKB approximation. Because of the tunneling process, the Hamiltonian is not diagonal in the $|\pm\rangle$ basis. Rather, we will have $a \equiv \langle +|\mathcal{H}|+\rangle = \langle -|\mathcal{H}|-\rangle$ and $b \equiv \langle -|\mathcal{H}|+\rangle = \langle +|\mathcal{H}|-\rangle$, where $b \ll a$, since the tunneling amplitude is exponentially suppressed. Diagonalizing this Hamiltonian we immediately find that the eigenstates are the symmetric and antisymmetric combinations $|S\rangle = |+\rangle + |-\rangle$ and $|A\rangle = |+\rangle - |-\rangle$, with energies $a \pm b$, respectively. Therefore the degeneracy between these states is lifted by the fact that $b \neq 0$, and the true ground state is the combination with energy $a - |b|$.

Under a parity transformation $q \to -q$, $|S\rangle$ is invariant while $|A\rangle$ picks a minus sign. Recalling that physical states are defined up to an overall phase, we see that the true ground state of the Hamiltonian goes into itself under parity, and there is no SSB of the Z_2 symmetry.

Consider now a real scalar field $\phi(t, \mathbf{x})$. In natural units $\hbar = c = 1$ we denote with $\partial_\mu = \partial/\partial x^\mu$ where $x = (x^0, x^1, x^2, x^3) = (t, \mathbf{x})$, and $\mu = 0, 1, 2, 3, 4$. Consider the Lagrangian

$$\mathcal{L} = \partial^\mu\phi\partial_\mu\phi - \frac{\lambda^2}{2}[\phi^2 - \eta^2]^2 \,. \tag{5.3}$$

Here again we have a Z_2 symmetry $\phi \to -\phi$. The crucial difference is that the tunneling amplitude in this case is proportional to $\exp(-cV)$ with $c > 0$ a constant and V the spatial volume. This result can be understood physically by discretizing space, so that our field theory corresponds to a quantum mechanical system in which for each spatial point $\mathbf{x}$ we have a variable $q_{\mathbf{x}}(t) \equiv \phi(t, \mathbf{x})$, and in order to tunnel into the other vacuum each of the $q_{\mathbf{x}}$ must tunnel. Let the tunneling amplitude for a single variable $q_{\mathbf{x}}$ be proportional to $\exp(-c')$, for some constant $c' > 0$. The total amplitude is the product of the separate amplitudes so, if N is the number of lattice sites we find that the tunneling amplitude is proportional to $\prod_{\mathbf{x}} \exp(-c') = \exp(-c'N) = \exp(-cV)$. In an infinite volume this amplitude vanishes and there is no mixing between the two vacua. In other words, the effective height of the barrier is infinite and therefore we truly have two distinct sectors of the theory, i.e. two different Hilbert spaces H_+, H_- constructed above the two vacua

$|\pm\rangle$. There is no possibility to restore the symmetry via tunneling, and all local operators have vanishing matrix elements between a state in H_+ and a state in H_-.

A characteristic of SSB is the existence of an order parameter which takes a non-zero expectation value on the chosen vacuum. In the example of Jellium the order parameter is the Fourier transform of the density while in the previous example it was an element of Z_2, $\langle\phi\rangle/\eta = \pm 1$. In the following we will be interested in situations where the order parameter is a scalar field ϕ, real or complex. In any case, the order parameter is a quantity which is not invariant under the symmetry in question, so that a non-vanishing expectation value means that the symmetry is broken.

For a Lie group, we can restate the condition of SSB in terms of the action of the generators on the vacuum state. We denote by $U = \exp(i\theta^a T^a)$ a generic element of the symmetry group in question, and by T^a the generators. If the vacuum state is invariant, then for any value of the parameters θ^a we have $U|0\rangle = \exp(i\alpha)|0\rangle$ for some constant phase α and therefore all generators must annihilate the vacuum, so $T^a|0\rangle = 0$. Conversely, in order to have SSB, beside $T^a|0\rangle \neq 0$ we must also require that $T^a|0\rangle$ is not proportional to $|0\rangle$ itself.

5.1.2 SSB of global symmetries: Goldstone bosons

Consider the Lagrangian for a complex scalar field

$$\mathcal{L} = \partial^\mu\phi^*\partial_\mu\phi - V(|\phi|)\,, \tag{5.4}$$

where

$$V(|\phi|) = \frac{\lambda^2}{2}[|\phi|^2 - \eta^2]^2\,, \tag{5.5}$$

This is a double-well potential for $|\phi|$ and therefore it has a continuous set of minima; writing $\phi = |\phi|\exp(i\alpha)$, the vacua are characterized by $\langle|\phi|\rangle = \eta$ and $\langle\alpha\rangle$ arbitrary. The Lagrangian has a *global* $U(1)$ invariance $\phi \to \exp(i\theta)\phi$ with θ an arbitrary constant. The scalar field will choose one of these vacua, so that $\langle\alpha\rangle = \alpha_0$, and the $U(1)$ symmetry is spontaneously broken. Without loss of generality we can redefine α so that $\alpha_0 = 0$, and therefore on the vacuum $\langle\phi\rangle = \eta$. We want to understand the spectrum of the theory after SSB. This can be done studying the small oscillations around the vacuum. We therefore write

$$\phi(x) = \eta + [\chi(x) + i\psi(x)]/\sqrt{2}\,, \tag{5.6}$$

where χ and ψ are real fields. Observe that the set of vacua is a circle of radius η in the complex field plane, and since we are expanding around the point $(\mathrm{Re}\,\phi = \eta, \mathrm{Im}\,\phi = 0)$, χ is a fluctuation in the direction orthogonal to the manifold of vacua, while ψ is a fluctuation in the tangential direction, as shown in Fig. 5.2. In other words, $\eta + i\psi$, for ψ constant and infinitesimal, is another vacuum. A small displacement in the direction of ψ does not cost energy since we are moving along a flat direction of the potential (at least to lowest order, i.e. retaining terms quadratic in ψ in the Lagrangian and neglecting cubic and higher-order terms). Instead with a small displacement in the direction of χ we feel an approximately quadratic rise of the potential, so this fluctuation costs energy. It is therefore clear that, after quantization, ψ is associated to a massless mode, while χ is a massive mode. To check this formally, we insert Eq. (5.6) into the Lagrangian (5.4), and we find

$$\mathcal{L} = \frac{1}{2}\partial^\mu\chi\partial_\mu\chi + \frac{1}{2}\partial^\mu\psi\partial_\mu\psi - \frac{\lambda^2}{8}[(2\sqrt{2}\eta)\chi + \chi^2 + \psi^2]^2\,, \tag{5.7}$$

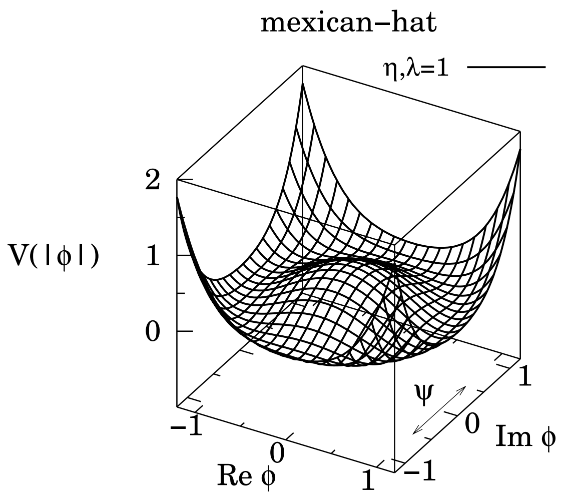


Figure 5.2: The mexican-hat potential of Eq. (5.5). Also shown is the direction in field space parametrized by ψ the Goldstone mode. The Higgs mode χ (not shown) is orthogonal to ψ.

We see indeed that in this Lagrangian there is a mass term for χ, $m_\chi^2/2 = (\lambda^2/8)(2\sqrt{2}\eta)^2 = \lambda^2\eta^2$. but there is no term of the form $(1/2)m_\psi^2\psi^2$, so ψ is massless. In conclusion, in this model the $U(1)$ symmetry is spontaneously broken by the choice of vacuum, and at the same time a massless spin-0 boson appears in the spectrum. This is an example of a general theorem, the Goldstone theorem, which states that, given a field theory which is Lorentz invariant, local, and has a Hilbert space with a positive definite scalar product, if a continuous global symmetry is spontaneously broken, then in the expansion around the symmetry-breaking vacuum there appears a massless particle for each generator that breaks the symmetry. This particle is called a Goldstone (or Nambu–Goldstone) particle.

As in the above example, also in the general case the emergence of massless particles corresponds to the possibility of moving, in field space, in the direction of the manifold of vacua. The dimensionality of the manifold of vacua is equal to the number of generators which break the symmetry. In fact, setting the vacuum energy to zero, by definition we have $\mathcal{H}|0\rangle = 0$. Since T^a is the generator of a symmetry transformation, it satisfies $[T^a, \mathcal{H}] = 0$ and therefore $\mathcal{H}(T^a|0\rangle) = T^a\mathcal{H}|0\rangle = 0$. So, if $T^a|0\rangle \neq 0$ (and if it is not proportional to $|0\rangle$ itself) we have found a new state with the minimum energy, i.e. another vacuum state. This is the origin of the fact that we have a Goldstone particle for each generator which breaks the symmetry.

The Goldstone theorem further states that the quantum numbers of the Goldstone particles are the same as the corresponding generator. In most cases, the global symmetry transformations are internal transformations in the field space which do not act on the Lorentz indices of the fields. For instance, in the above example the symmetry which is broken is $U(1)$, or, equivalently, an $O(2)$ rotation symmetry in the space $(\mathrm{Re}\,\phi, \mathrm{Im}\,\phi)$. These rotations do not touch Lorentz indices, and therefore the generators are Lorentz scalars. Correspondingly, the associated massless particle is a spin-0 boson (however, in supersymmetry the generators exchange fermions with bosons and carry half integer spin. As a consequence, the Goldstone particles associated to global supersymmetry breaking are fermions).

5.1.3 SSB of local symmetries: Abelian gauge theories

To illustrate the effect of SSB on a theory with a local symmetry we start again from the Lagrangian (5.4), but now we gauge the $U(1)$ symmetry. Therefore we introduce a $U(1)$ gauge field A_μ and we take as Lagrangian

$$\mathcal{L} = (D^\mu \phi)^* D_\mu \phi - V(|\phi|) - \frac{1}{4} F_{\mu\nu} F^{\mu\nu} \,, \tag{5.8}$$

with $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$ and

$$D_\mu \phi = (\partial_\mu + iqA_\mu)\phi \,. \tag{5.9}$$

To understand the physical content of the theory, it is convenient to write the complex field ϕ in terms of its modulus and a phase, and to expand the modulus around η,

$$\phi(x) = |\phi(x)| e^{i\alpha(x)} = \left(\eta + \frac{1}{\sqrt{2}}\varphi(x)\right) e^{i\alpha(x)} \,. \tag{5.10}$$

Now observe that, since under the $U(1)$ *local* transformation ϕ transforms as $\phi(x) \to \exp[iq\theta(x)]\phi(x)$, with $\theta(x)$ the parameter of the gauge transformation, we can fix the gauge freedom setting $\alpha(x) = 0$ in Eq. (5.10). In other words, we have used the gauge freedom to remove one degree of freedom from the complex field ϕ, so that we are left with just a single real scalar field φ. The phase $\alpha(x)$ parametrizes the manifold of vacua, so it is the field that, in the case of global symmetries, describes the Goldstone boson. We see that when we break a local symmetry the Goldstone boson is eliminated from the physical spectrum by gauge invariance. After setting $\alpha(x) = 0$, using Eqs. (5.9) and (5.10) we get

$$D_\mu \phi = \frac{1}{\sqrt{2}} \partial_\mu \varphi + iq \left(\eta + \frac{1}{\sqrt{2}}\varphi(x)\right) A_\mu \,, \tag{5.11}$$

and, substituting into Eq. (5.8)

$$\mathcal{L} = \frac{1}{2}\partial^\mu\varphi\partial_\mu\varphi - \lambda^2 \left(\eta^2\varphi^2 + \frac{1}{\sqrt{2}}\eta\varphi^3 + \frac{1}{8}\varphi^4\right) + q^2 \left(\eta + \frac{1}{\sqrt{2}}\varphi(x)\right)^2 A_\mu A^\mu - \frac{1}{4} F_{\mu\nu} F^{\mu\nu} \,, \tag{5.12}$$

We recognize a standard kinetic term for a real massive scalar field φ, with mass $m_\varphi^2 = 2\lambda^2\eta^2$. For the gauge field, the quadratic term is now

$$\mathcal{L}_A = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} + \frac{1}{2} m_A^2 A_\mu A^\mu \,, \tag{5.13}$$

with $m_A^2 = 2q^2\eta^2$. Taking the variation of Eq. (5.13) we find the equation of motion $\partial_\mu F^{\mu\nu} + m_A^2 A^\nu = 0$ which is known as the Proca equation. Contracting it with ∂_ν we have $\partial_\mu\partial_\nu F^{\mu\nu} + m_A^2 \partial_\nu A^\nu = 0$. Since $\partial_\mu\partial_\nu F^{\mu\nu} = 0$ and $m_A \neq 0$, we find $\partial_\nu A^\nu = 0$. Using this condition $\partial_\mu F^{\mu\nu}$ becomes $\Box A^\nu$ and the Proca equation reduces to $(\Box + m_A^2)A^\nu = 0$. Therefore the Proca equation describes a massive spin-1 gauge boson.

From this example we learn that the spontaneous breaking of a local symmetry does not produce Goldstone bosons, but instead the gauge field has acquired a mass proportional to the vacuum expectation value of the scalar field. In this context the scalar field ϕ is called a *Higgs*

field, and the mechanism that produces a mass for the gauge boson is called the *Higgs mechanism*.
It is interesting to compare the number of degrees of freedom with and without SSB. If in the potential we set $\eta = 0$, then there is no SSB; the scalar field has two real components. We cannot use the gauge invariance to eliminate the phase θ as before, because when $\eta = 0$ the decomposition (5.10) of ϕ in terms of two real fields φ, θ is not well defined: in fact in this case $\varphi = \sqrt{2}|\phi|$ and therefore $\varphi \geqslant 0$, so it is no longer a scalar field which can freely perform at least infinitesimal fluctuations around $\varphi = 0$. Rather, gauge invariance can be used to eliminate the longitudinal components of A_μ and the remaining gauge field has two physical degrees of freedom, the two transverse polarizations. In total we have two physical degrees of freedom from the Higgs field and two from the gauge field. After SSB, the scalar field has just one real component, but the gauge field is massive, and a massive spin-1 particle has three degrees of freedom. In total, we have $1 + 3 = 4$ degrees of freedom. So, the Higgs mechanism implies a reshuffling of the degrees of freedom. The field that, in the case of a global symmetry, was a Goldstone boson, is turned into the third polarization state of a massive spin-1 particle.

One might ask what do we really gain by giving a mass to the gauge boson with the Higgs mechanism, rather than adding by hand a mass term $(1/2)m_A^2 A_\mu A^\mu$ to the Lagrangian (a mass term for the gauge fieldgenerated by SSB is called a soft mass term, in contrast to a term added by hand, which is called a hard mass term). The point is that, in the Higgs mechanism, the Lagrangian is gauge invariant, which is not the case if we instead add by hand a mass term. The breaking of the symmetry takes place at the level of the vacuum. It can be shown that such a spontaneous breaking preserves a number of good properties of the unbroken theory, and in particular the theory is still renormalizable. Intuitively, this comes from the fact that at very high energies $E \gg \eta$ we can neglect η and the ultraviolet properties of the theory are the same as in the case $\eta = 0$. If we instead break the gauge symmetry by hand the theory is not renormalizable.

5.2 The experiment

Although the Higgs field would exist everywhere, proving its existence was far from easy. In principle, it can be proved to exist by detecting its excitations, which manifest as Higgs particles (the Higgs boson), but these are extremely difficult to produce and detect, due to the energy required to produce them and their very rare production even if the energy is sufficient. It was therefore several decades before the first evidence of the Higgs boson could be found. Particle colliders, detectors, and computers capable of looking for Higgs bosons took more than 30 years (c. 1980–2010) to develop.

The importance of this fundamental question led to a 40-year search, and the construction of one of the world's most expensive and complex experimental facilities to date, CERN's Large Hadron Collider, in an attempt to create Higgs bosons and other particles for observation and study. On 4 July 2012, the discovery of a new particle with a mass between 125 and 127 GeV/c^2 was announced; physicists suspected that it was the Higgs boson. Since then, the particle has been shown to behave, interact, and decay in many of the ways predicted for Higgs particles by the Standard Model, as well as having even parity and zero spin [51], two fundamental attributes of a Higgs boson. This also means it is the first elementary scalar particle discovered in nature.

By March 2013, the existence of the Higgs boson was confirmed, and therefore, the concept of some type of Higgs field throughout space is strongly supported. On 14 March 2013 CERN confirmed the following:

“CMS and ATLAS have compared a number of options for the spin-parity of this particle, and these all prefer no spin and even parity [two fundamental criteria of a Higgs boson consistent with the Standard Model]. This, coupled with the measured interactions of the new particle with other particles, strongly indicates that it is a Higgs boson."

This also makes the particle the first elementary scalar particle to be discovered in nature.

The presence of the field, now confirmed by experimental investigation, explains why some fundamental particles have mass, despite the symmetries controlling their interactions implying that they should be massless. It also resolves several other long-standing puzzles, such as the reason for the extremely short distance travelled by the weak force bosons, and therefore the weak force's extremely short range.

As of 2018, in-depth research shows the particle continuing to behave in line with predictions for the Standard Model Higgs boson. More studies are needed to verify with higher precision that the discovered particle has all of the properties predicted, or whether, as described by some theories, multiple Higgs bosons exist.

The nature and properties of this field are now being investigated further, using more data collected at the LHC.

Chapter 6

The Casimir effect

In quantum field theory, the *Casimir effect* is a physical force acting on the macroscopic boundaries of a confined space (see Fig. 6.1) which arises from the quantum fluctuations of the field. It is named after the Dutch physicist Hendrik Casimir, who predicted the effect for electromagnetic systems in 1948.

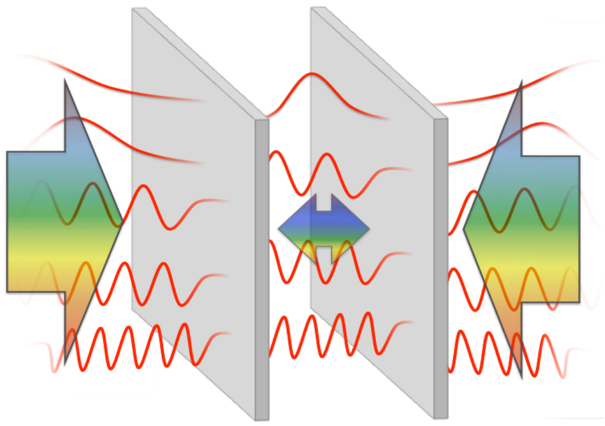


Figure 6.1: Casimir forces on parallel metal plates.

It was not until 1997 that a direct experiment by S. Lamoreaux quantitatively measured the Casimir force to within 5% of the value predicted by the theory [52].

The Casimir effect can be understood by the idea that the presence of macroscopic material interfaces, such as conducting metals and dielectrics, alters the vacuum expectation value of the energy of the second-quantized electromagnetic field. Since the value of this energy depends on the shapes and positions of the materials, the Casimir effect manifests itself as a force between such objects.

The typical example is of the two uncharged conductive plates in a vacuum, placed a few nanometers apart. In a classical description, the lack of an external field means that there is no field between the plates, and no force would be measured between them [53]. When this field is instead studied using the quantum electrodynamic vacuum, it is seen that the plates do affect the virtual photons which constitute the field, and generate a net force – either an attraction or a repulsion depending on the specific arrangement of the two plates. Although the Casimir effect can be expressed in terms of virtual particles interacting with the objects, it is best described and more easily calculated in terms of the zero-point energy of a quantized field in the intervening space between the objects. This force has been measured and is a striking example of an effect captured formally by second quantization [54].

The treatment of boundary conditions in these calculations has led to some controversy. In fact, "Casimir's original goal was to compute the van der Waals force between polarizable molecules" of the conductive plates. Thus it can be interpreted without any reference to the zero-point energy (vacuum energy) of quantum fields [55].

Because the strength of the force falls off rapidly with distance, it is measurable only when the distance between the objects is extremely small. On a submicron scale, this force becomes so strong that it becomes the dominant force between uncharged conductors. In fact, at separations of 10 nm – about 100 times the typical size of an atom – the Casimir effect produces the equivalent of about 1 atmosphere of pressure (the precise value depending on surface geometry and other factors) [54].

There are few instances wherein the Casimir effect can give rise to repulsive forces between uncharged objects. Evgeny Lifshitz showed (theoretically) that in certain circumstances (most commonly involving liquids), repulsive forces can arise [56]. Timothy Boyer showed in his work published in 1968 [57] that a conductor with spherical symmetry will also show this repulsive force, and the result is independent of radius. Further work shows that the repulsive force can be generated with materials of carefully-chosen dielectrics [58].

6.1 The mathematical observation

The causes of the Casimir effect are described by quantum field theory, which states that all of the various fundamental fields, such as the electromagnetic field, must be quantized at each and every point in space. In a simplified view, a "field" in physics may be envisioned as if space were filled with interconnected vibrating balls and springs, and the strength of the field can be visualized as the displacement of a ball from its rest position. Vibrations in this field propagate and are governed by the appropriate wave equation for the particular field in question. The second quantization of quantum field theory requires that each such ball-spring combination be quantized, that is, that the strength of the field be quantized at each point in space. At the most basic level, the field at each point in space is a simple harmonic oscillator, and its quantization places a quantum harmonic oscillator at each point. Excitations of the field correspond to the elementary particles of particle physics. However, even the vacuum has a vastly complex structure, so all calculations of quantum field theory must be made in relation to this model of the vacuum.

The vacuum has, implicitly, all of the properties that a particle may have: spin [59], or polarization in the case of light, energy, and so on. On average, most of these properties cancel out: the vacuum is, after all, "empty" in this sense. One important exception is the vacuum energy or the vacuum expectation value of the energy. The quantization of a simple harmonic oscillator states that the lowest possible energy or zero-point energy that such an oscillator may have is

$$E = \frac{1}{2}\hbar\omega\,. \tag{6.1}$$

Summing over all possible oscillators at all points in space gives an infinite quantity. Since only differences in energy are physically measurable (with the notable exception of gravitation, which remains beyond the scope of quantum field theory), this infinity may be considered a feature of the mathematics rather than of the physics. This argument is the underpinning of the theory of renormalization. Dealing with infinite quantities in this way was a cause of widespread unease among quantum field theorists before the development in the 1970s of the renormalization group, a mathematical formalism for scale transformations that provides a natural basis for the process.

When the scope of the physics is widened to include gravity, the interpretation of this formally infinite quantity remains problematic. There is currently no compelling explanation as to why it should not result in a cosmological constant that is many orders of magnitude larger than observed [60]. However, since we do not yet have any fully coherent quantum theory of gravity, there is likewise no compelling reason as to why it should instead actually result in the value of the cosmological constant that we observe.

In the original calculation done by Casimir, he considered the space between a pair of conducting metal plates at distance a apart. In this case, the standing waves are particularly easy to calculate, because the transverse component of the electric field and the normal component of the magnetic field must vanish on the surface of a conductor. Assuming the plates lie parallel to the xy-plane, the standing waves are

$$\psi_n(x,y,z;t) = e^{-i\omega_n t} e^{ik_x x + ik_y y} \sin(k_n z) \,, \tag{6.2}$$

where ψ stands for the electric component of the electromagnetic field, and, for brevity, the polarization and the magnetic components are ignored here. Here, k_x and k_y are the wave numbers in directions parallel to the plates, and $k_n = \frac{n\pi}{a}$ is the wave-number perpendicular to the plates. Here, n is an integer, resulting from the requirement that ψ vanish on the metal plates. The frequency of this wave is

$$\omega_n = c\sqrt{{k_x}^2 + {k_y}^2 + \frac{n^2\pi^2}{a^2}} \,, \tag{6.3}$$

where c is the speed of light. The vacuum energy is then the sum over all possible excitation modes. Since the area of the plates is large, we may sum by integrating over two of the dimensions in k-space. The assumption of periodic boundary conditions yields,

$$\langle E \rangle = \frac{\hbar}{2} 2 \int \frac{A dk_x dk_y}{(2\pi)^2} \sum_{n=1}^{\infty} \omega_n \,, \tag{6.4}$$

where A is the area of the metal plates, and a factor of 2 is introduced for the two possible polarizations of the wave. This expression is clearly infinite, and to proceed with the calculation, it is convenient to introduce a regulator (discussed in greater detail below). The regulator will serve to make the expression finite, and in the end will be removed. The zeta-regulated version of the energy per unit-area of the plate is

$$\frac{\langle E(s) \rangle}{A} = \hbar \int \frac{dk_x dk_y}{(2\pi)^2} \sum_{n=1}^{\infty} \omega_n \omega_n^{-s} \,. \tag{6.5}$$

In the end, the limit $s \to 0$ is to be taken. Here s is just a complex number. This integral/sum is finite for s real and larger than 3. The sum has a pole at $s = 3$, but may be analytically continued to $s = 0$, where the expression is finite. The above expression simplifies to:

$$\frac{\langle E(s) \rangle}{A} = \frac{\hbar c^{1-s}}{4\pi^2} \sum_n \int_0^\infty 2\pi q dq \left(q^2 + \frac{\pi^2 n^2}{a^2} \right)^{(1-s)/2} \,, \tag{6.6}$$

where polar coordinates $q^2 = k_x^2 + k_y^2$ were introduced to turn the double integral into a single integral. The q in front is the Jacobian, and the 2π comes from the angular integration. The integral converges if Re$[s] > 3$, resulting in

$$\frac{\langle E(s)\rangle}{A} = -\frac{\hbar c^{1-s}\pi^{2-s}}{2a^{3-s}(3-s)}\sum_n \frac{1}{n^{s-3}}\,. \tag{6.7}$$

The sum diverges for s in a neighborhood of zero, but if the damping of large-frequency excitations corresponding to analytic continuation of the Riemann zeta function to $s = 0$ is assumed to make sense physically in some way, then one has

$$\frac{\langle E\rangle}{A} = \lim_{s\to 0}\frac{\langle E(s)\rangle}{A} = -\frac{\hbar c\pi^2}{6a^3}\zeta(-3)\,. \tag{6.8}$$

But $\zeta(-3) = \dfrac{1}{120}$ and so one obtains

$$\frac{\langle E\rangle}{A} = \frac{-\hbar c\pi^2}{720a^3}\,. \tag{6.9}$$

The analytic continuation has evidently lost an additive positive infinity, somehow exactly accounting for the zero-point energy (not included above) outside the slot between the plates, but which changes upon plate movement within a closed system. The Casimir force per unit area F_c/A for idealized, perfectly conducting plates with vacuum between them is

$$\frac{F_c}{A} = -\frac{d}{da}\frac{\langle E\rangle}{A} = -\frac{\hbar c\pi^2}{240a^4}\,. \tag{6.10}$$

The force is negative, indicating that the force is attractive: by moving the two plates closer together, the energy is lowered. The presence of $\hbar$ shows that the Casimir force per unit area F_c/A is very small, and that furthermore, the force is inherently of quantum-mechanical origin.

By integrating the equation above it is possible to calculate the energy required to separate to infinity the two plates as:

$$U_E(a) = \int F(a)\,da = \int -\hbar c\pi^2\frac{A}{240a^4}\,da = \hbar c\pi^2\frac{A}{720a^3}\,. \tag{6.11}$$

In Casimir's original derivation [61], a moveable conductive plate is positioned at a short distance a from one of two widely separated plates (distance L apart). The 0-point energy on both sides of the plate is considered. Instead of the above ad hoc analytic continuation assumption, non-convergent sums and integrals are computed using Euler–Maclaurin summation with a regularizing function (e.g., exponential regularization) not so anomalous as ω_n^{-s} in the above.

6.2 The experiment

One of the first experimental tests was conducted by Marcus Sparnaay at Philips in Eindhoven (Netherlands), in 1958, in a delicate and difficult experiment with parallel plates, obtaining results not in contradiction with the Casimir theory [62, 63], but with large experimental errors.

The Casimir effect was measured more accurately in 1997 by Steve K. Lamoreaux of Los Alamos National Laboratory [52], and by Umar Mohideen and Anushree Roy of the University of California, Riverside.[31] In practice, rather than using two parallel plates, which would require phenomenally accurate alignment to ensure they were parallel, the experiments use one plate that is flat and another plate that is a part of a sphere with a very large radius.

In 2001, a group (Giacomo Bressi, Gianni Carugno, Roberto Onofrio and Giuseppe Ruoso) at the University of Padua (Italy) finally succeeded in measuring the Casimir force between parallel plates using microresonators [64].

In 2013, a conglomerate of scientists from Hong Kong University of Science and Technology, University of Florida, Harvard University, Massachusetts Institute of Technology, and Oak Ridge National Laboratory demonstrated a compact integrated silicon chip that can measure the Casimir force [65]. The integrated chip defined by electron-beam lithography does not need extra alignment, making it an ideal platform for measuring Casimir force between complex geometries. In 2017 and 2021, the same group from Hong Kong University of Science and Technology demonstrated the non-monotonic Casimir force [66] and distance-independent Casimir force [67], respectively, using this on-chip platform.

Chapter 7

The Aharonov-Bohm effect

The *Aharonov-Bohm effect*, sometimes called the Ehrenberg-Siday-Aharonov-Bohm effect, is a quantum mechanical phenomenon in which an electrically charged particle is affected by an electromagnetic potential $(\phi, \mathbf{A})$, despite being confined to a region in which both the magnetic field $\mathbf{B}$ and electric field $\mathbf{E}$ are zero [68]. The underlying mechanism is the coupling of the electromagnetic potential with the complex phase of a charged particle's wave function, and the Aharonov–Bohm effect is accordingly illustrated by interference experiments as shown in Fig. 7.1.

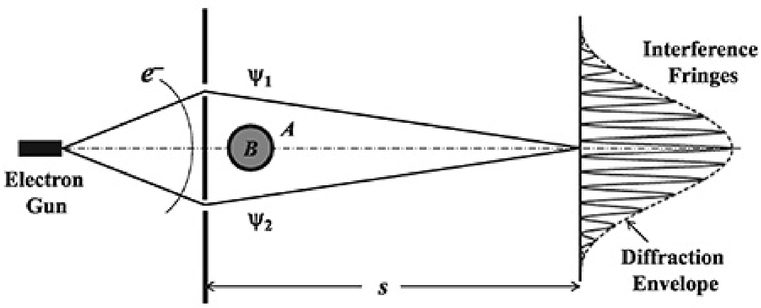


Figure 7.1: Aharonov-Bohm effect illustrated by interference experiments.

The most commonly described case, sometimes called the Aharonov-Bohm solenoid effect, takes place when the wave function of a charged particle passing around a long solenoid experiences a phase shift as a result of the enclosed magnetic field, despite the magnetic field being negligible in the region through which the particle passes and the particle's wavefunction being negligible inside the solenoid. This phase shift has been observed experimentally [69]. There are also magnetic Aharonov-Bohm effects on bound energies and scattering cross sections, but these cases have not been experimentally tested. An electric Aharonov–Bohm phenomenon was also predicted, in which a charged particle is affected by regions with different electrical potentials but zero electric field, but this has no experimental confirmation yet [69]. A separate "molecular" Aharonov-Bohm effect was proposed for nuclear motion in multiply connected regions, but this has been argued to be a different kind of geometric phase as it is "neither nonlocal nor topological", depending only on local quantities along the nuclear path [70].

Werner Ehrenberg (1901–1975) and Raymond E. Siday first predicted the effect in 1949 [71]. Yakir Aharonov and David Bohm published their analysis in 1959 [68]. After publication of the 1959 paper, Bohm was informed of Ehrenberg and Siday's work, which was acknowledged and credited in Bohm and Aharonov's subsequent 1961 paper [72]. The effect was confirmed experimentally, with a very large error, while Bohm was still alive. By the time the error was down to a respectable value, Bohm had died [73].

7.1 The mathematical observation

The Lagrangian for a charged particle of mass m and charge q in an electromagnetic field equivalently describes the dynamics of the particle in terms of its energy, rather than the force exerted on it. The classical expression is given by:

$$L = \frac{m}{2}\dot{\mathbf{r}}\cdot\dot{\mathbf{r}} + q\mathbf{A}\cdot\dot{\mathbf{r}} - q\phi\,, \tag{7.1}$$

where $\mathbf{A}$ and ϕ are the potential fields as above such that

$$\begin{aligned} \mathbf{E} &= -\nabla\phi - \frac{\partial \mathbf{A}}{\partial t}\,, &(7.2)\\ \mathbf{B} &= \nabla\times\mathbf{A}\,, &(7.3) \end{aligned}$$

where $\mathbf{E}$ and $\mathbf{B}$ are the electric and magnetic fields respectively. The quantity $V = q(\phi - \mathbf{A}\cdot\dot{\mathbf{r}})$ can be thought as a velocity-dependent potential function. Using Lagrange's equations, the equation for the Lorentz force given above can be obtained as follows.

Lagrange's equations are

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{x}} = \frac{\partial L}{\partial x}\,, \tag{7.4}$$

(same for y and z). So calculating the partial derivatives:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{x}} &= m\ddot{x} + q\frac{\mathrm{d}A_x}{\mathrm{d}t} &(7.5)\\ &= m\ddot{x} + q\left(\frac{\partial A_x}{\partial t} + \frac{\partial A_x}{\partial x}\dot{x} + \frac{\partial A_x}{\partial y}\dot{y} + \frac{\partial A_x}{\partial z}\dot{z}\right)\,, &(7.6)\\ \frac{\partial L}{\partial x} &= -q\frac{\partial \phi}{\partial x} + q\left(\frac{\partial A_x}{\partial x}\dot{x} + \frac{\partial A_y}{\partial x}\dot{y} + \frac{\partial A_z}{\partial x}\dot{z}\right)\,, &(7.7) \end{aligned}$$

equating and simplifying:

$$m\ddot{x} + q\left(\frac{\partial A_x}{\partial t} + \frac{\partial A_x}{\partial x}\dot{x} + \frac{\partial A_x}{\partial y}\dot{y} + \frac{\partial A_x}{\partial z}\dot{z}\right) = -q\frac{\partial \phi}{\partial x} + q\left(\frac{\partial A_x}{\partial x}\dot{x} + \frac{\partial A_y}{\partial x}\dot{y} + \frac{\partial A_z}{\partial x}\dot{z}\right)\,, \tag{7.8}$$

$$\begin{aligned} F_x &= -q\left(\frac{\partial \phi}{\partial x} + \frac{\partial A_x}{\partial t}\right) + q\left[\dot{y}\left(\frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y}\right) + \dot{z}\left(\frac{\partial A_z}{\partial x} - \frac{\partial A_x}{\partial z}\right)\right] &(7.9)\\ &= qE_x + q[\dot{y}(\nabla\times\mathbf{A})_z - \dot{z}(\nabla\times\mathbf{A})_y] &(7.10)\\ &= qE_x + q[\dot{\mathbf{r}}\times(\nabla\times\mathbf{A})]_x &(7.11)\\ &= qE_x + q(\dot{\mathbf{r}}\times\mathbf{B})_x\,, &(7.12) \end{aligned}$$

and similarly for the y and z directions. Hence the force equation is:

$$\mathbf{F} = q(\mathbf{E} + \dot{\mathbf{r}} \times \mathbf{B}), \tag{7.13}$$

which is the well known Lorentz force. The canonical or conjugate momenta are then

$$p_i = \frac{\partial L}{\partial \dot{r}_i} \quad i = 1, 2, 3, \tag{7.14}$$

where $\mathbf{r} = (r_1, r_2, r_3) = (x, y, z)$. So that $\mathbf{p} = m\mathbf{v} + q\mathbf{A}$. Or in quantum mechanics $\mathbf{p} = -i\hbar\nabla + q\mathbf{A}$.

The magnetic Aharonov-Bohm effect can be seen as a result of the requirement that quantum physics be invariant with respect to the gauge choice for the electromagnetic potential ($\phi \to \phi - \partial f/\partial t, \mathbf{A} \to \mathbf{A} + \nabla f$ with f any twice continuously differentiable function that depends on position and time), of which the magnetic vector potential $\mathbf{A}$ forms part. Electromagnetic theory implies that a particle with electric charge q travelling along some path P in a region with zero magnetic field $\mathbf{B}$, but non-zero $\mathbf{A}$ (by $\mathbf{B} = \mathbf{0} = \nabla \times \mathbf{A}$), acquires a phase shift φ, given in SI units by

$$\varphi = \frac{q}{\hbar} \int_P \mathbf{A} \cdot d\mathbf{r}. \tag{7.15}$$

Therefore, particles, with the same start and end points, but travelling along two different routes will acquire a phase difference $\Delta\varphi$ determined by the magnetic flux $\Phi_{\mathbf{B}} = \iint_S \nabla \times \mathbf{A} \cdot d\mathbf{S}$ through the area between the paths (via Stokes' theorem and $\nabla \times \mathbf{A} = \mathbf{B}$), and given by:

$$\Delta\varphi = \frac{q\Phi_{\mathbf{B}}}{\hbar}. \tag{7.16}$$

In quantum mechanics the same particle can travel between two points by a variety of paths. Therefore, this phase difference can be observed by placing a solenoid between the slits of a double-slit experiment (or equivalent, see Fig. 7.1). An ideal solenoid (i.e. infinitely long and with a perfectly uniform current distribution) encloses a magnetic field $\mathbf{B}$, but does not produce any magnetic field outside of its cylinder, and thus the charged particle (e.g. an electron) passing outside experiences no magnetic field $\mathbf{B}$. However, there is a (curl-free) vector potential $\mathbf{A}$ outside the solenoid with an enclosed flux, and so the relative phase of particles passing through one slit or the other is altered by whether the solenoid current is turned on or off. This corresponds to an observable shift of the interference fringes on the observation plane.

The same phase effect is responsible for the quantized-flux requirement in superconducting loops. This quantization occurs because the superconducting wave function must be single valued: its phase difference $\Delta\varphi$ around a closed loop must be an integer multiple of 2π (with the charge $q = 2e$ for the electron Cooper pairs), and thus the flux must be a multiple of $h/2e$. The superconducting flux quantum was actually predicted prior to Aharonov and Bohm, by F. London in 1948 using a phenomenological model [74].

The Aharonov-Bohm effect is important conceptually because it bears on three issues apparent in the recasting of (Maxwell's) classical electromagnetic theory as a gauge theory, which before the advent of quantum mechanics could be argued to be a mathematical reformulation with no physical consequences. The Aharonov-Bohm thought experiments and their experimental realization imply that the issues were not just philosophical. The three issues are: *i.* whether potentials are "physical" or just a convenient mathematical tool for calculating force fields; *ii.* whether action principles are fundamental; *iii.* the principle of locality.

The Aharonov–Bohm effect shows that the local **E** and **B** fields do not contain full information about the electromagnetic field, and the electromagnetic four-potential, $(\phi, \mathbf{A})$, must be used instead. By Stokes' theorem, the magnitude of the Aharonov-Bohm effect can be calculated using the electromagnetic fields alone, or using the four-potential alone. But when using just the electromagnetic fields, the effect depends on the field values in a region from which the test particle is excluded. In contrast, when using just the electromagnetic four-potential, the effect only depends on the potential in the region where the test particle is allowed. Therefore, one must either abandon the principle of locality, which most physicists are reluctant to do, or accept that the electromagnetic four-potential offers a more complete description of electromagnetism than the electric and magnetic fields can. On the other hand, the Aharonov-Bohm effect is crucially quantum mechanical; quantum mechanics is well known to feature non-local effects (albeit still disallowing superluminal communication).

In classical electromagnetism the two descriptions were equivalent. With the addition of quantum theory, though, the electromagnetic potentials ϕ and **A** are seen as being more fundamental. Despite this, all observable effects end up being expressible in terms of the electromagnetic fields, **E** and **B**. This is interesting because, while you can calculate the electromagnetic field from the four-potential, due to gauge freedom the reverse is not true. Richard Feynman in "The Feynman Lectures on Physics" [75] Vol. 2, pp. 15-25 states:

"knowledge of the classical electromagnetic field acting locally on a particle is not sufficient to predict its quantum-mechanical behavior. [...] is the vector potential a "real" field? [...] a real field is a mathematical device for avoiding the idea of action at a distance. [...] for a long time it was believed that **A** was not a "real" field. [...] there are phenomena involving quantum mechanics which show that in fact **A** is a "real" field in the sense that we have defined it [...] **E** and **B** are slowly disappearing from the modern expression of physical laws; they are being replaced by **A** [the vector potential] and ϕ [the scalar potential]"

7.2 The experiment

The first claimed experimental confirmation was by Robert G. Chambers in 1960 [76, 77], in an electron interferometer with a magnetic field produced by a thin iron whisker, and other early work is summarized in Olariu and Popèscu (1984) [78]. However, subsequent authors questioned the validity of several of these early results because the electrons may not have been completely shielded from the magnetic fields [79]. An early experiment in which an unambiguous Aharonov-Bohm effect was observed by completely excluding the magnetic field from the electron path (with the help of a superconducting film) was performed by Tonomura et al. in 1986 [80]. The effect's scope and application continues to expand. Webb et al. (1985) [81] demonstrated Aharonov-Bohm oscillations in ordinary, non-superconducting metallic rings. Bachtold et al. (1999) [82] detected the effect in carbon nanotubes.

Chapter 8

The quasicrystal

A quasiperiodic crystal, or *quasicrystal*, is a structure that is ordered but not periodic. A quasicrystalline pattern can continuously fill all available space, but it lacks translational symmetry [83]. While crystals, according to the classical crystallographic restriction theorem, can possess only two-, three-, four-, and six-fold rotational symmetries, the Bragg diffraction pattern of quasicrystals shows sharp peaks with other symmetry orders – for instance, five-fold.

Aperiodic tilings were discovered by mathematicians in the early 1960s, and, some twenty years later, they were found to apply to the study of natural quasicrystals. The discovery of these aperiodic forms in nature has produced a paradigm shift in the field of crystallography. In crystallography the quasicrystals were predicted in 1981 by a five-fold symmetry study of Alan Lindsay Mackay [84], – that also brought in 1982, with the crystallographic Fourier transform of a Penrose tiling (see Fig. 8.1) [85], the possibility of identifying quasiperiodic order in a material through diffraction.

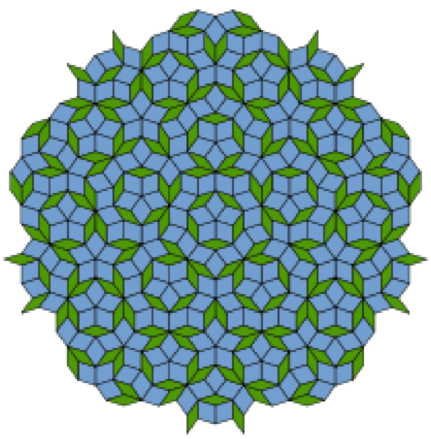


Figure 8.1: A Penrose tiling.

Quasicrystals had been investigated and observed earlier, but, until the 1980s, they were disregarded in favor of the prevailing views about the atomic structure of matter. For example in his 1944 book "What Is Life? The Physical Aspect of the Living Cell" [86], Erwin Schrödinger introduced the idea of an "aperiodic crystal" that contained genetic information in its configuration of covalent chemical bonds. In the 1950s, this idea stimulated enthusiasm for discovering the genetic molecule. Although the existence of some form of hereditary information had been hypothesized since 1869, its role in reproduction and its helical shape were still unknown at the time of Schrödinger's book. In retrospect, Schrödinger's aperiodic crystal can be viewed as a well-reasoned theoretical prediction of what biologists should have been looking for during their search for genetic material. In 1953 James D. Watson, and Francis Crick jointly proposed the double helix structure of DNA based on, amongst other theoretical insights, X-ray diffraction experiments by Rosalind Franklin. They both credited Schrödinger's book with presenting an early theoretical description of how the storage of genetic information would work, and each independently acknowledged the book as a source of inspiration for their initial researches.

In 2009, after a dedicated search, a mineralogical finding, icosahedrite, offered evidence for the existence of natural quasicrystals [87].

Roughly, an ordering is non-periodic if it lacks translational symmetry, which means that a shifted copy will never match exactly with its original. The more precise mathematical definition is that there is never translational symmetry in more than n – 1 linearly independent directions, where n is the dimension of the space filled, e.g., the three- dimensional tiling displayed in a quasicrystal may have translational symmetry in two directions. Symmetrical diffraction patterns result from the existence of an indefinitely large number of elements with a regular spacing, a property loosely described as long-range order. Experimentally, the aperiodicity is revealed in the unusual symmetry of the diffraction pattern, that is, symmetry of orders other than two, three, four, or six.

8.1 The mathematical observation

The crystallographic restriction theorem in its basic form was based on the observation that the rotational symmetries of a crystal are usually limited to 2-fold, 3-fold, 4-fold, and 6-fold.

Crystals are modeled as discrete lattices, generated by a list of independent finite translations. Because discreteness requires that the spacings between lattice points have a lower bound, the group of rotational symmetries of the lattice at any point must be a finite group (alternatively, the point is the only system allowing for infinite rotational symmetry). The strength of the theorem is that not all finite groups are compatible with a discrete lattice; in any dimension, we will have only a finite number of compatible groups.

The special cases of 2 dimensions (wallpaper groups) and 3 dimension (space groups) are most heavily used in applications, and they can be treated together (when the dimension of the lattice rises to four or more, rotations need no longer be planar). Consider two lattice points A and B separated by a translation vector $\mathbf{r}$. Consider an angle α such that a rotation of angle α about any lattice point is a symmetry of the lattice. Rotating about point B by α maps point A to a new point A'. Similarly, rotating about point A by α maps B to a point B'. Since both rotations mentioned are symmetry operations, A' and B' must both be lattice points. Due to periodicity of the crystal, the new vector $\mathbf{r}'$ which connects them must be equal to an integer multiple of $\mathbf{r}$: $\mathbf{r}' = m\mathbf{r}$, with m integer. The four points A, B, B', A' form a trapezium. Therefore, the length of $\mathbf{r}'$ is also given by: $r' = 2r\cos\alpha - r$. Combining the two equations gives:

$$\cos\alpha = \frac{m+1}{2} = \frac{M}{2}\,, \tag{8.1}$$

where $M = m + 1$ is also an integer. Bearing in mind that $|\cos\alpha| \leqslant 1$ we have allowed integers $M \in \{-2, -1, 0, 1, 2\}$. Solving for possible values of α reveals that the only values in the 0° to

$180°$ range are $0°, 60°, 90°, 120°$, and $180°$. In radians, the only allowed rotations consistent with lattice periodicity are given by $2\pi/n$, where $n = 1, 2, 3, 4, 6$. This corresponds to 1-, 2-, 3-, 4-, and 6-fold symmetry, respectively, and therefore excludes the possibility of 5-fold or greater than 6-fold symmetry.

A Penrose tiling is an example of an aperiodic tiling. Here, a tiling is a covering of the plane by non-overlapping polygons or other shapes, and aperiodic means that shifting any tiling with these shapes by any finite distance, without rotation, cannot produce the same tiling. However, despite their lack of translational symmetry, Penrose tilings may have both reflection symmetry and fivefold rotational symmetry. Penrose tilings are named after mathematician and physicist Roger Penrose, who investigated them in the 1970s. Penrose tilings are self-similar: they may be converted to equivalent Penrose tilings with different sizes of tiles, using processes called inflation and deflation. The pattern represented by every finite patch of tiles in a Penrose tiling occurs infinitely many times throughout the tiling. They are quasicrystals: implemented as a physical structure a Penrose tiling will produce diffraction patterns with Bragg peaks and five-fold symmetry, revealing the repeated patterns and fixed orientations of its tiles. The study of these tilings has been important in the understanding of physical materials that also form quasicrystals.

8.2 The experiment

In 1982 materials scientist Dan Shechtman observed that certain aluminium-manganese alloys produced the unusual diffractograms which today are seen as revelatory of quasicrystal structures. Due to fear of the scientific community's reaction, it took him two years to publish the results [88] for which he was awarded the Nobel Prize in Chemistry in 2011. "His discovery of quasicrystals revealed a new principle for packing of atoms and molecules," stated the Nobel Committee and pointed that "this led to a paradigm shift within chemistry." On 25 October 2018, Luca Bindi and Paul Steinhardt were awarded the Aspen Institute 2018 Prize for collaboration and scientific research between Italy and the United States, after they discovered icosahedrite, the first quasicrystal known to occur naturally [87]. Its composition is $Al_{63}Cu_{24}Fe_{13}$ and was approved by the International Mineralogical Association in 2010. Analysis indicates it may be meteoritic in origin, possibly delivered from a carbonaceous chondrite asteroid. It is also believed that natural quasicrystals are formed by rapid quenching of a meteorite heated during an impact-induced shock.

Chapter 9

Some non-scientific discoveries

The astronomer Galileo Galilei in his “Il Saggiatore” wrote that “[The universe] is written in the language of mathematics, and its characters are triangles, circles, and other geometric figures.” Artists who strive and seek to study nature must first, in Galileo’s view, fully understand mathematics. Mathematicians, conversely, have sought to interpret and analyse art through the lens of geometry and rationality. The mathematician Felipe Cucker suggests that mathematics, and especially geometry, is a source of rules for ‘rule-driven artistic creation’, though not the only one.

Mathematics and art are related in a variety of ways. Mathematics has itself been described as an art motivated by beauty. Mathematics can be discerned in arts such as music, dance, painting, architecture, sculpture, and textiles.

Mathematics and art have a long historical relationship. Artists have used mathematics since the 4th century BC when the Greek sculptor Polykleitoswrote his Canon, prescribing proportions conjectured to have been based on the ratio $1/\sqrt{2}$ for the ideal male nude. Persistent popular claims have been made for the use of the golden ratio in ancient art and architecture, without reliable evidence. In the Italian Renaissance, Luca Pacioli wrote the influential treatise “De divina proportione” (1509) [89], illustrated with woodcuts by Leonardo da Vinci, on the use of the golden ratio in art. Another Italian painter, Piero della Francesca, developed Euclid’s ideas on perspective in treatises such as “De Prospectiva Pingendi”, and in his paintings. The engraver Albrecht Dürer made many references to mathematics in his work “Melencolia I”. In modern times, the graphic artist M. C. Escher made intensive use of tessellation and hyperbolic geometry, with the help of the mathematician H. S. M. Coxeter, while the De Stijl movement led by Theo van Doesburg and Piet Mondrian explicitly embraced geometrical forms. Mathematics has inspired textile arts such as quilting, knitting, cross- stitch, crochet, embroidery, weaving, Turkish and other carpet-making, as well as kilim. In Islamic art, symmetries are evident in forms as varied as Persian girih and Moroccan zellige tilework, Mughal jali pierced stone screens, and widespread muqarnas vaulting.

Mathematics has directly influenced art with conceptual tools such as linear perspective, the analysis of symmetry, and mathematical objects such as polyhedra and the Möbius strip. Magnus Wenninger creates colourful stellated polyhedra, originally as models for teaching. Mathematical concepts such as recursion and logical paradox can be seen in paintings by René Magritte and in engravings by M. C. Escher. Computer art often makes use of fractals including the Mandelbrot set, and sometimes explores other mathematical objects such as cellular automata. Controversially, the artist David Hockney has argued that artists from the Renaissance onwards made use of the camera lucida to draw precise representations of scenes; the architect Philip Steadman similarly argued that Vermeer used the camera obscura in his distinctively observed paintings.

Other relationships include the algorithmic analysis of artworks by X-ray fluorescence spectroscopy, the finding that traditional batiks from different regions of Java have distinct fractal dimensions, and stimuli to mathematics research, especially Filippo Brunelleschi's theory of perspective, which eventually led to Girard Desargues's projective geometry. A persistent view, based ultimately on the Pythagorean notion of harmony in music, holds that everything was arranged by Number, that God is the geometer of the world, and that therefore the world's geometry is sacred (see Chapter 10).

In mathematics, two quantities are in the *golden ratio* if their ratio is the same as the ratio of their sum to the larger of the two quantities. Expressed algebraically, for quantities a and b with $a > b > 0$,

$$\frac{a+b}{a} = \frac{a}{b} =: \varphi \,, \tag{9.1}$$

where the Greek letter φ represents the golden ratio. It is an irrational number (see Fig. 9.1)

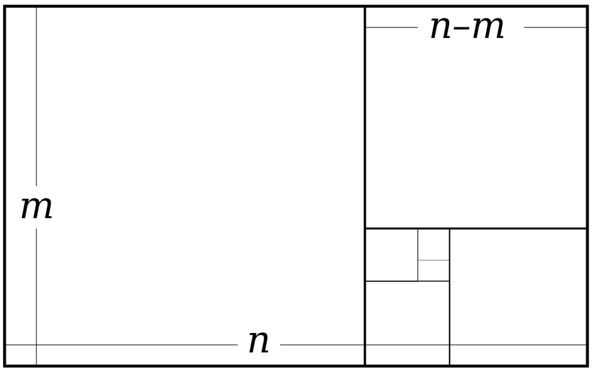


Figure 9.1: If φ were rational, then it would be the ratio of sides of a rectangle with integer sides (the rectangle comprising the entire diagram). But it would also be a ratio of integer sides of the smaller rectangle (the rightmost portion of the diagram) obtained by deleting a square. The sequence of decreasing integer side lengths formed by deleting squares cannot be continued indefinitely because the positive integers have a lower bound, so φ cannot be rational.

that is a solution to the quadratic equation $\varphi^2 - \varphi - 1 = 0$, which follows immediately from Eq. (9.1) with a value of

$$\varphi = \frac{1+\sqrt{5}}{2} = 1.618033988749\ldots\,. \tag{9.2}$$

The golden ratio is also called the golden mean or golden section (Latin: sectio aurea).Other names include extreme and mean ratio, medial section, divine proportion (Latin: proportio divina), divine section (Latin: sectio divina), golden proportion, golden cut, and golden number.

The mathematics of the golden ratio is intimately interconnected with that of the Fibonacci sequence. The Fibonacci numbers, commonly denoted F_n, form a sequence, the Fibonacci sequence, in which each number is the sum of the two preceding ones, $F_n = F_{n-1} + F_{n-2}$ for $n > 1$. The sequence commonly starts from $F_0 = 0$ and $F_1 = 1$. The Fibonacci numbers were first described in Indian mathematics, as early as 200 BC in work by Pingala on enumerating possible patterns of Sanskrit poetry formed from syllables of two lengths. They are named after the Italian mathematician Leonardo of Pisa, later known as Fibonacci, who introduced the sequence to Western European mathematics in his 1202 book Liber Abaci. They appear in biological settings, such as branching in trees, the arrangement of leaves on a stem, the fruit sprouts of a pineapple, the flowering of an artichoke, an uncurling fern, and the arrangement of a pine cone's bracts. It follows immediately that

$$\lim_{n\to\infty} \frac{F_{n+1}}{F_n} = \varphi \,. \tag{9.3}$$

This convergence was found in 1608 by Johannes Kepler and holds regardless of the starting values F_0 and F_1, unless $F_1 = -F_0/\varphi$. Like every sequence defined by a linear recurrence with constant coefficients, the Fibonacci numbers have a closed-form expression. It has become known as Binet's formula, named after French mathematician Jacques Philippe Marie Binet, though it was already known by Abraham de Moivre and Daniel Bernoulli:

$$F_n = \frac{\varphi^n - \psi^n}{\varphi - \psi} = \frac{\varphi^n - \psi^n}{\sqrt{5}} \,, \tag{9.4}$$

where $\psi = 1 - \varphi = -1/\varphi$ is the golden ratio conjugate.

The first known decimal approximation of the (inverse) golden ratio was stated as "about 0.6180340" in 1597 by Michael Maestlin of the University of Tübingen in a letter to Kepler, his former student. The same year, Kepler wrote to Maestlin of the Kepler triangle, which combines the golden ratio with the Pythagorean theorem. Kepler said of these: 'Geometry has two great treasures: one is the theorem of Pythagoras, the other the division of a line into extreme and mean ratio. The first we may compare to a mass of gold, the second we may call a precious jewel.'

Mathematicians since Euclid have studied the properties of the golden ratio, including its appearance in the dimensions of a regular pentagon and in a golden rectangle, which may be cut into a square and a smaller rectangle with the same aspect ratio. The golden ratio has also been used to analyze the proportions of natural objects. The golden ratio appears in some patterns in nature, including the spiral arrangement of leaves and other parts of vegetation.

The zome construction system (see Fig. 9.2), developed by Steve Baer in the late 1960s, is based on the symmetry system of the icosahedron/dodecahedron, and uses the golden ratio ubiquitously. Between 1973 and 1974, Roger Penrose developed Penrose tiling, a pattern related to the golden ratio both in the ratio of areas of its two rhombic tiles and in their relative frequency within the pattern. This led to Dan Shechtman's early 1980s discovery of quasicrystals, some of which exhibit icosahedral symmetry (see Chapter 8).

Some 20th-century artists and architects, including Le Corbusier and Salvador Dalí, have proportioned their works to approximate the golden ratio, believing this to be aesthetically pleasing. These often appear in the form of the golden rectangle, in which the ratio of the longer side to the shorter is the golden ratio. Le Corbusier, famous for his contributions to the modern international style, centered his design philosophy on systems of harmony and proportion. Le Corbusier's faith in the mathematical order of the universe was closely bound to the golden ratio and the Fibonacci series, which he described as "rhythms apparent to the eye and clear in their relations with one another. And these rhythms are at the very root of human activities. They resound in man by an organic inevitability, the same fine inevitability which causes the tracing out of the Golden Section by children, old men, savages and the learned."

Le Corbusier explicitly used the golden ratio in his Modulor system for the scale of architectural proportion. He saw this system as a continuation of the long tradition of Vitruvius,

Figure 9.2: Steve Baer's zome house.

Leonardo da Vinci's "Vitruvian Man", the work of Leon Battista Alberti, and others who used the proportions of the human body to improve the appearance and function of architecture.

In addition to the golden ratio, Le Corbusier based the system on human measurements, Fibonacci numbers, and the double unit. He took suggestion of the golden ratio in human proportions to an extreme: he sectioned his model human body's height at the navel with the two sections in golden ratio, then subdivided those sections in golden ratio at the knees and throat; he used these golden ratio proportions in the Modulor system. Le Corbusier's 1927 Villa Stein in Garches exemplified the Modulor system's application. The villa's rectangular ground plan, elevation, and inner structure closely approximate golden rectangles (see Fig. 9.3).

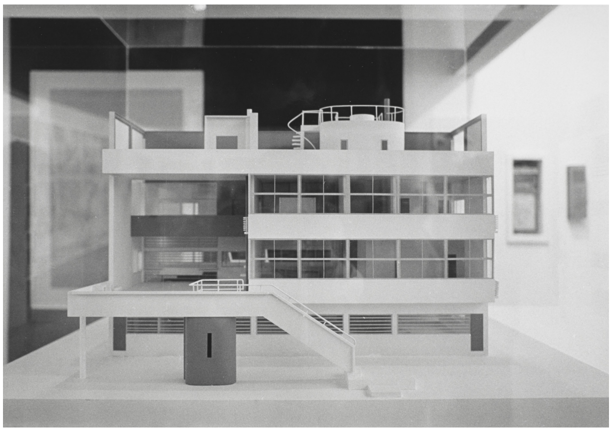


Figure 9.3: Installation view, "Le Corbusier 'Les Terrasses' The Villa Stein at Garches 1927" part of the exhibition, "Four Americans in Paris: The Collection of Gertrude Stein and her Family" (December 19, 1970-March 1, 1971. Photographic Archive. The Museum of Modern Art Archives, New York. Photograph by Ludwig Glaeser).

"Gödel, Escher, Bach: an Eternal Golden Braid", is a Pulitzer Prize-winning 1979 book by Douglas Hofstadter [90] which by exploring common themes in the lives and works of logician

Kurt Gödel, artist Maurits Cornelis Escher, and composer Johann Sebastian Bach, the book expounds concepts fundamental to mathematics, symmetry, and intelligence. Through short stories, illustrations, and analysis, the book discusses how systems can acquire meaningful context despite being made of "meaningless" elements. It also discusses self-reference and formal rules, isomorphism, what it means to communicate, how knowledge can be represented and stored, the methods and limitations of symbolic representation, and even the fundamental notion of 'meaning' itself.

In response to confusion over the book's theme, Hofstadter emphasized that the book is not about the relationships of mathematics, art, and music – but rather about how cognition emerges from hidden neurological mechanisms. One point in the book presents an analogy about how individual neurons in the brain coordinate to create a unified sense of a coherent mind by comparing it to the social organization displayed in a colony of ants.

Artists, however, do not necessarily take geometry literally. As Douglas Hofstadter writes in his book:"The difference between an Escher drawing and non-Euclidean geometry is that in the latter, comprehensible interpretations can be found for the undefined terms, resulting in a comprehensible total system, whereas for the former, the end result is not reconcilable with one's conception of the world, no matter how long one stares at the pictures." Hofstadter discusses the seemingly paradoxical lithograph "Print Gallery" by M. C. Escher; it depicts a seaside town containing an art gallery which seems to contain a painting of the seaside town, there being a 'strange loop, or tangled hierarchy' to the levels of reality in the image. The artist himself, Hofstadter observes, is not seen; his reality and his relation to the lithograph are not paradoxical. The image's central void has also attracted the interest of mathematicians Bart de Smit and Hendrik Lenstra, who propose that it could contain a Droste effect copy of itself, rotated and shrunk; this would be a further illustration of recursion beyond that noted by Hofstadter.

Italo Calvino was an Italian writer and journalist. Amongst his best known works the "Cosmicomics" is a collection of twelve short stories by Italo Calvino first published in Italian in 1965 and in English in 1968. The stories were originally published between 1964 and 1965 in the Italian periodicals "Il Caffè" and "Il Giorno". Each story takes a scientific 'fact' (though sometimes a falsehood by today's understanding), and builds an imaginative story around it. An always-extant being called Qfwfq narrates all of the stories save two. Every story is a memory of an event in the history of the universe. Qfwfq also narrates some stories in Calvino's "t zero". All of the stories feature non-human characters which have been heavily anthropomorphized. For example in the story "The Form of Space" as the unnamed narrator 'falla' through space, he cannot help but notice that his trajectory is parallel to that of a beautiful woman, Ursula H'x, and that of lieutenant Fenimore, who is also in love with Ursula. The narrator dreams of the shape of space changing, so that he may touch Ursula (or fight with Fenimore).

9.1 Poetry

The British philosopher, logician and social critic Bertrand Russell expressed his sense of mathematical beauty in these words [91]:

> "Mathematics, rightly viewed, possesses not only truth, but supreme beauty – a beauty cold and austere, like that of sculpture, without appeal to any part of our weaker nature, without the gorgeous trappings of painting or music, yet sublimely pure, and capable of a stern perfection such as only the greatest art can show. The true spirit of delight, the exaltation, the sense of being more than Man, which is the touchstone of the highest excellence, is to be found in mathematics as surely as poetry."

The Fibonacci numbers were first described in Indian mathematics [92], as early as 200 BC in work by Acharya Pingala, an ancient Indian poet and mathematician (the author of the Chandaḥśāstra (also called Pingala-sutras), the earliest known treatise on Sanskrit prosody), on enumerating possible patterns of Sanskrit poetry formed from syllables of two lengths.

Infinity has attracted the imagination of many. “L’infinito” is a poem written by Giacomo Leopardi probably in the autumn of 1819. The poem is a product of Leopardi’s yearning to travel beyond his restrictive home town of Recanati and experience more of the world which he had studied. It is considered one of the most beautiful poem in italian poetry of all times. In the poem the ‘hedge’ excludes the sight from the infinity, because the infinity is something that we can imagine but not verify (see Section 10.1). Our own life for example is finite, has a beginning and an end. The infinite is an abstract subject that often is shifted from mathematics to poetry. Leopardi has various reflections on the infinity in his “Zibaldone” a collection of personal impressions, aphorisms, philosophical observations, philological analyses, literary criticism and various types of notes which was published posthumously in seven volumes in 1898 thanks to a special governmental commission presided over by Giosuè Carducci in occasion of the centennial anniversary of the poet’s birth.

The English schoolmaster, theologian, and Anglican priest Edwin A. Abbott in his satirical novella “Flatland: A Romance of Many Dimensions” [93], written pseudonymously by ‘A Square’, used the fictional two-dimensional world of Flatland to comment on the hierarchy of Victorian culture, but the novella’s more enduring contribution is its examination of dimensions.

Linguistics is the scientific study of human language. It entails a comprehensive, systematic, objective, and precise analysis of all aspects of language, particularly its nature and structure. As linguistics is concerned with both the cognitive and social aspects of language, it is considered a scientific field as well as an academic discipline; it has been classified as a social science, natural science, cognitive science, or part of the humanities.

Traditional areas of linguistic analysis correspond to phenomena found in human linguistic systems, such as syntax (rules governing the structure of sentences); semantics (meaning); morphology (structure of words); phonetics (speech sounds and equivalent gestures in sign languages); phonology (the abstract sound system of a particular language); and pragmatics (how social context contributes to meaning). Subdisciplines such as evolutionary linguistics (the study of the origins and evolution of language) and psycholinguistics (the study of psychological factors in human language) bridge many of these divisions. Computational linguistics is an interdisciplinary field concerned with the computational modelling of natural language, as well as the study of appropriate computational approaches to linguistic questions. In general, computational linguistics draws upon linguistics, computer science, artificial intelligence, mathematics, logic, philosophy, cognitive science, cognitive psychology, psycholinguistics, anthropology and neuroscience, among others.

Linguistics encompasses many branches and subfields that span both theoretical and practical applications. Theoretical linguistics (including traditional descriptive linguistics) is concerned with understanding the fundamental nature of language and developing a general theoretical framework for describing it. Applied linguistics seeks to utilise the scientific findings of the study of language for practical purposes, such as developing methods of improving language education and literacy.

Linguistic phenomena may be studied through a variety of perspectives: synchronically (describing a language at a specific point of time) or diachronically (through historical development); in monolinguals or multilinguals; children or adults; as they are learned or already acquired; as abstract objects or cognitive structures; through texts or oral elicitation; and through mechanical data collection versus fieldwork.

Noam Chomsky is an American linguist, philosopher, cognitive scientist, historical essayist, social critic, and political activist. Sometimes called "the father of modern linguistics".

9.2 Painting and sculpture

Mathematics can provide tools for artists, such as the rules of linear perspective as described by Brook Taylor and Johann Lambert, or the methods of descriptive geometry, now applied in software modelling of solids, dating back to Albrecht Dürer and Gaspard Monge. Artists from Luca Pacioli in the Middle Ages and Leonardo da Vinci and Albrecht Dürer in the Renaissance have made use of and developed mathematical ideas in the pursuit of their artistic work. The use of perspective began, despite some embryonic usages in the architecture of Ancient Greece, with Italian painters such as Giotto in the 13th century; rules such as the vanishing point were first formulated by Brunelleschi in about 1413, his theory influencing Leonardo and Dürer. Isaac Newton's work on the optical spectrum influenced Goethe's Theory of Colours and in turn artists such as Philipp Otto Runge, J. M. W. Turner, the Pre-Raphaelites and Wassily Kandinsky. Artists may also choose to analyse the symmetry of a scene. Tools may be applied by mathematicians who are exploring art, or artists inspired by mathematics, such as M. C. Escher (inspired by H. S. M. Coxeter) and the architect Frank Gehry, who more tenuously argued that computer aided design enabled him to express himself in a wholly new way.

Polykleitos the elder (450–420 BC) was a Greek sculptor from the school of Argos, and a contemporary of Phidias. His works and statues consisted mainly of bronze and were of athletes. According to the philosopher and mathematician Xenocrates, Polykleitos is ranked as one of the most important sculptors of classical antiquity for his work on the "Doryphorus" and the statue of Hera in the "Heraion of Argos". While his sculptures may not be as famous as those of Phidias, they are much admired. In his "Canon", a treatise he wrote designed to document the 'perfect' body proportions of the male nude, Polykleitos gives us a mathematical approach towards sculpturing the human body. The "Canon" itself has been lost but it is conjectured that Polykleitos used a sequence of proportions where each length is that of the diagonal of a square drawn on its predecessor, $1/\sqrt{2}$. The influence of the "Canon" of Polykleitos is immense in Classical Greek, Roman, and Renaissance sculpture, many sculptors following Polykleitos's prescription. While none of Polykleitos's original works survive, Roman copies demonstrate his ideal of physical perfection and mathematical precision. Some scholars argue that Pythagorean thought influenced the "Canon" of Polykleitos. The Canon applies the basic mathematical concepts of Greek geometry, such as the ratio, proportion, and symmetria (Greek for 'harmonious proportions') and turns it into a system capable of describing the human form through a series of continuous geometric progressions.

The rudiments of perspective arrived with Giotto (1266/7 - 1337), who attempted to draw in perspective using an algebraic method to determine the placement of distant lines. In 1415, the Italian architect Filippo Brunelleschi and his friend Leon Battista Alberti demonstrated the geometrical method of applying perspective in Florence, using similar triangles as formulated by Euclid, to find the apparent height of distant objects. Brunelleschi's own perspective paintings are lost, but Masaccio's painting of the Holy Trinity shows his principles at work.

The Italian painter Paolo Uccello (1397–1475) was fascinated by perspective, as shown in his paintings of "The Battle of San Romano" (1435-1460): broken lances lie conveniently along perspective lines.

The painter Piero della Francesca (1415–1492) exemplified this new shift in Italian Renaissance thinking. He was an expert mathematician and geometer, writing books on solid geometry and perspective, including "De prospectiva pingendi" (On Perspective for Painting), "Trattato d'Abaco" (Abacus Treatise), and "De quinque corporibus regularibus" (On the Five Regular

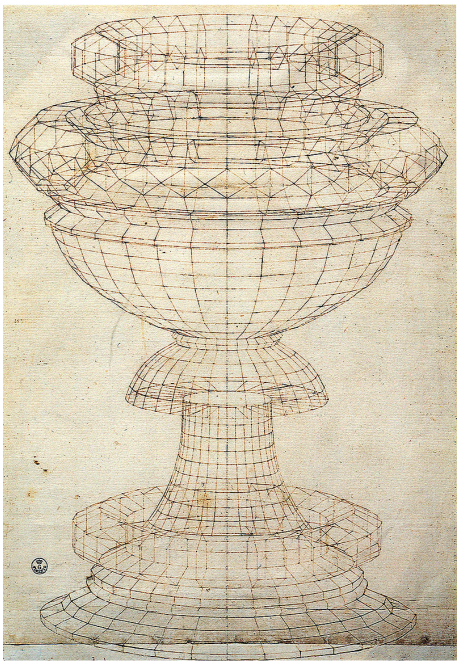


Figure 9.4: Wireframe drawing of a vase as a solid of revolution by Paolo Uccello (15th century).

Solids). The historian Vasari in his Lives of the Painters calls Piero the "greatest geometer of his time, or perhaps of any time." Piero's interest in perspective can be seen in his paintings including the "Polyptych of Perugia", the "San Agostino altarpiece" and "The Flagellation of Christ". His work on geometry influenced later mathematicians and artists including Luca Pacioli in his "De divina proportione" and "Leonardo da Vinci". Piero studied classical mathematics and the works of Archimedes. He was taught commercial arithmetic in 'abacus schools'; his writings are formatted like abacus school textbooks, perhaps including Leonardo Pisano (Fibonacci)'s 1202 "Liber Abaci". Linear perspective was just being introduced into the artistic world. Alberti explained in his 1435 "De pictura": "light rays travel in straight lines from points in the observed scene to the eye, forming a kind of pyramid with the eye as vertex." A painting constructed with linear perspective is a cross-section of that pyramid. In "De Prospectiva Pingendi", Piero transforms his empirical observations of the way aspects of a figure change with point of view into mathematical proofs. His treatise starts in the vein of Euclid: he defines the point as "the tiniest thing that is possible for the eye to comprehend". He uses deductive logic to lead the reader to the perspective representation of a three-dimensional body.

Leonardo da Vinci's illustrations of polyhedra in "Divina proportione" (see Fig. 9.5) have led some to speculate that he incorporated the golden ratio in his paintings. Leonardo's drawings are probably the first illustrations of skeletonic solids. These, such as the rhombicuboctahedron, were among the first to be drawn to demonstrate perspective by being overlaid on top of each other. Da Vinci studied Pacioli's "Summa", from which he copied tables of proportions. In "Mona Lisa" and "The Last Supper", Da Vinci's work incorporated linear perspective with a vanishing point to provide apparent depth. "The Last Supper" is constructed in a tight ratio of 12:6:4:3, as is Raphael's "The School of Athens", which includes Pythagoras with a tablet of ideal ratios, sacred to the Pythagoreans. In "Vitruvian Man", Leonardo expressed the ideas of

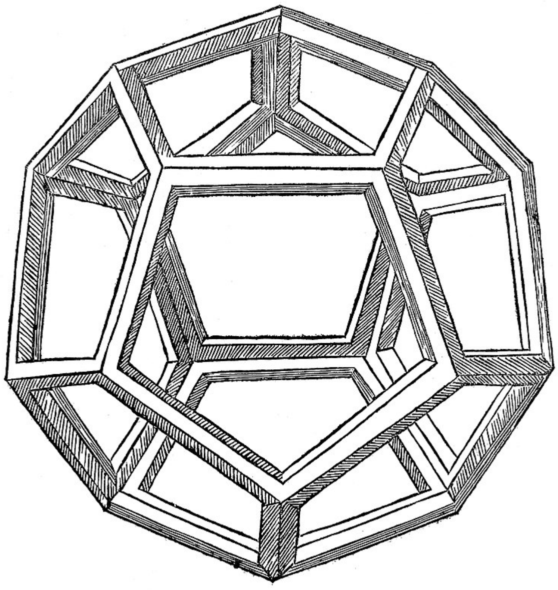


Figure 9.5: Da Vinci's illustration of a dodecahedron from Pacioli's "Divina proportione" (1509).

the Roman architect Vitruvius, innovatively showing the male figure twice, and centring him in both a circle and a square.

As early as the 15th century, curvilinear perspective found its way into paintings by artists interested in image distortions. Jan van Eyck's 1434 "Arnolfini Portrait" contains a convex mirror with reflections of the people in the scene, while Parmigianino's "Self-portrait in a Convex Mirror", 1523–1524, shows the artist's largely undistorted face at the centre, with a strongly curved background and artist's hand around the edge.

Albrecht Dürer (1471–1528) was a German Renaissance printmaker who made important contributions to polyhedral literature in his 1525 book, "Underweysung der Messung" (Education on Measurement), meant to teach the subjects of linear perspective, geometry in architecture, Platonic solids, and regular polygons. Dürer was likely influenced by the works of Luca Pacioli and Piero della Francesca during his trips to Italy. While the examples of perspective in "Underweysung der Messung" are underdeveloped and contain inaccuracies, there is a detailed discussion of polyhedra. Dürer is also the first to introduce in text the idea of polyhedral nets, polyhedra unfolded to lie flat for printing. Dürer published another influential book on human proportions called "Vier Bücher von Menschlicher Proportion" (Four Books on Human Proportion) in 1528. Dürer's well-known engraving "Melencolia I" depicts a frustrated thinker sitting by a truncated triangular trapezohedron and a magic square. These two objects, and the engraving as a whole, have been the subject of more modern interpretation than the contents of almost any other print, including a two-volume book by Peter-Klaus Schuster, and an influential discussion in Erwin Panofsky's monograph of Dürer.

Salvador Dalí, influenced by the works of Matila Ghyka, explicitly used the golden ratio in his masterpiece, The Sacrament of the Last Supper (see Fig. 9.6). The dimensions of the canvas are a golden rectangle. A huge dodecahedron, in perspective so that edges appear in golden ratio to one another, is suspended above and behind Jesus and dominates the composition. Salvador

Figure 9.6: "The Sacrament of the Last Supper" by Salvador Dalí (1955, Oil on canvas, 267 cm × 166.7 cm National Gallery of Art, Washington DC).

Dalí's last painting, "The Swallow's Tail" (1983), was part of a series inspired by René Thom's catastrophe theory. Salvador Dalí's 1954 painting "Corpus Hypercubus" uniquely depicts the cross of Christ as an unfolded three-dimensional net for a hypercube, also known as a tesseract: the unfolding of a tesseract into these eight cubes is analogous to unfolding the sides of a cube into a cross shape of six squares, here representing the divine perspective with a four-dimensional regular polyhedron. The painting shows the figure of Christ in front of the tessaract; he would normally be shown fixed with nails to the cross, but there are no nails in the painting. Instead, there are four small cubes in front of his body, at the corners of the frontmost of the eight tessaract cubes. The mathematician Thomas Banchoff states that Dalí was trying to go beyond the three-dimensional world, while the poet and art critic Kelly Grovier says that "The painting seems to have cracked the link between the spirituality of Christ's salvation and the materiality of geometric and physical forces. It appears to bridge the divide that many feel separates science from religion."

Maurits Cornelis Escher was a Dutch graphic artist who made mathematically inspired woodcuts, lithographs, and mezzotints (see Fig. 9.7). Despite wide popular interest, Escher was for most of his life neglected in the art world, even in his native Netherlands. He was 70 before a retrospective exhibition was held. In the late twentieth century, he became more widely appreciated, and in the twenty-first century he has been celebrated in exhibitions around the world. His work features mathematical objects and operations including impossible objects, explorations of infinity, reflection, symmetry, perspective, truncated and stellated polyhedra, hyperbolic geometry, and tessellations. Although Escher believed he had no mathematical ability, he interacted with the mathematicians George Pólya, Roger Penrose, Harold Coxeter and crystallographer Friedrich Haag, and conducted his own research into tessellation. Early in his career, he drew inspiration from nature, making studies of insects, landscapes, and plants such as lichens, all of which he used as details in his artworks. He traveled in Italy and Spain, sketching buildings, townscapes, architecture and the tilings of the Alhambra and the Mezquita of Cordoba, and became steadily more interested in their mathematical structure. The mathematics of tessellation, polyhedra, shaping of space, and self-reference provided Escher with a lifetime's worth of materials for his woodcuts. In the "Alhambra Sketch", Escher showed that art can be created with

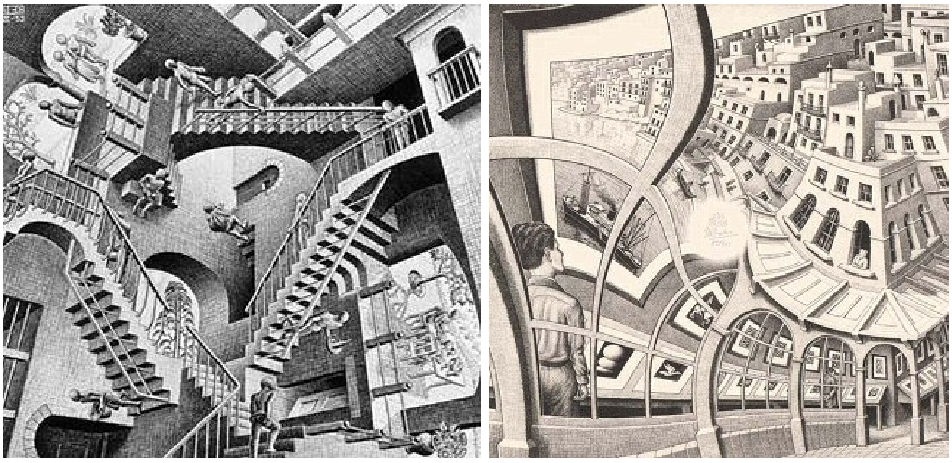


Figure 9.7: Left: "Relativity" is a lithograph print (27.7 cm × 29.2 cm) by the Dutch artist M. C. Escher, first printed in December 1953. The first version of this work was a woodcut made earlier that same year. It depicts a world in which the normal laws of gravity do not apply. In the world of Relativity, there are three sources of gravity, each being orthogonal to the two others. Each inhabitant lives in one of the gravity wells, where normal physical laws apply. Right: Lithograph "Print Gallery" by M. C. Escher, 1956.

polygons or regular shapes such as triangles, squares, and hexagons. Escher used irregular polygons when tiling the plane and often used reflections, glide reflections, and translations to obtain further patterns. Many of his works contain impossible constructions, made using geometrical objects which set up a contradiction between perspective projection and three dimensions, but are pleasant to the human sight. Escher's "Ascending and Descending" is based on the "impossible staircase" created by the medical scientist Lionel Penrose and his son the mathematician Roger Penrose. Some of Escher's many tessellation drawings were inspired by conversations with the mathematician H. S. M. Coxeter on hyperbolic geometry. Escher was especially interested in five specific polyhedra, which appear many times in his work. The Platonic solids – tetrahedrons, cubes, octahedrons, dodecahedrons, and icosahedrons – are especially prominent in "Order and Chaos" and "Four Regular Solids". These stellated figures often reside within another figure which further distorts the viewing angle and conformation of the polyhedrons and provides a multifaceted perspective artwork. The visual intricacy of mathematical structures such as tessellations and polyhedra have inspired a variety of mathematical artworks. Stewart Coffin makes polyhedral puzzles in rare and beautiful woods; George W. Hart works on the theory of polyhedra and sculpts objects inspired by them; Magnus Wenninger makes 'especially beautiful' models of complex stellated polyhedra. The distorted perspectives of anamorphosis have been explored in art since the sixteenth century, when Hans Holbein the Younger incorporated a severely distorted skull in his 1533 painting "The Ambassadors". Many artists since then, including Escher, have make use of anamorphic tricks.

The artists Theo van Doesburg and Piet Mondrian founded the De Stijl movement, which they wanted to "establish a visual vocabulary comprised of elementary geometrical forms comprehensible by all and adaptable to any discipline". Many of their artworks visibly consist of ruled squares and triangles, sometimes also with circles. De Stijl artists worked in painting, furniture, interior design and architecture. After the breakup of De Stijl, Van Doesburg founded the Avant-garde Art Concret movement, describing his 1929-1930 "Arithmetic Composition" (see

Fig. 9.8), a series of four black squares on the diagonal of a squared background, as "a structure that can be controlled, a definite surface without chance elements or individual caprice", yet "not lacking in spirit, not lacking the universal and not [...] empty as there is everything which fits the internal rhythm". The art critic Gladys Fabre observes that two progressions are at work in the painting, namely the growing black squares and the alternating backgrounds.

Figure 9.8: From left to right, top to bottom: Theo van Doesburg's "Six Moments in the Development of Plane to Space", 1926 or 1929; "Mathematical sculpture" by Bathsheba Grossman, 2007; "3D Fraktal 03/H/dd" by Hartmut Skerbisch, 2003; Detail of Fibonacci world in artwork by Samuel Monnier, 2009; Computer art image produced by Desmond Paul Henry's "Drawing Machine 1", 1962.

Giotto's "Stefaneschi Triptych", 1320, illustrates recursion in the form of 'mise en abyme'; the central panel of the triptych contains, lower left, the kneeling figure of Cardinal Stefaneschi, holding up the triptych as an offering.

Giorgio de Chirico's metaphysical paintings such as his 1917 "Great Metaphysical Interior" explore the question of levels of representation in art by depicting paintings within his paintings.

Art can exemplify logical paradoxes, as in some paintings by the surrealist René Magritte, which can be read as semiotic jokes about confusion between levels. In "La condition humaine" (1933) (see Fig. 9.9), Magritte depicts an easel (on the real canvas), seamlessly supporting a view through a window which is framed by 'real' curtains in the painting. Similarly, Escher's "Print Gallery" (1956) is a print which depicts a distorted city which contains a gallery which recursively contains the picture, and so ad infinitum (see Fig. 9.7). Magritte made use of spheres and cuboids to distort reality in a different way, painting them alongside an assortment of houses in his 1931 “Mental Arithmetic” as if they were children’s building blocks, but house-sized. The Guardian observed that the ‘eerie toytown image’ prophesied Modernism’s usurpation of ‘cosy traditional forms’, but also plays with the human tendency to seek patterns in nature.

Figure 9.9: René Magritte’s “La condition humaine” 1933.

The Spanish painter and sculptor Pablo Palazuelo focused on the investigation of form. He developed a style that he described as the geometry of life and the geometry of all nature. Consisting of simple geometric shapes with detailed patterning and coloring, in works such as “Angular I” and “Automnes”, Palazuelo expressed himself in geometric transformations.

The mathematician and theoretical physicist Henri Poincaré’s “Science and Hypothesis” was widely read by the Cubists, including Pablo Picasso (see Fig. 9.10) and Jean Metzinger. Being thoroughly familiar with Bernhard Riemann’s work on non-Euclidean geometry, Poincaré was more than aware that Euclidean geometry is just one of many possible geometric configurations, rather than as an absolute objective truth. The possible existence of a fourth dimension inspired artists to question classical Renaissance perspective: non-Euclidean geometry became a valid alternative. The concept that painting could be expressed mathematically, in colour and form, contributed to Cubism, the art movement that led to abstract art. Metzinger, in 1910, wrote that: “[Picasso] lays out a free, mobile perspective, from which that ingenious mathematician Maurice Princet has deduced a whole geometry”.

The impulse to make teaching or research models of mathematical forms naturally creates objects that have symmetries and surprising or pleasing shapes. Some of these have inspired artists such as the Dadaists Man Ray, Marcel Duchamp and Max Ernst, and following Man Ray, Hiroshi Sugimoto. Man Ray photographed some of the mathematical models in the Institut Henri Poincaré in Paris, including “Objet mathematique” (Mathematical object) (see Fig. 9.11). He noted that this represented Enneper surfaces with constant negative curvature, derived from the pseudo-sphere [94]. This mathematical foundation was important to him, as it allowed him to deny that the object was ‘abstract’, instead claiming that it was as real as the urinal that

Figure 9.10: Pablo Picasso's 1907 painting "Les Demoiselles d'Avignon" uses a fourth dimension projection to show a figure both full face and in profile.

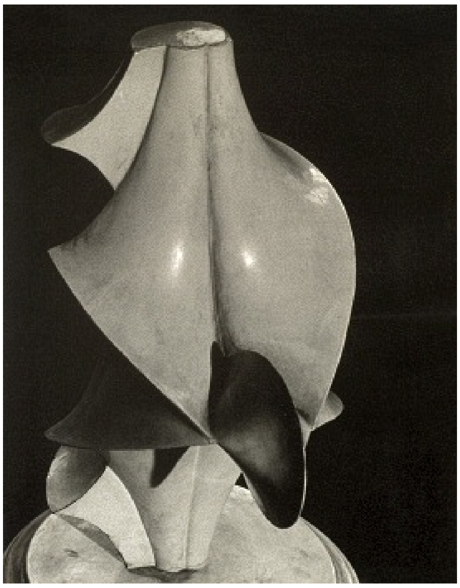


Figure 9.11: Man Ray's 1934 "Objet mathematique".

Duchamp made into a work of art. Man Ray admitted that the object's formula "meant nothing to me, but the forms themselves were as varied and authentic as any in nature." He used his photographs of the mathematical models as figures in his series he did on Shakespeare's plays, such as his 1934 painting "Antony and Cleopatra". The art reporter Jonathan Keats, writing in ForbesLife, argues that Man Ray photographed "the elliptic paraboloids and conic points in the same sensual light as his pictures of Kiki de Montparnasse", and "ingeniously repurposes the cool calculations of mathematics to reveal the topology of desire".

The mathematics of topology has inspired several artists in modern times. The sculptor John Robinson created works such as Gordian Knot and Bands of Friendship, displaying knot theory in polished bronze. Other works by Robinson explore the topology of toruses. "Genesis" is based on Borromean rings – a set of three circles, no two of which link but in which the whole structure cannot be taken apart without breaking. The sculptor Helaman Ferguson creates complex surfaces and other topological objects. His works are visual representations of mathematical objects; The "Eightfold Way" is based on the projective special linear group PSL(2,7), a finite group of 168 elements. The sculptor Bathsheba Grossman similarly bases her work on mathematical structures. The artist Nelson Saiers incorporates mathematical concepts and theorems in his art from toposes and schemes to the four color theorem and the irrationality of π.

The drip painting works of the modern artist Jackson Pollock are distinctive in their fractal dimension. His 1948 "Number 14" has a coastline- like dimension of 1.45, while his later paintings had successively higher fractal dimensions and accordingly more elaborate patterns. One of his last works, "Blue Poles", took six months to create, and has the fractal dimension of 1.72.

The artist Richard Wright argues that mathematical objects that can be constructed can be seen either 'as processes to simulate phenomena' or as works of 'computer art'. He considers the nature of mathematical thought, observing that fractals were known to mathematicians for a century before they were recognised as such. Wright concludes by stating that it is appropriate to subject mathematical objects to any methods used to "come to terms with cultural artifacts like art, the tension between objectivity and subjectivity, their metaphorical meanings and the character of representational systems." He gives as instances an image from the Mandelbrot set, an image generated by a cellular automaton algorithm, and a computer-rendered image, and discusses, with reference to the Turing test, whether algorithmic products can be art. Sasho Kalajdzievski's "Math and Art: An Introduction to Visual Mathematics" takes a similar approach, looking at suitably visual mathematics topics such as tilings, fractals and hyperbolic geometry.

Some of the first works of computer art were created by Desmond Paul Henry's "Drawing Machine 1", an analogue machine based on a bombsight computer and exhibited in 1962. The machine was capable of creating complex, abstract, asymmetrical, curvilinear, but repetitive line drawings. More recently, Hamid Naderi Yeganeh has created shapes suggestive of real world objects such as fish and birds, using formulae that are successively varied to draw families of curves or angled lines. Artists such as Mikael Hvidtfeldt Christensen create works of generative or algorithmic art by writing scripts for a software system such as "Structure Synth": the artist effectively directs the system to apply a desired combination of mathematical operations to a chosen set of data.

9.3 Music

Examples of the use of mathematics in music include the stochastic music of Iannis Xenakis, the Fibonacci sequence in Tool's "Lateralus", counterpoint of Johann Sebastian Bach, polyrhythmic structures (as in Igor Stravinsky's "The Rite of Spring"), the Metric modulation of Elliott Carter, permutation theory in serialism beginning with Arnold Schoenberg, and application of Shepard tones in Karlheinz Stockhausen's "Hymnen". They also include the application of Group theory to transformations in music in the theoretical writings of David Lewin.

Ernõ Lendvai analyzes Béla Bartók's works as being based on two opposing systems, that of the golden ratio and the acoustic scale [95].

French composer Erik Satie used the golden ratio in several of his pieces, including Sonneries de la Rose+Croix.

The golden ratio is also apparent in the organization of the sections in the music of Debussy's Reflets dans l'eau (Reflections in Water), from Images (1st series, 1905), in which "the sequence of keys is marked out by the intervals 34, 21, 13 and 8, and the main climax sits at the phi position" [96]. The musicologist Roy Howat has observed that the formal boundaries of Debussy's La Mer correspond exactly to the golden section [97].

Though Heinz Bohlen proposed the non-octave-repeating 833 cents scale based on combination tones, the tuning features relations based on the golden ratio. As a musical interval the ratio 1.618... is 833.090... cents

Johann Sebastian Bach enriched established German styles through his mastery of counterpoint, harmonic and motivic organisation. Many of his works employ the genres of canon and fugue. "The Art of Fugue", BWV 1080, is an incomplete musical work of unspecified instrumentation by Johann Sebastian Bach. Written in the last decade of his life, The Art of Fugue is the culmination of Bach's experimentation with monothematic instrumental works. This work consists of 14 fugues and four canons in D minor, each using some variation of a single principal subject, and generally ordered to increase in complexity. "Fuga a 3 Soggetti" ("fugue in three subjects"), also referred to as the "Unfinished Fugue", was contained in a handwritten manuscript bundled with the autograph manuscript Mus. ms. autogr. P200. It breaks off abruptly in the middle of its third section, with an only partially written measure 239. This autograph carries a note in the handwriting of Carl Philipp Emanuel Bach, stating "Über dieser Fuge, wo der Name B A C H im Contrasubject angebracht worden, ist der Verfasser gestorben." ("While working on this fugue, which introduces the name BACH in the countersubject, the composer died.") This account is disputed by modern scholars, as the manuscript is clearly written in Bach's own hand, and thus dates to a time before his deteriorating health and vision would have prevented his ability to write, probably 1748–1749. The "Unfinished Fugue" is the 14th one as 14 is the number corresponding to the sum B+A+C+H. Douglas Hofstadter's book "Gödel, Escher, Bach" discusses the unfinished fugue and Bach's supposed death during composition as a tongue-in-cheek illustration of Austrian logician Kurt Gödel's first incompleteness theorem. According to Gödel, the very power of a 'sufficiently powerful' formal mathematical system can be exploited to 'undermine' the system, by leading to statements that assert such things as "I cannot be proven in this system". In Hofstadter's discussion, Bach's great compositional talent is used as a metaphor for a 'sufficiently powerful' formal system; however, Bach's insertion of his own name 'in code' into the fugue is not, even metaphorically, a case of Gödelian self-reference; and Bach's failure to finish his self-referential fugue serves as a metaphor for the unprovability of the Gödelian assertion, and thus for the incompleteness of the formal system.

Sylvestre and Costa [98] reported a mathematical architecture of "The Art of Fugue", based on bar counts, which shows that the whole work was conceived on the basis of the Fibonacci series and the golden ratio.

Musicology is the scholarly analysis and research-based study of music. Musicology departments traditionally belong to the humanities, although some music research is scientific in focus (psychological, sociological, acoustical, neurological, computational). Some geographers and anthropologists have an interest in musicology so the social sciences also have an academic interest. A scholar who participates in musical research is a musicologist.

Musicology traditionally is divided in three main branches: historical musicology, systematic musicology and ethnomusicology. Historical musicologists mostly study the history of the western classical music tradition, though the study of music history need not be limited to that. Ethnomusicologists draw from anthropology to understand how and why people make music. Systematic musicology includes music theory, aesthetics, pedagogy, musical acoustics, the science and technology of musical instruments, and the musical implications of physiology, psychology, sociology, philosophy and computing. Cognitive musicology is the set of phenomena surrounding the cognitive modeling of music. When musicologists carry out research using computers, their research often falls under the field of computational musicology.

9.4 Logic

We report here a discussion given by Elio Fabri from the Department of Physics of the University of Pisa about the *Gödel's incompleteness theorems*. These are two theorems of mathematical logic that are concerned with the limits of provability in formal axiomatic theories. These results, published by Kurt Gödel in 1931, are important both in mathematical logic and in the philosophy of mathematics. The theorems are widely, but not universally, interpreted as showing that Hilbert's program to find a complete and consistent set of axioms for all mathematics is impossible. The first incompleteness theorem states that no consistent system of axioms whose theorems can be listed by an effective procedure (i.e., an algorithm) is capable of proving all truths about the arithmetic of natural numbers. For any such consistent formal system, there will always be statements about natural numbers that are true, but that are unprovable within the system. The second incompleteness theorem, an extension of the first, shows that the system cannot demonstrate its own consistency.

“The story begins in the first years of the 20th century, when following the research of Cantor, Russel, etc. becomes relevant the discussion on the possibility to “rigorously prove” something in mathematics, or also to state precisely what is a mathematical proof. Or to prove that an axiomatic system is not contradictory.

Hilbert proposes this route: rigorously formalize the mathematical proofs. The technique to be followed is very familiar today, thanks to the developments of information technology, which at the time was still far to come. First of all, it is a question of writing all the mathematical propositions in a precise language, completely unambiguous, as “strings” of symbols of a certain alphabet. In this spirit, propositions are not to be seen as endowed with meaning (interpreted), but only for their syntactic structure, that is, for the relations between symbols. Again, the computer analogy helps: a program written in a language like BASIC, C, PASCAL has meaning for us; but the compiler that has to translate it into machine language takes care first of all of the syntactic correctness (are there the “;” in the right place? are the opening brackets also closed? have the variables been declared? etc.).

Once we have reduced the propositions to strings, we must introduce the “axioms”, which are only starting propositions, which we do not want to prove. Then we must give the “rules of deduction”, that is the procedures by which it is permissible to “deduce” one proposition from another. For example: if the proposition “$a = a$” is an axiom (or if it has already been proved) it is permissible to replace the variable “a” with any correct expression: I can therefore take as proved equation “$b + c/7 = b + c/7$”. There are (we need) more “clever” deduction rules than this substitution rule, but it is not necessary here to specify what they may be. A finite sequence of applications of deduction rules starting from the axioms ends in a final proposition: this proposition is a “theorem”, and the sequence is called the “proof” of the theorem.

Once this is done (a theory has been formalized, which can be Euclidean geometry, elementary arithmetic, or other) we can ask ourselves some questions, of which the central ones are: 1) is the theory consistent? 2) is the theory complete? Consistent is the same as "not contradictory": it means that it is not possible for a proposition and its negation to be theorems (therefore demonstrable). Complete means that it is possible to prove all "true" propositions. This second assertion must be explained, because the term "true" belongs to a different level of discourse.

The fact is that the theory that we have formalized usually has an interpretation (geometry speaks of lines, angles ...) and in this interpretation certain propositions may be true or false, regardless of whether we have succeeded in finding a proof of them. [1]

The two central questions I mentioned above have another common feature: they are meta-mathematical questions. I am not asking myself if "the sum of the squares of the catheti ..." that is, if a certain mathematical proposition is true; I am asking if a certain property of a mathematical theory holds. That is, I look at mathematics "from the outside", "from above". Hilbert's formalistic objective can be summarized as follows: formalize not only mathematics, but also metamathematics, that is, to give formal demonstrations of the answers to those two questions (and other similar ones).

And here comes Gödel, to break the eggs in the basket ... I can only report in "popular" form the result of his work, which appears in 1931, with the title Über formal unentscheidbare Sätze der "Principia mathematica" und verwandter Systeme (on the formally undecidable propositions of the "Principia mathematica" and related systems). I draw attention to two words: "formally undecidable"; then we'll see why. The "Principia mathematica" is a treatise on mathematical logic by Whitehead and Russell: a cornerstone of the research I mentioned at the beginning, but we don't need to deal with it.

Basically it is about this: with a laborious as well as ingenious work, Gödel manages to construct a proposition of the formal system of arithmetic, which he will indicate with G, whose interpretation is simply "**G is not an arithmetic theorem**". But be careful not to draw hasty conclusions: this is the interpretation, which on the other hand could also be false; but how are we on the formal level? Let's consider the two possibilities: a) G is a theorem, that is, it has a formal proof. Then we would have proved a theorem which says that this proof cannot exist: contradiction. b) G is not a theorem, i.e. the proof does not exist. Then the interpretation of G is true, and we have a true proposition that cannot be proved. So either the arithmetic is contradictory, or it is incomplete. There is no escape from one of the two horns of the dilemma. More: if the arithmetic is not contradictory, not only is G not a theorem, but neither is its negation, which is a false proposition: this is expressed by saying that G is undecidable. (A proposition P is generally called undecidable when neither P nor its negation is demonstrable.) Note that incompleteness is not a problem, as it is not for geometry. If you remove the postulate of parallels from Euclidean geometry, we know that this cannot be proved starting from the others. We then have two ways: a) add the new postulate (and we have Euclidean geometry) b) add a different postulate, which negates that (and we have non-Euclidean geometry).

I stop here on the exposition of Gödel's theorem.

Just a comment: very often Gödel's theorem is quoted inappropriately, forgetting the key word of the title. Do you remember? "formally undecidable". Everything revolves around that formally, which we now know what it means. Gödel has shown us that there are formally undecidable propositions (and not only in arithmetic, but in all formal systems useful for mathematics). But he didn't stop us from giving non-formal demonstrations, as we always do. The price you

[1] I know well that a big problem is touched upon here, which leads us not only to semantics, as soon as we speak of interpretation, but also to philosophy. For example: does it make sense to speak of a truth different from that which can be expressed with formal demonstrations? I am satisfied with this observation: anyone who does mathematics thinks mostly in a non-formal way, and is convinced of the validity of his conclusions. So there is a truth, at least in practice, that is not reduced to the formal one.

pay is that we cannot be certain of the rigor that a formal demonstration guarantees. Or in other words: in a non-formal proof we are probably using richer deduction rules than we know how to formalize.

Therefore it is not fair to say that Goldbach's conjecture (or other propositions: it was also assumed of Fermat's theorem, now proved) can be undecidable, referring to Gödel's theorem. In fact, for almost all the theorems of mathematics that we know, no one has ever even tried to write formal proofs, so the possible non-existence of such proofs does not change anything. And in fact, after Gödel, mathematicians did not change jobs, they continued to collect salaries (including Gödel) and so on. Nor does the difficulty of finding a proof (or proof of negation) of a conjecture say anything about its formal undecidability; and "non-formal" undecidability does not even make sense to speak.

Am I saying that Gödel's theorem is useless? Not at all. First, it clarified the logical status of mathematical theories and proofs: what can be and what cannot be expected / achieved. Secondly – this is the mischievous point of view of a physicist, I admit – he should have taught mathematicians a little modesty: they can no longer look down on other scientists, believing themselves to be alone and true. custodians of an unassailable rigor. On the other hand, I reject the interpretations (dear to some philosophers) regarding the alleged proof of the limits of human reason, the unknowability of the truth, and similar nonsense. These interpretations are simply forcing, deriving from a combination of ignorance and bad faith: ignorance, because none of those who speak like this have ever tried to understand Gödel's theorem seriously (and they wouldn't be able to!); bad faith, because they only try to turn something that has another meaning to the advantage of their own thesis (always unscientific). ”

In his book "La ribellione del numero" [99] Paolo Zellini narrates that "We are of divine race and we possess the power to create" wrote a great mathematician, Richard Dedekind, in a letter of 1888. That phrase corresponds to the climate of general intoxication and euphoria that reigned in mathematics at the time. With the non-Euclidean geometries of Lobacevskij and Riemann, with the transfinite numbers of Cantor it seemed that the doors of a boundless 'paradise' had opened, teeming with unprecedented 'mental entities', which existed side by side, obeying the only condition not to be contradictory. Then, suddenly, within a few years, between 1897 and 1901, the first 'paradoxes' began to emerge, signaling as many dead ends in set theory and in Russell's new logical-mathematical construction. It was the first sign of a devastating 'rebellion of numbers': as if the formula revealed that it had a nature of its own, perhaps incompatible with that of the mind that had just made it explicit. Mathematicians were immediately tempted to shake off such troublesome difficulties as irrelevant. Indeed, precisely in the first decades of the century we witness the development of the most ambitious challenge ever supported by mathematics: Hilbert's project of total axiomatization. But soon that grandiose feat also showed its cracks. Finally, the late and definitive revenge of paradoxes came in 1931 with Gödel's theorem, which demonstrated the insuperability of those paradoxes. Since then it can be said that due to the 'crisis of the foundations', what happened to so many other discoveries of the Modern: what had presented itself as a dramatic and distressing novelty has become part of normal life. The quicksand that one day paralyzed with fear seems to have changed into a public park, where shrewd gardeners have drawn paths that allow you to avoid the points where you immediately sink.

Chapter 10

God

The human being can encounter something that is rationally known or something that is unknown. In the latter case he usually behaves irrationally before than rationally eventually guided by fear or faith. It is like when visiting for the first time an unknown house we are faced in front of opening a closed door. Before any rational thinking our imagination starts producing all kinds of hypothesis based on our previous experiences. In the same way we can say that mathematics is a manifestation of the rational thinking and religion is a manifestation of the irrational behavior. Therefore mathematics is universal, the same for any human being on the planet, whereas religion can manifest in several different representations.'God' is the name we conventionally give to the unknown or to something that we are not able to explain. The first thing that a young boy asks to the father trying to explain how life came in the universe is: "but father then who created the universe". Then the father may as well answer "we don't know yet" or simply "God" depending whether he wants to stimulate the rational thinking of the son or the irrational one. The "Divine Comedy" written by Dante Alighieri [9] allegorically represents the soul's journey towards God, beginning with the recognition and rejection of sin (Inferno), followed by the penitent Christian life (Purgatorio), which is then followed by the soul's ascent to God (Paradiso).

We give here some examples in which God is lumped to some mathematical concepts:

A strand of art from Ancient Greece onwards sees God as the geometer of the world, and the world's geometry therefore as sacred (see Fig. 10.1). The belief that God created the universe according to a geometric plan has ancient origins. Plutarch attributed the belief to Plato, writing that 'Plato said God geometrizes continually' ("Convivialium disputationum", liber 8,2). This image has influenced Western thought ever since. The Platonic concept derived in its turn from a Pythagorean notion of harmony in music, where the notes were spaced in perfect proportions, corresponding to the lengths of the lyre's strings; indeed, the Pythagoreans held that everything was arranged by Number. In the same way, in Platonic thought, the regular or Platonic solids dictate the proportions found in nature, and in art. An illumination in the 13th-century "Codex Vindobonensis" shows God drawing out the universe with a pair of compasses, which may refer to a verse in the Old Testament: 'When he established the heavens I was there: when he set a compass upon the face of the deep' (Proverbs 8:27). In 1596, the mathematical astronomer Johannes Kepler modelled the universe as a set of nested Platonic solids, determining the relative sizes of the orbits of the planets. William Blake's "Ancient of Days" (depicting Urizen, Blake's embodiment of reason and law) and his painting of the physicist Isaac Newton, naked, hunched and drawing with a compass, use the symbolism of compasses to critique conventional reason and materialism as narrow-minded. Salvador Dalí's 1954 "Crucifixion" (Corpus Hypercubus) depicts the cross as a hypercube, representing the divine perspective with four dimensions rather than the usual three. In Dalí's "The Sacrament of the Last Supper" (1955) Christ and his disciples are pictured inside a giant dodecahedron (see Section 9.2).

Figure 10.1: From left to right: God the geometer. "Codex Vindobonensis", c. 1220; Johannes Kepler's Platonic solid model of planetary spacing in the Solar System from "Mysterium Cosmographicum", 1596; William Blake's "The Ancient of Days", 1794.

"Divina proportione" (Divine proportion), a three-volume work by Luca Pacioli, was published in 1509 [89]. Pacioli, a Franciscan friar, was known mostly as a mathematician, but he was also trained and keenly interested in art. Divina proportione explored the mathematics of the golden ratio (see Chapter 9). Though it is often said that Pacioli advocated the golden ratio's application to yield pleasing, harmonious proportions, Livio points out that the interpretation has been traced to an error in 1799, and that Pacioli actually advocated the Vitruvian system of rational proportions [100]. Pacioli also saw Catholic religious significance in the ratio, which led to his work's title.

"God Created the Integers: The Mathematical Breakthroughs That Changed History" [101] is a 2005 anthology, edited by Stephen Hawking, of "excerpts from thirty-one of the most important works in the history of mathematics." The title of the book is a reference to a quotation attributed to mathematician Leopold Kronecker, who once wrote that "God made the integers; all else is the work of man."

In the mainstream media, the Higgs boson (see Chapter 5) has often been called the 'God particle' from the 1993 book "The God Particle" by Nobel Laureate Leon Lederman [102], although the nickname is not endorsed by many physicists.

In the physics of soft matter when studying functionalized colloidal solutions [103] a *Janus fluid* is one made of Janus particles immersed in a solvent. A *Janus particle* like the Roman God Janus, depicted in Fig. 10.2, is one that has two faces with two different functionalities. Originally, the term Janus particle was coined by C. Casagrande *et al.* in 1988 [104] to describe spherical glass particles with one of the hemispheres hydrophilic and the other hydrophobic. The Janus particles are commonly found amongst the soft matter colloidal particles [105]. Today an unprecedented development in particle synthesis is generating a whole new set of colloidal particles, characterized by different sizes, shapes, patterns, particle patchiness, and functionalities

Figure 10.2: A statue representing Janus Bifrons in the Vatican Museums.

[106]. The concept of "Janus particles" was first raised by P.-G. de Gennes 20 years ago in his Nobel Prize lecture [107].

In his book "La matematica degli dèi e gli algoritmi degli uomini" [108] Paolo Zellini, an Italian mathematician, asks if numbers are an invention of the mind or a discovery by which the mind ascertains the existence of something that is in the world. Question that mathematicians have tried to answer for centuries and which can also be formulated as follows: what kind of reality should be attributed to numbers? Another fundamental question arises: how can it happen that something, while growing in size (and nothing grows like numbers), remains the same? Question similar to that on the identity of things subject to metamorphosis. And comparable to those that physicists place on the constitution of matter.

10.1 The infinite

There are two notions of infinite, two 'twins': the *potential infinite*, the good twin, that is the ability of being able to think in a logically correct way about objects that can be infinitely big as for example the set of the natural numbers, and the *actual infinite*, the obscure twin, already conceived by the Greeks, that is the attitude of considering the infinite as something that is given and can be kept all under our sight. It had already been recognized at the time of the Greeks that one encounters some paradoxes when dealing with this second notion of infinite and these had been consolidated in the more modern times. Galileo for example noted that there is a one-to-one correspondence (a bijective correspondence) between the natural numbers and the even numbers that albeit still infinite seems too be only half of the whole naturals since there are also the odd numbers. So this is in conflict with our intuitions coming from the 'finite world'. So in the 'infinite world' it can happen that one set can be 'as big as' one of its subsets. The first reaction when we faced these paradoxes was just to ignore them and just admit that the actual infinite cannot be treated being essentially illogical. For example the great mathematician Gauss use to say that the infinite is just a saying and we cannot do any theory out of it. But at the end of 1800 the genial mathematician Georg Cantor showed that one can nonetheless give a rich and sensible theory of the infinite. The idea of Cantor was to 'measure' the infinite. In mathematics the number of elements, the cardinality, of the family of subsets (the power set) of a finite set of n elements is 2^n. Therefore the natural numbers can be identified with the family of subsets of the finite sets. But we can also consider the family of subsets of an infinite set: for example, the family of subsets of the set of natural numbers can be put in a one-to-one correspondence with the set of real numbers. The family of subsets is of fundamental importance in the theory of infinite sets. In fact, in the transfinite arithmetic defined by Georg Cantor, the 'exponentiation' operation, in the sense of identifying the cardinality (a measure of the 'number of elements' of a given set) of the family of subsets of a given infinite set, is the only way to advance in the succession of transfinite cardinal numbers, and therefore obtaining an infinite of an higher order. In the above example, we pass from the cardinality of the discrete (or cardinality of the numerable or countable), that is, of the sets for which it is possible to establish a one-to-one correspondence with the naturals, such as integers, rationals (discovered by Pythagoras) and each of their Cartesian products, to the cardinality of the continuum proper to reals. The next cardinality is the one of the functions (real functions of one real variable) that is obtained by considering the power set of the reals.

In his book "Breve storia dell'infinito" [109] the mathematician Paolo Zellini has told about the technical evolution of the mathematical notion of infinity and at the same time has wrapped it up in those rich mythical, theological and literary speculations that have always accompanied it. He tells how the first Western word to designate infinity is 'apeiron', the 'limitless', as it already appears in Anaximander. But the Greek infinity, from the Presocratics to the Aristotelian arrangement, precisely because it was considered a 'divine, immortal and indestructible' principle, is handled with extreme caution in the procedures of discursive thought. And it will always be a question, then, of an *infinite potential*, conceived in the sign of 'negation' and 'privation' (Aristotle's stéresis). The contest between the finite and the infinite thus appeared as one of the forms of the ultimate contest of all: that between the One and the Many. The number, synonymous with measure and harmony, served as a mysterious point of mediation between the limit and the unlimited. From ancient Greece to today, the sequence of the metamorphoses of infinity will be dizzying. The development of mathematics is intertwined with radical changes in the way of conceiving the cosmic and mental reality of infinity. Gradually we will see the emergence of what is the great attraction and temptation of Western thought: the *actual infinity*, which the Greeks had avoided and is now taking on an increasingly central role. In the last, burning stretch of this story, which goes from Leibniz to Bolzano and Cantor, we will see ever renewed attempts to "explicitly indicate the infinite with 'something' ", until this 'something' proves to be "susceptible to be manipulated as a tangible sign of algebraic mechanics". A subjugating cosmic reality thus turns into a thin mark on paper. But once we reach, with the Cantorian theory of the transfinite, the flowering of an unprecedented kind of mathematics, the insidious holes of paradoxes and antinomies will immediately begin to open, which will undermine the very foundations of science. From this crisis, in which we are still immersed, will derive the most relevant epistemological discoveries of our time.

Two questions regarding contemporary theological and philosophical studies are often overlooked: "Is God infinite or finite?" and, "What does it mean to say that God is infinite?" In "The Infinity of God" [110], Benedikt Paul Göcke and Christian Tapp bring together prominent scholars to discuss God's infinitude from philosophical and theological perspectives. Each contributor deals with a particular aspect of the infinity of God, employing the methods of analytic theology and analytic philosophy. The essays in the first section examine historical issues from a systematic point of view. The contributors focus on the Cappadocian Fathers, Thomas Aquinas, Leibniz, Kant, Hegel, Bolzano, and Cantor. The second section deals with particular issues concerning the relation between God's infinity and both the finitude of the world and the classical attributes of God: eternity, simplicity, omnipresence, omnipotence, omniscience, and moral perfection.

"The Library of Babel" is a short story by Argentine author and librarian Jorge Luis Borges, conceiving of a universe in the form of a vast library containing all possible 410-page books of a certain format and character set. "There is a concept that corrupts and alters all the others. I am not speaking of Evil, whose limited empire is Ethics; I speak of the Infinite", so wrote J. L. Borges.

We report here an article by Devdutt Pattanaik, an Indian mythologist, speaker, illustrator and author, known for his writing on Hindu sacred lore, legends, folklore, fables and parables. His work focuses largely on the areas of religion, mythology, and management. He has written books on the relevance of sacred stories, symbols and rituals in modern times:

"If you travel to the North Indian state of Uttar Pradesh, and visit a place called Deo-garh, which literally means citadel of the gods, you will find the ruins of a Hindu temple, one of the oldest, at least 1500 years old, built by the kings of the Gupta dynasty. On its walls, there is the image of a man reclining on the coils of a serpent with many hoods, surrounded by his wife and many warriors and sages. Its clearly inspired by a scene from the royal court. But it is clearly a celestial scene, visualisation of the moment when the world was created.

For Hindus, the world is created when Narayana awakes. Narayana is the god reclining on the serpent with multiple-hoods. When he is in dreamless slumber, the world does not exist. When he awakens, the world comes into being. Narayana is thus a visual representation of human consciousness, which awakening heralds the creation of our world.

What is interesting is the serpent on whose coils Narayana reclines. Its name is: Adi-Ananta-Sesha, which literally means Primal-Limitless-Residue, which is numerically visualised as One-Infinity-Zero. For with consciousness, we become aware of the first moment of beginnings, of limitless possibilities, and of nothingness that existed before the first moment.

The Hindu worldview has always been obsessed with infinity (everything-ness) and zero (nothing-ness) and with the number one (the beginning). More than Hindu, it is the Indic worldview, the substratum of thought which gave rise to three major ideas: Hinduism, Buddhism and Jainism, all of which speak of rebirth, cyclical time, and a world where there are no boundaries. Buddhism came up with ideas such as nirvana (oblivion) and shunya (which literally means zero). Jainism spoke of a world of endless possibilities (an-ekanta-vada).

This is in stark contrast to the Greek worldview where the world begins as chaos until the gods create order. And with order comes definitions, boundaries, certainty, and predictability. It is also different from the Abrahamic worldview where God creates the world out of nothingness and the world he creates in seven days has a definite expiry date: the Apocalypse. The Greek and Abrahamic worldviews inform what we call the Western worldview today that is obsessed with organisation, and is terrified of disorder, and unpredictability, something Indians are used to and rather comfortable with, even thriving in it.

The story goes that when Alexander, the Great, after having conquered Persia, visited India, he met a sage on the banks of the river Indus, who he referred to as a gymno-sophist or naked wise man, in Greek. This sage sat on a rock and spent all day staring at the sky. Alexander asked him what he was doing and the sage replied, "Experiencing nothingness." The sage asked Alexander what was he doing. Alexander replied, "I am conquering the world." Both laughed. Each one thought the other was a fool. For Alexander, the sage was wasting his one and only life doing nothing. For the sage, Alexander was wasting his time trying to conquer a world that has no limits, with a sense of urgency that made no sense when one lives infinite lives. Belief in one life, which is the hallmark of Greek worldview, and later Abrahamic, results us in valuing achievements. But belief in rebirth, hence infinite lives, which is the hallmark of Indic worldview, makes achievements meaningless, and puts the focus on wisdom and understanding. When the denominator of life is one, the world is different from when the denominator of life is infinity.

India's philosophical obsession with infinity and zero led to mathematicians not just conceptualising the idea of zero, but also giving it a form (a dot), and finally using it in a decimal system. This happened around the same time that the Gupta kings built the temple in Deogarh. The mathematician Brahmagupta, 638 AD, is associated with giving form to the number zero, and formulating the first rules with its usage. The rise of the decimal system enabled the writing of vast numbers, of huge value, a practice that has been traced to even Vedic texts written around 1000 BC, values that are not seen in any other parts of the world.

The Arab sea-merchants who frequented the coasts of India, and who dominated the rich spice and textile trade then (before the European sea-farers took over in the 16th century) saw value in this system and took it with them to Arabia. The Arab mathematician Khowarizimi suggested use of a little circle for zero. This circle was called 'sifr' which means 'empty', which eventually became 'zero'. Zero travelled from Arabia through Persia and Mesopotamia to Europe during the Crusades. In Spain, Fibonacci found it useful to do equations without using the abacus. Italian government was suspicious of this Arabic numbering system and so outlawed it. But the merchants used it secretly, which is why 'sifr' became 'cipher', meaning 'code'. It comes as a shock to many people that the modern use of the number zero is less than thousand years old, and that it became popular less than 500 years ago. Had it not been for the arrival of zero, neither would the Cartesian coordinate system nor calculus have developed in the 16th century. Zero enabled people to conceptualize large numbers and helped in book keeping and accounting. In the 20th century, came the binary system which forms the foundation of modern computing. All because some wild Indian sages conceptualised the universe and their gods in terms of zero and infinity."

Anselm of Canterbury, also called Anselm of Aosta after his birthplace and Anselm of Bec after his monastery, was an Italian Benedictine monk, abbot, philosopher and theologian of the Catholic Church, who held the office of Archbishop of Canterbury from 1093 to 1109. After his death, he was canonized as a saint; his feast day is 21 April. He thought that imagining something of infinitely big is categorically excluded because only God is such. "Cur Deus Homo" ("Why God was a Man") was written by Anselm from 1095 to 1098 once he was already archbishop of Canterbury as a response for requests to discuss the Incarnation. It takes the form of a dialogue between Anselm and Boso, one of his students. Its core is a purely rational argument for the necessity of the Christian mystery of atonement, the belief that Jesus's crucifixion was necessary to atone for mankind's sin. Anselm argues that, owing to the Fall and mankind's fallen nature ever since, humanity has offended God. Divine justice demands restitution for sin but human beings are incapable of providing it, as all the actions of men are already obligated to the furtherance of God's glory. Further, God's infinite justice demands infinite restitution for the impairment of his infinite dignity. The enormity of the offense led Anselm to reject personal acts of atonement, even Peter Damian's flagellation, as inadequate and ultimately vain. Instead, full recompense could only be made by God, which His infinite mercy inclines Him to provide. Atonement for humanity, however, could only be made through the figure of Jesus, as a sinless being both fully divine and fully human. Taking it upon himself to offer his own life on our behalf, his crucifixion accrues infinite worth, more than redeeming mankind and permitting it to enjoy a just will in accord with its intended nature. This interpretation is notable for permitting divine justice and mercy to be entirely compatible and has exercised immense influence over church doctrine, largely supplanting the earlier theory developed by Origen and Gregory of Nyssa that had focused primarily on Satan's power over fallen man. Cur Deus Homo is often accounted Anselm's greatest work, but the legalist and amoral nature of the argument, along with its neglect of the individuals actually being redeemed, has been criticized both by comparison with the treatment by Abelard and for its subsequent development in Protestant theology.

10.2 The Dice

In the journal Physics Today of June 1994 Robert March from the University of Wisconsin, comments as follows the book "The Broken Dice and Other Mathematical Tales of Chance" by Ivar Ekeland [111]:

"

If the two centuries of Newtonian scientific hegemony can be called the Era of Determinism, the 20th century seems destined to go down as the beginning of the Era of Chance. Chance first entered physics via the practical necessity of treating the macroscopic effects of atoms by statistical methods and soon became imbedded within the heart of quantum theory. Over the past two decades, chance has invaded the classical domain through studies of chaos; in "Broken Dice" [111], Ivar Ekeland deals with it from the point of view of a highly literate mathematician. All thoughtful human beings realize that much of what matters in their lives is at the mercy of fate, and they have personal strategies for dealing with this predicament. These strategies often have more to do with psychological comfort than with rational calculation: Most people will accept a high level of familiar risk rather than take their chances with the unknown. Any rational assessment, for example, would lead one to fear the drive to the airport more than the flight that follows, but that is not how our minds deal with risk. What Ekeland is trying to do in his book is form a bridge between the scientific and personal dimensions of chance, with the latter exemplified by fables chosen mainly from Scandinavian folklore. The book draws its title from a fabled toss of the dice by King Olaf Haraldsson of Norway to decide the ownership of an island. The King of Sweden having already thrown two sixes, Olafs situation seemed hopeless. But on his toss, one die came up a six, while the other split in two, displaying a six and a one. The fable takes this highly improbable event as a clear sign of the divine favor that ultimately led to sainthood for the revered 11th-century monarch. As a mathematician, Ekeland is particularly intrigued by the emergence of chance within his orderly discipline. Addressing the problems of the pseudorandom-number generators employed in computing, he uses very simple algorithms to illustrate how easily these generators lapse into periodicity or bias. Even though every number in the sequence is predetermined from the start, the best of them can pass an impressive battery of tests for randomness. This introduces a central theme of the book: how apparent randomness can arise from determinism, and vice versa. In one particularly dramatic example, Ekeland cites Thucydides's account of the lifting of the siege of Syracuse as the result of an Athenian commander's hesitation when faced with the spectacle of a lunar eclipse. This battle was a major factor contributing to the ultimate defeat of Athens in the Peloponnesian War. It is this eclipse, seemingly random to the ancient Greeks, that enables us to date the event precisely to 27 August 413 BC. Ekeland shows that in contests that require a blind guess at an opponent's strategy, such as the "rock-scissors-paper" trial of the familiar children's game, a random choice will always prove the optimum strategy in the long run. He also illustrates the applicability of stochastic calculus, a descendant of Albert Einstein's treatment of Brownian motion, to such diverse topics as radar filters and investment strategy. In a chapter on chaos, Ekeland introduces some laudably simple examples, while also briefly describing the Lorenz and Smale attractors. He traces the connection of chaos to fractals and shows how reasoning of this sort by Andrei Kolmogorov led to a successful treatment of turbulence. Henri Poincare first drew attention to the chaotic nature of the classical three-body problem and opined that such nonintegrable cases represent the rule rather than the exception in nature. Ekeland thus credits Poincare as the true father of chaos theory. Here we learn that our solar system, seemingly so stable over the span of recorded history, is almost certainly chaotic on a time scale of 108 years, making it a near-miracle that Earth has remained reasonably hospitable to life long enough for our species to evolve. ”

10.3 Is mathematics in our brain or is it in nature?

We report here a roundtable between four scientists who debate ideas on whether math is an inherent part of our reality, or merely something our brains use to cope with and explain our environment. Does math come from the brain or from the universe?

The discovery of the Higgs boson particle, which was predicted by mathematical formulas, shows the power of math to describe and predict the world around us–from the helical structure of DNA and the spirals of galaxies, to how rapidly epidemics spread and our universe is expanding. But is that because everything in our world is inherently mathematical and follows precise rules? Or do we tend to see mathematical patterns everywhere because of the way our brains embroider an orderly overlay over what we experience? The origins of math has become a hot topic of debate as neuroscientists continue to uncover mathematical abilities we seem to be born with, and have pinpointed regions in the brain responsible for mathematical thinking. Other scientists are finding that certain math capabilities vary by culture and depend on how we interact with the world. Both types of findings suggest math is a human construct rather than a phenomenon that determines how the cosmos is constructed.

To explore this debate on the origins of math and why it matters, The Kavli Foundation (TKF)[1] led a roundtable discussion with two physicists (physicist 1: Simeon Hellerman—Associate Professor at the Kavli Institute for Physics and Mathematics of the Universe at the University of Tokyo, Japan; physicist 2: Max Tegmark—Professor of Physics at the Massachusetts Institute of Technology, and member of MIT's Kavli Institute for Astrophysics and Space Research), a neuroscientist (Brian Butterworth—Emeritus Professor of Cognitive Neuropsychology at the Institute of Cognitive Neuroscience at University College London), and a cognitive scientist (Rafael Núñez—Professor of Cognitive Science at the University of California, San Diego and member of UCSD's Kavli Institute for Brain and Mind):

“

TKF: Let us begin by having Dr. Tegmark discuss his hypothesis that the universe is inherently mathematical in the way it is constructed and doesn't know any other way to behave other than according to mathematical rules.

Physicist 2: It's actually an old idea. Even Galileo exclaimed that the universe is a grand book written in the language of mathematics because he was wowed by all the astronomic regularities discovered in his time, for instance the precise circular or elliptical orbits of planets. Long afterwards, a whole set of subatomic particles were predicted by mathematical principles and then discovered, the Higgs boson being the most recent example. Even the possibility of curved shape of outer space was predicted centuries earlier by non-Euclidian geometry. So nature is

[1] The Kavli Foundation is dedicated to advancing science for the benefit of humanity. The foundation's mission is to stimulate basic research in astrophysics, nanoscience, neuroscience and theoretical physics; strengthen the relationship between science and society; and honor scientific discoveries with The Kavli Prize.

clearly giving us hints that the universe is mathematical. I've taken it to the extreme by proposing that our entire physical reality isn't just described by math, but that it is a mathematical structure, having no properties besides mathematical properties.

TKF: How can that be?

Physicist 2: Someone might say that their cat has the properties of being cute, black and slightly neurotic. But a cat is actually a very elaborate arrangement of particles such as electrons, which are purely mathematical objects in the sense that they have no properties other than mathematical properties—numbers like one-half, minus one, and one, for example, that we physicists have named spin, electric charge and lepton number. If we zoom out from this microscopic scale to the largest scale of our cosmos, we realize that space itself is also purely mathematical, with no properties other than mathematical properties as dimensionality, curvature and topology.

TKF: As a fellow physicist Dr. Hellerman, do you agree with Dr. Tegmark's mathematical universe hypothesis?

Physicist 1: I think many physicists, including myself, agree that there should be some complete description of the universe and the laws of nature. Implicit in that assumption is the universe is intrinsically mathematical. The discovery of the Higgs boson, which completes what physicists call the Standard Model, and others have called "the theory of everything" is a great example of this idea that there should exist some complete and mathematically consistent description of nature. Although the principles in the Standard Model hold true for atomic or subatomic particles—at the level of quantum physics–there may be inconsistencies in how they relate to fundamental physics, such as the law of gravity. But increasingly there is compelling evidence for bringing gravity into the model and still having a single internally consistent mathematical framework.

Physicist 2: Generally we feel we are taking a step forward when we get more out of a model than is put into it. The Higgs boson is a beautiful example of something that came out that we did not have to put in. Many mathematicians feel that they don't invent mathematical structures, they just discover them—that these mathematical structures exist independently of humans.

TKF: Dr. Butterworth, you hypothesize that math is a construct that stems from our brain. What supports that hypothesis?

Neuroscientist: We've pinpointed an area of the human brain where there's a specialist neural network that responds to counting the number of objects in a set. This area of the brain can recognize numbers across modalities. In other words, it can recognize three cats, three tones or three wishes. A similar area in the brain of the monkey does the same job. We even discovered the guppy — a small fish with a tiny brain — has one system for detecting small sets of up to four objects and one for larger sets. My argument is that we evolved a brain-based system for detecting and comparing the number of items in groups. Humans have developed symbolism for these numbers and elaborated on them to create the kinds of mathematics that Max and Simeon need to describe the universe. Numbers are not necessarily a property of the universe, but rather a very powerful way of describing some aspects of the universe.

TKF: So you disagree with Dr. Tegmark's notion that electrons are merely numbers?

Neuroscientist: Yes, because in order to have a physical explanation for phenomena, you have to have a cause for it. But how can a number be a cause? It's true that you can use numbers to describe electron properties, but that doesn't mean those numbers are actually a property of that physical object. Twoness is a property of a set of objects, such as two cups, or two electrons. But it is independent of the kinds of objects that are in the set for which it is a property. A set of two cups is different from a set of two electrons so twoness can't have the same causal property for cups and electrons.

TKF: Dr. Núẽz, what is your response to these hypotheses, given that your research has detected cultural differences in math abilities and suggests many mathematical principles are learned from our interactions with the world?

Cognitivist: I agree with Brian that numbers are not properties of the universe, but rather that they reflect the biological grounding for how people make sense of the world. Math is a form of human imagination that is not only brain-based but also culturally shaped – and this is crucial. It's true that without a brain we can't do math, but it's also true that we need a brain to play the piano or tennis or go snowboarding. And none of these actions are genetically determined. We need a brain for all of them, but we also need a sophisticated cultural apparatus that shapes how those basic brain functions are recruited and expressed. Brain areas support the invention of mathematical principles, but these principles don't come straight out of a particular area of the brain.

TKF: Can you give an example that supports the notion that math can be culturally shaped?

Cognitivist: Take the mathematical notion that '0 factorial = 1'. This 'truth' doesn't exist anywhere in the universe, and it doesn't come out straight from brain activity. But in the culture of mathematical practice, certain mathematicians realized that they needed this 'truth' for certain things to work out, and adopted it. In modern mathematics this is routinely done via formal definitions and axioms. These are outcomes of cultural practice – not merely conventional, but highly constrained cultural practices. In the domain of numbers, I've done research in remote areas of the world, such as Papua New Guinea, and in the highlands of the Andes. Some cultures operate with precise numbers concepts and others don't have the concepts for, say, the numbers 8 or 11 – their languages don't have words that discriminate those numbers from something like 9 or 10. When you investigate these niches of cultural practices, you see some fundamental notions of number that are not present, such as precision, for example.

Neuroscientist: Are you saying that math is a cultural invention, which is sort of arbitrary?

Cognitivist: No, because culture is not arbitrary. Cultural practices are constrained, among others, by the biology of the individuals that make up the culture. Speech accents, for example, are related to cultural (linguistic) practices that are not genetically determined – nothing in my genes says that my native language is Spanish and that I speak English with a Spanish accent. And humans can't just produce any arbitrary sound they want in any frequency – because they are highly constrained biologically. So it's not purely arbitrary.

Neuroscientist: You said that if you don't have the word for nine, you're not going to have the concept of nine. But John Locke, the British philosopher of the 17th century, reported talking to Amazonian Indians that had no number words beyond 5. Yet, if he asked them to explain to him about larger numbers, these Indians would hold up their fingers as well as the fingers of other people present in order to show what these larger numbers were. So they had a concept of all these numbers, even though they didn't have any words for them. Our own research in Australian cultures that have no counting words shows that if you present in a culturally appropriate way, you'll find these children have the same concepts of numbers and arithmetic that kids brought up speaking English do.

Cognitivist: I agree that we could have an idea of a regular polygon with 103 sides, even though we don't have a name for it. But I don't think that this is the essence of the question. In fact, I don't think that the origin of math is ultimately about numbers. Instead, it is much more about logical constraints, postulates and axioms, inferential mechanisms, and so on. A good accountant who does a lot of number crunching does not make for a good mathematician. Number may play a role, but is not necessarily the cornerstone of mathematics. And we have lots of different logical principles or axioms from which to choose, each of which may be internally consistent but inconsistent with others. So you can't just say, for example, that a particular statement about infinity is true in the universe because its truth status will depend on the axioms you start with and those are concocted from the human imagination, which is mediated by language and culturally shaped. There is no inherent single form of logic in the universe. Humans operate with different types of logics in different contexts and for different purposes.

Physicist 1: But we know that given the usual rules of logical inference, it's possible to construct all operations involving numbers. So we can agree that whole numbers and the laws of all forms of geometries are consistent and universal, whether or not they may be realized in nature.

Neuroscientist: It's not clear that you can derive the properties of numbers from logic alone or that having an arithmetic technology is necessary for logic. It may make doing complicated shades of logical reasoning easier. In any case, formal logic is not going to turn out to be sufficient to give you any of the kinds of mathematics that we are interested in, even the relatively simple arithmetic we're familiar with. I think formal reasoning stems from our frontal lobes of the brain and there are some axioms about numbers that come from the parietal lobes of the brain. The frontal lobe operates on these numerical concepts in order to give us what we understand as the rest of mathematics.

Physicist 2: When different cultures evolve, they aren't all going to come up with the concepts and words for all the different mathematical structures, but I think they will all come up with some of the most useful concepts. All cultures find it useful to distinguish between one and two, so they can know if they left one kid behind in the forest–ducks are really good at keeping track of how many babies they have swimming after them–whereas studying abstract algebra may not be something important to all cultures.

Cognitivist: That's right. Beginning with Galileo's time, the math that was created and developed became intimately intertwined with physics so it fit the phenomena humans observed in nature. For centuries now, we have been cherry-picking the math that has been useful and discarded the math that hasn't. At this point, contemporary physics can no longer exist without the mathematics that goes with it. You ascribe the number properties as if they are in the universe, but in fact in mathematics there are all kinds of choices that have been made beforehand for that very mathematics to be what it is. For example, set theory says that the empty set is a subset of every set, even though we don't see that fact physically materialized anywhere in the universe. Yet, we now realize that such 'truth' is 'needed' and therefore we make it true. This kind of cherry-picking has happened all over the history of math, essentially after the 19th century with the invention of non-Euclidian geometry, which altered certain postulates and axioms set previously, and with the creation of modern new logical systems.

Physicist 2: The fantastic twist of this is that non-Euclidian geometry was invented almost 200 years ago when physicists thought it didn't describe our own physical space, which they thought was flat, not curved, so two parallel lines could never cross. Then Einstein came along and after studying non-Euclidean geometry supposed space was curved and that this suggested light would bend around the Sun, which it does, and that there could be black holes, which were later found. Don't you think it's surprising that such mathematics could predict things in nature that we later found?

Cognitivist: Yes, at first glance it seems surprising, but when you dig in a little more you realize that not all the tools that mathematicians have invented have been useful in physics in finding new things. We humans are pretty good at trying to make sense of things and excel at developing new tools for such purposes. You are giving examples for cases in which mathematics does work apparently in nature. But, how about all those cases for which it doesn't, including for making precise weather predictions? The saga of mathematics in science has been to invent new mathematical tools that help make testable predictions and to keep those that work, while discarding those that aren't useful. But there are tons of other things in pure mathematics that aren't testable or useful in empirical science proper.

Neuroscientist: What about things that can only be described using probability, such the position of an electron at any point in time. How does that fit in with your hypothesis Max?

Physicist 2: Quantum mechanics famously threw that monkey wrench into the old idea of causality when it turned out there are certain experiments where you can't say for sure what's going to happen. But you can take a purely mathematical description, known as the Schrödinger equation, and say that it always applies to everything, so there is no random or indeterminate thing about that. It just means that the actual full reality is bigger than the reality that we can see.

TKF: Are you saying that to us it feels subjective and random, but above it all there is this order that we just can't perceive?

Physicist 2: Yes. It's like if they put one clone of you in a room labeled A and the original you in a room labeled B. When you come out the next morning and look at your room label, you can't predict whether you are going to see A or B because you have no way of knowing whether you're the clone. So it's going to seem subjectively random to you whether you come out of room A or room B. But someone who is observing both you and your clone will be able to predict that if your clone comes out of room A, than your original version will come out of room B.

TKF: Let's end our discussion by talking about why we need to understand the origins of math. Are there practical implications for each theory that you've suggested?

Neuroscientist: Understanding the origins of math is important for education. If we have an innate system that underlies much of our mathematical abilities, then things can go wrong with it's genetic transmission in the brain, so there will be some people who aren't going to be able to learn this arithmetic in the usual way. You have to find different ways to teach these people, just like you have to find different ways to teach dyslexics to read.

Physicist 2: If math is inherent out in the universe, then mathematics can give us hints for solving future problems in physics. If we really believe that nature is fundamentally mathematical, then we should look for mathematical patterns and regularities when we come across phenomena that we don't understand. This problem-solving approach has been at the heart of physics' success for the past 500 years.

Physicist 1: I agree with Max and want to add that, in the physical sciences, the gold standard of a theory is that it predicts qualitatively new phenomena. If we thought math was so culture bound and flexible that it could describe whatever you observe – maybe there's a Higgs boson, maybe not, and math can describe either situation on a democratic basis – then there would be a great deal in physics we wouldn't bother doing and we never would have had the successes that we've had.

Cognitivist: I agree with Brian that understanding the origins of math has a tremendous impact on what education could or should be. It also has implications for understanding other cultures' beliefs and logics. Many wars are due to not understanding another culture's logic. Logical systems embody mathematical principles that are incorporated in our legal systems and religions, both of which prescribe behavior. Understanding the origins of math will help us understand human nature better.

Physicist 2: I've really enjoyed this interdisciplinary conversation. Perhaps the reason why Simeon and I are more gung-ho about nature being mathematical than the neuroscientists is that it's much easier to study and mathematically describe one little electron than to study the zillions of electrons that comprise the human brain. There is beautiful complexity there and we have a lot of work cut out for us, even if nature is ultimately mathematical at the root.

Neuroscientist: There are still some unanswered questions. For example, would the Higgs boson exist if there wasn't the mathematics to describe it? Perhaps this is a question best solved after a few drinks. ”

10.4 The anthropic principle

The *anthropic principle* is the principle that there is a restrictive lower bound on how statistically probable our observations of the universe are, given that we could only exist in the particular type of universe capable of developing and sustaining sentient life. Proponents of the anthropic principle argue that it explains why this universe has the age and the fundamental physical constants necessary to accommodate conscious life, since if either had been different, we would not have been around to make observations. Anthropic reasoning is often used to deal with the notion that the universe seems to be fine tuned.

There are many different formulations of the anthropic principle. But the underlying principles can be divided into"weak" and "strong" forms, depending on the types of cosmological claims they entail. The weak anthropic principle (WAP), such as the one defined by Brandon Carter, states that the universe's ostensible fine tuning is the result of selection bias (specifically survivorship bias). Sometimes such arguments draw upon some notion of the multiverse for there to be a statistical population of universes to select from. However, a single vast universe is sufficient for most forms of the WAP that do not specifically deal with fine tuning. The strong anthropic principle (SAP), as proposed by John D. Barrow and Frank Tipler, states that the universe is in some sense compelled to eventually have conscious and sapient life emerge within it.

We report here a part of chapter 8 of the book "A Brief History of Time: From the Big Bang to Black Holes" of Stephen Hawking [112]:

"

If the universe is indeed spatially infinite, or if there are infinitely many universes, there would probably be some large regions somewhere that started out in a smooth and uniform manner. It is a bit like the well-known horde of monkeys hammering away on typewriters – most of what they write will be garbage, but very occasionally by pure chance they will type out one of Shakespeare's sonnets. Similarly, in the case of the universe, could it be that we are living in a region that just happens by chance to be smooth and uniform? At first sight this might seem very improbable, because such smooth regions would be heavily outnumbered by chaotic and irregular regions. However, suppose that only in the smooth regions were galaxies and stars formed and were conditions right for the development of complicated self-replicating organisms like ourselves who were capable of asking the question: why is the universe so smooth? This is an example of the application of what is known as the anthropic principle, which can be paraphrased as, 'We see the universe the way it is because we exist.'

There are two versions of the anthropic principle, the weak and the strong. The weak anthropic principle states that in a universe that is large or infinite in space and/or time, the conditions necessary for the development of intelligent life will be met only in certain regions that are limited in space and time. The intelligent beings in these regions should therefore not be surprised if they observe that their locality in the universe satisfies the conditions that are necessary for their existence. It is a bit like a rich person living in a wealthy neighborhood not seeing any poverty.

One example of the use of the weak anthropic principle is to 'explain' why the big bang occurred about ten thousand million years ago – it takes about that long for intelligent beings to evolve. As explained above, an early generation of stars first had to form. These stars converted some of the original hydrogen and helium into elements like carbon and oxygen, out of which we are made. The stars then exploded as supernovas, and their debris went to form other stars and planets, among them those of our Solar System, which is about five thousand million years old. The first one or two thousand million years of the earth's existence were too hot for the development of anything complicated. The remaining three thousand million years or so have been taken up by the slow process of biological evolution, which has led from the simplest organisms to beings who are capable of measuring time back to the big bang.

Few people would quarrel with the validity or utility of the weak anthropic principle. Some, however, go much further and propose a strong version of the principle. According to this theory, there are either many different universes or many different regions of a single universe, each with its own initial configuration and, perhaps, with its own set of laws of science. In most of these universes the conditions would not be right for the development of complicated organisms; only in the few universes that are like ours would intelligent beings develop and ask the question, 'Why is the universe the way we see it?' The answer is then simple: if it had been different, we would not be here!

The laws of science, as we know them at present, contain many fundamental numbers, like the size of the electric charge of the electron and the ratio of the masses of the proton and the electron. We cannot, at the moment at least, predict the values of these numbers from theory – we have to find them by observation. It may be that one day we shall discover a complete unified theory that predicts them all, but it is also possible that some or all of them vary from universe to universe or within a single universe. The remarkable fact is that the values of these numbers seem to have been very finely adjusted to make possible the development of life. For example, if the electric charge of the electron had been only slightly different, stars either would have been unable to burn hydrogen and helium, or else they would not have exploded. Of course, there might be other forms of intelligent life, not dreamed of even by writers of science fiction, that did not require the light of a star like the sun or the heavier chemical elements that are made in stars and are flung back into space when the stars explode. Nevertheless, it seems clear that there are relatively few ranges of values for the numbers that would allow the development of any form of intelligent life. Most sets of values would give rise to universes that, although they might be very beautiful, would contain no one able to wonder at that beauty. One can take this either as evidence of a divine purpose in Creation and the choice of the laws of science or as support for the strong anthropic principle.

There are a number of objections that one can raise to the strong anthropic principle as an explanation of the observed state of the universe. First, in what sense can all these different universes be said to exist? If they are really separate from each other, what happens in another universe can have no observable consequences in our own universe. We should therefore use the principle of economy and cut them out of the theory. If, on the other hand, they are just different regions of a single universe, the laws of science would have to be the same in each region, because otherwise one could not move continuously from one region to another. In this case the only difference between the regions would be their initial configurations and so the strong anthropic principle would reduce to the weak one.

A second objection to the strong anthropic principle is that it runs against the tide of the whole history of science. We have developed from the geocentric cosmologies of Ptolemy and his forebears, through the heliocentric cosmology of Copernicus and Galileo, to the modern picture in which the earth is a medium-sized planet orbiting around an average star in the outer suburbs of an ordinary spiral galaxy, which is itself only one of about a million million galaxies in the observable universe. Yet the strong anthropic principle would claim that this whole vast construction exists simply for our sake. This is very hard to believe. Our Solar System is certainly a prerequisite for our existence, and one might extend this to the whole of our galaxy to allow for an earlier generation of stars that created the heavier elements. But there does not seem to be any need for all those other galaxies, nor for the universe to be so uniform and similar in every direction on the large scale.

One would feel happier about the anthropic principle, at least in its weak version, if one could show that quite a number of different initial configurations for the universe would have evolved to produce a universe like the one we observe. If this is the case, a universe that developed from some sort of random initial conditions should contain a number of regions that are smooth and uniform and are suitable for the evolution of intelligent life. On the other hand, if the initial state of the universe had to be chosen extremely carefully to lead to something like what we see around us, the universe would be unlikely to contain any region in which life would appear. In the hot big bang model described above, there was not enough time in the early universe for heat to have flowed from one region to another. This means that the initial state of the universe would have had to have exactly the same temperature everywhere in order to account for the fact that the microwave background has the same temperature in every direction we look. The initial rate of expansion also would have had to be chosen very precisely for the rate of expansion still to be so close to the critical rate needed to avoid recollapse. This means that the initial state of the universe must have been very carefully chosen indeed if the hot big bang model was correct right back to the beginning of time. It would be very difficult to explain why the universe should have begun in just this way, except as the act of a God who intended to create beings like us.

In an attempt to find a model of the universe in which many different initial configurations could have evolved to something like the present universe, a scientist at the Massachusetts Institute of Technology, Alan Guth, suggested that the early universe might have gone through a period of very rapid expansion. This expansion is said to be 'inflationary,' meaning that the universe at one time expanded at an increasing rate rather than the decreasing rate that it does today. According to Guth, the radius of the universe increased by a million million million million million (1 with thirty zeros after it) times in only a tiny fraction of a second.

Guth suggested that the universe started out from the big bang in a very hot, but rather chaotic, state. These high temperatures would have meant that the particles in the universe would be moving very fast and would have high energies. As we discussed earlier, one would expect that at such high temperatures the strong and weak nuclear forces and the electromagnetic force would all be unified into a single force. As the universe expanded, it would cool, and particle energies would go down. Eventually there would be what is called a phase transition and the symmetry between the forces would be broken: the strong force would become different from the weak and electromagnetic forces. One common example of a phase transition is the freezing of water when you cool it down. Liquid water is symmetrical, the same at every point and in every direction. However, when ice crystals form, they will have definite positions and will be lined up in some direction. This breaks water's symmetry.

In the case of water, if one is careful, one can 'supercool' it: that is, one can reduce the temperature below the freezing point (0°C) without ice forming. Guth suggested that the universe might behave in a similar way: the temperature might drop below the critical value without the symmetry between the forces being broken. If this happened, the universe would be in an unstable state, with more energy than if the symmetry had been broken. This special extra energy can be shown to have an antigravitational effect: it would have acted just like the cosmological constant that Einstein introduced into general relativity when he was trying to construct a static model of the universe. Since the universe would already be expanding just as in the hot big bang model, the repulsive effect of this cosmological constant would therefore have made the universe expand at an ever-increasing rate. Even in regions where there were more matter particles than average, the gravitational attraction of the matter would have been outweighed by the repulsion of the effective cosmological constant. Thus these regions would also expand in an accelerating inflationary manner. As they expanded and the matter particles got farther apart, one would be left with an expanding universe that contained hardly any particles and was still in the supercooled state. Any irregularities in the universe would simply have been smoothed out by the expansion, as the wrinkles in a balloon are smoothed away when you blow it up. Thus the present smooth and uniform state of the universe could have evolved from many different non-uniform initial states.

In such a universe, in which the expansion was accelerated by a cosmological constant rather than slowed down by the gravitational attraction of matter, there would be enough time for light to travel from one region to another in the early universe. This could provide a solution to the problem, raised earlier, of why different regions in the early universe have the same properties. Moreover, the rate of expansion of the universe would automatically become very close to the critical rate determined by the energy density of the universe. This could then explain why the rate of expansion is still so close to the critical rate, without having to assume that the initial rate of expansion of the universe was very carefully chosen.

The idea of inflation could also explain why there is so much matter in the universe. There are something like ten million million million million million million million million million million million million million million (1 with eighty zeros after it) particles in the region of the universe that we can observe. Where did they all come from? The answer is that, in quantum theory, particles can be created out of energy in the form of particle/antiparticle pairs. But that just raises the question of where the energy came from. The answer is that the total energy of the universe is exactly zero. The matter in the universe is made out of positive energy. However, the matter is all attracting itself by gravity. Two pieces of matter that are close to each other have less energy than the same two pieces a long way apart, because you have to expend energy to separate them against the gravitational force that is pulling them together. Thus, in a sense, the gravitational field has negative energy. In the case of a universe that is approximately uniform in space, one can show that this negative gravitational energy exactly cancels the positive energy represented by the matter. So the total energy of the universe is zero.

Now twice zero is also zero. Thus the universe can double the amount of positive matter energy and also double the negative gravitational energy without violation of the conservation of energy. This does not happen in the normal expansion of the universe in which the matter energy density goes down as the universe gets bigger. It does happen, however, in the inflationary expansion because the energy density of the supercooled state remains constant while the universe expands: when the universe doubles in size, the positive matter energy and the negative gravitational energy both double, so the total energy remains zero. During the inflationary phase, the universe increases its size by a very large amount. Thus the total amount of energy available to make particles becomes very large. As Guth has remarked, 'It is said that there's no such thing as a free lunch. But the universe is the ultimate free lunch.'

The universe is not expanding in an inflationary way today. Thus there has to be some mechanism that would eliminate the very large effective cosmological constant and so change the rate of expansion from an accelerated one to one that is slowed down by gravity, as we have today. In the inflationary expansion one might expect that eventually the symmetry between the forces would be broken, just as supercooled water always freezes in the end. The extra energy of the unbroken symmetry state would then be released and would reheat the universe to a temperature just below the critical temperature for symmetry between the forces. The universe would then go on to expand and cool just like the hot big bang model, but there would now be an explanation of why the universe was expanding at exactly the critical rate and why different regions had the same temperature.

In Guth's original proposal the phase transition was supposed to occur suddenly, rather like the appearance of ice crystals in very cold water. The idea was that 'bubbles' of the new phase of broken symmetry would have formed in the old phase, like bubbles of steam surrounded by boiling water. The bubbles were supposed to expand and meet up with each other until the whole universe was in the new phase. The trouble was, as I and several other people pointed out, that the universe was expanding so fast that even if the bubbles grew at the speed of light, they would be moving away from each other and so could not join up. The universe would be left in a very nonuniform state, with some regions still having symmetry between the different forces. Such a model of the universe would not correspond to what we see.

In October 1981, I went to Moscow for a conference on quantum gravity. After the conference I gave a seminar on the inflationary model and its problems at the Sternberg Astronomical

Institute. Before this, I had got someone else to give my lectures for me, because most people could not understand my voice. But there was not time to prepare this seminar, so I gave it myself, with one of my graduate students repeating my words. It worked well, and gave me much more contact with my audience. In the audience was a young Russian, Andrei Linde, from the Lebedev Institute in Moscow. He said that the difficulty with the bubbles not joining up could be avoided if the bubbles were so big that our region of the universe is all contained inside a single bubble. In order for this to work, the change from symmetry to broken symmetry must have taken place very slowly inside the bubble, but this is quite possible according to grand unified theories. Linde's idea of a slow breaking of symmetry was very good, but I later realized that his bubbles would have to have been bigger than the size of the universe at the time! I showed that instead the symmetry would have broken everywhere at the same time, rather than just inside bubbles. This would lead to a uniform universe, as we observe. I was very excited by this idea and discussed it with one of my students, Ian Moss. As a friend of Linde's, I was rather embarrassed, however, when I was later sent his paper by a scientific journal and asked whether it was suitable for publication. I replied that there was this flaw about the bubbles being bigger than the universe, but that the basic idea of a slow breaking of symmetry was very good. I recommended that the paper be published as it was because it would take Linde several months to correct it, since anything he sent to the West would have to be passed by Soviet censorship, which was neither very skilful nor very quick with scientific papers. Instead, I wrote a short paper with Ian Moss in the same journal in which we pointed out this problem with the bubble and showed how it could be resolved.

The day after I got back from Moscow I set out for Philadelphia, where I was due to receive a medal from the Franklin Institute. My secretary, Judy Fella, had used her not inconsiderable charm to persuade British Airways to give herself and me free seats on a Concorde as a publicity venture. However, I was held up on my way to the airport by heavy rain and I missed the plane. Nevertheless, I got to Philadelphia in the end and received my medal. I was then asked to give a seminar on the inflationary universe at Drexel University in Philadelphia. I gave the same seminar about the problems of the inflationary universe, just as in Moscow.

A very similar idea to Linde's was put forth independently a few months later by Paul Steinhardt and Andreas Albrecht of the University of Pennsylvania. They are now given joint credit with Linde for what is called 'the new inflationary model,' based on the idea of a slow breaking of symmetry. (The old inflationary model was Guth's original suggestion of fast symmetry breaking with the formation of bubbles.)

The new inflationary model was a good attempt to explain why the universe is the way it is. However, I and several other people showed that, at least in its original form, it predicted much greater variations in the temperature of the microwave background radiation than are observed. Later work has also cast doubt on whether there could be a phase transition in the very early universe of the kind required. In my personal opinion, the new inflationary model is now dead as a scientific theory, although a lot of people do not seem to have heard of its demise and are still writing papers as if it were viable. A better model, called the chaotic inflationary model, was put forward by Linde in 1983. In this there is no phase transition or supercooling. Instead, there is a spin 0 field, which, because of quantum fluctuations, would have large values in some regions of the early universe. The energy of the field in those regions would behave like a cosmological constant. It would have a repulsive gravitational effect, and thus make those regions expand in an inflationary manner. As they expanded, the energy of the field in them would slowly decrease until the inflationary expansion changed to an expansion like that in the hot big bang model. One of these regions would become what we now see as the observable universe. This model has all the advantages of the earlier inflationary models, but it does not depend on a dubious phase transition, and it can moreover give a reasonable size for the fluctuations in the temperature of the microwave background that agrees with observation.

This work on inflationary models showed that the present state of the universe could have arisen from quite a large number of different initial configurations. This is important, because it shows that the initial state of the part of the universe that we inhabit did not have to be chosen with great care. So we may, if we wish, use the weak anthropic principle to explain why the universe looks the way it does now. It cannot be the case, however, that every initial configuration would have led to a universe like the one we observe. One can show this by considering a very different state for the universe at the present time, say, a very lumpy and irregular one. One could use the laws of science to evolve the universe back in time to determine its configuration at earlier times. According to the singularity theorems of classical general relativity, there would still have been a big bang singularity. If you evolve such a universe forward in time according to the laws of science, you will end up with the lumpy and irregular state you started with. Thus there must have been initial configurations that would not have given rise to a universe like the one we see today. So even the inflationary model does not tell us why the initial configuration was not such as to produce something very different from what we observe. Must we turn to the anthropic principle for an explanation? Was it all just a lucky chance? That would seem a counsel of despair, a negation of all our hopes of understanding the underlying order of the universe.

In order to predict how the universe should have started off, one needs laws that hold at the beginning of time. If the classical theory of general relativity was correct, the singularity theorems that Roger Penrose and I proved show that the beginning of time would have been a point of infinite density and infinite curvature of space-time. All the known laws of science would break down at such a point. One might suppose that there were new laws that held at singularities, but it would be very difficult even to formulate such laws at such badly behaved points, and we would have no guide from observations as to what those laws might be. However, what the singularity theorems really indicate is that the gravitational field becomes so strong that quantum gravitational effects become important: classical theory is no longer a good description of the universe. So one has to use a quantum theory of gravity to discuss the very early stages of the universe. As we shall see, it is possible in the quantum theory for the ordinary laws of science to hold everywhere, including at the beginning of time: it is not necessary to postulate new laws for singularities, because there need not be any singularities in the quantum theory.

We don't yet have a complete and consistent theory that combines quantum mechanics and gravity. However, we are fairly certain of some features that such a unified theory should have. One is that it should incorporate Feynman's proposal to formulate quantum theory in terms of a sum over histories. In this approach, a particle does not have just a single history, as it would in a classical theory. Instead, it is supposed to follow every possible path in space-time, and with each of these histories there are associated a couple of numbers, one representing the size of a wave and the other representing its position in the cycle (its phase). The probability that the particle, say, passes through some particular point is found by adding up the waves associated with every possible history that passes through that point. When one actually tries to perform these sums, however, one runs into severe technical problems. The only way around these is the following peculiar prescription: one must add up the waves for particle histories that are not in the 'real' time that you and I experience but take place in what is called imaginary time. Imaginary time may sound like science fiction but it is in fact a well-defined mathematical concept. If we take any ordinary (or 'real') number and multiply it by itself, the result is a positive number. (For example, 2 times 2 is 4, but so is -2 times -2.) There are, however, special numbers (called imaginary numbers) that give negative numbers when multiplied by themselves. (The one called i, when multiplied by itself, gives -1, 2i multiplied by itself gives -4, and so on.)

One can picture real and imaginary numbers in the following way. The real numbers can be represented by a line going from left to right, with zero in the middle, negative numbers like -1, -2, etc. on the left, and positive numbers, 1, 2, etc. on the right. Then imaginary numbers are represented by a line going up and down the page, with i, 2i, etc. above the middle, and -i, -2i, etc. below. Thus imaginary numbers are in a sense numbers at right angles to ordinary real numbers.

To avoid the technical difficulties with Feynman's sum over histories, one must use imaginary time. That is to say, for the purposes of the calculation one must measure time using imaginary numbers, rather than real ones. This has an interesting effect on space-time: the distinction between time and space disappears completely. A space-time in which events have imaginary values of the time coordinate is said to be Euclidean, after the ancient Greek Euclid, who founded the study of the geometry of two-dimensional surfaces. What we now call Euclidean space-time is very similar except that it has four dimensions instead of two. In Euclidean space-time there is no difference between the time direction and directions in space. On the other hand, in real space-time, in which events are labeled by ordinary, real values of the time coordinate, it is easy to tell the difference – the time direction at all points lies within the light cone, and space directions lie outside. In any case, as far as everyday quantum mechanics is concerned, we may regard our use of imaginary time and Euclidean space-time as merely a mathematical device (or trick) to calculate answers about real space-time.

A second feature that we believe must be part of any ultimate theory is Einstein's idea that the gravitational field is represented by curved spacetime: particles try to follow the nearest thing to a straight path in a curved space, but because space-time is not flat their paths appear to be bent, as if by a gravitational field. When we apply Feynman's sum over histories to Einstein's view of gravity, the analogue of the history of a particle is now a complete curved space-time that represents the history of the whole universe. To avoid the technical difficulties in actually performing the sum over histories, these curved space-times must be taken to be Euclidean. That is, time is imaginary and is indistinguishable from directions in space. To calculate the probability of finding a real space-time with some certain property, such as looking the same at every point and in every direction, one adds up the waves associated with all the histories that have that property.

In the classical theory of general relativity, there are many different possible curved space-times, each corresponding to a different initial state of the universe. If we knew the initial state of our universe, we would know its entire history. Similarly, in the quantum theory of gravity, there are many different possible quantum states for the universe. Again, if we knew how the Euclidean curved space-times in the sum over histories behaved at early times, we would know the quantum state of the universe.

In the classical theory of gravity, which is based on real space-time, there are only two possible ways the universe can behave: either it has existed for an infinite time, or else it had a beginning at a singularity at some finite time in the past. In the quantum theory of gravity, on the other hand, a third possibility arises. Because one is using Euclidean space-times, in which the time direction is on the same footing as directions in space, it is possible for space-time to be finite in extent and yet to have no singularities that formed a boundary or edge. Space-time would be like the surface of the earth, only with two more dimensions. The surface of the earth is finite in extent but it doesn't have a boundary or edge: if you sail off into the sunset, you don't fall off the edge or run into a singularity. (I know, because I have been round the world!)

If Euclidean space-time stretches back to infinite imaginary time, or else starts at a singularity in imaginary time, we have the same problem as in the classical theory of specifying the initial state of the universe: God may know how the universe began, but we cannot give any particular reason for thinking it began one way rather than another. On the other hand, the quantum theory of gravity has opened up a new possibility, in which there would be no boundary to space-time and so there would be no need to specify the behavior at the boundary. There would be no singularities at which the laws of science broke down, and no edge of space-time at which one would have to appeal to God or some new law to set the boundary conditions for space-time. One could say: 'The boundary condition of the universe is that it has no boundary.' The universe would be completely self-contained and not affected by anything outside itself. It would neither be created nor destroyed. It would just BE.

It was at the conference in the Vatican mentioned earlier that I first put forward the suggestion that maybe time and space together formed a surface that was finite in size but did not have any boundary or edge. My paper was rather mathematical, however, so its implications for the role of God in the creation of the universe were not generally recognized at the time (just as well for me). At the time of the Vatican conference, I did not know how to use the 'no boundary' idea to make predictions about the universe. However I spent the following summer at the University of California, Santa Barbara. There a friend and colleague of mine, Jim Hartle, worked out with me what conditions the universe must satisfy if space-time had no boundary. When I returned to Cambridge, I continued this work with two of my research students, Julian Luttrel and Jonathan Halliwell.

I'd like to emphasize that this idea that time and space should be finite 'without boundary' is just a proposal: it cannot be deduced from some other principle. Like any other scientific theory, it may initially be put forward for aesthetic or metaphysical reasons, but the real test is whether it makes predictions that agree with observation. This, however, is difficult to determine in the case of quantum gravity, for two reasons. First, as will be explained in [a later] chapter, we are not yet sure exactly which theory successfully combines general relativity and quantum mechanics, though we know quite a lot about the form such a theory must have. Second, any model that described the whole universe in detail would be much too complicated mathematically for us to be able to calculate exact predictions. One therefore has to make simplifying assumptions and approximations – and even then, the problem of extracting predictions remains a formidable one.

Each history in the sum over histories will describe not only the space-time but everything in it as well, including any complicated organisms like human beings who can observe the history of the universe. This may provide another justification for the anthropic principle, for if all the histories are possible, then so long as we exist in one of the histories, we may use the anthropic principle to explain why the universe is found to be the way it is. Exactly what meaning can be attached to the other histories, in which we do not exist, is not clear. This view of a quantum theory of gravity would be much more satisfactory, however, if one could show that, using the sum over histories, our universe is not just one of the possible histories but one of the most probable ones. To do this, we must perform the sum over histories for all possible Euclidean space-times that have no boundary.

Under the 'no boundary' proposal one learns that the chance of the universe being found to be following most of the possible histories is negligible, but there is a particular family of histories that are much more probable than the others. These histories may be pictured as being like the surface of the earth, with the distance from the North Pole representing imaginary time and the size of a circle of constant distance from the North Pole representing the spatial size of the universe. The universe starts at the North Pole as a single point. As one moves south, the circles of latitude at constant distance from the North Pole get bigger, corresponding to the universe expanding with imaginary time (Fig. 10.3). The universe would reach a maximum size at the equator and would contract with increasing imaginary time to a single point at the South Pole. Even though the universe would have zero size at the North and South Poles, these points would not be singularities, any more than the North and South Poles on the earth are singular. The laws of science will hold at them, just as they do at the North and South Poles on the earth.

The history of the universe in real time, however, would look very different. At about ten

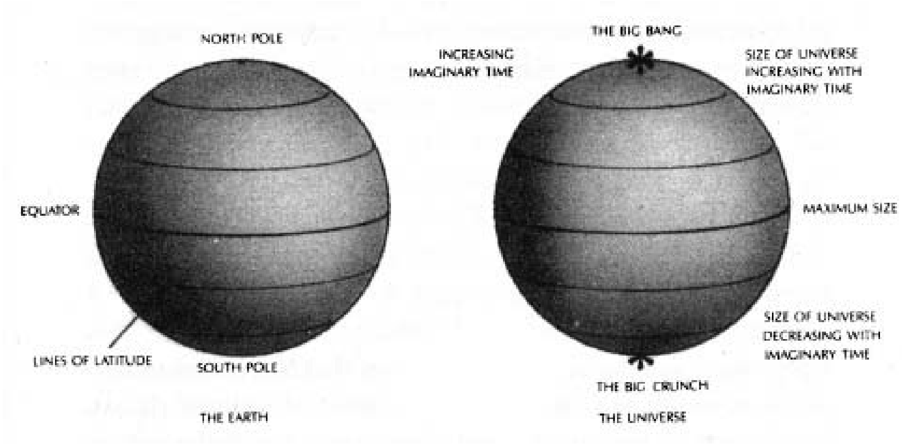


Figure 10.3: The universe in euclidean time with the topology of a sphere.

or twenty thousand million years ago, it would have a minimum size, which was equal to the maximum radius of the history in imaginary time. At later real times, the universe would expand like the chaotic inflationary model proposed by Linde (but one would not now have to assume that the universe was created somehow in the right sort of state). The universe would expand to a very large size and eventually it would collapse again into what looks like a singularity in real time. Thus, in a sense, we are still all doomed, even if we keep away from black holes. Only if we could picture the universe in terms of imaginary time would there be no singularities.

If the universe really is in such a quantum state, there would be no singularities in the history of the universe in imaginary time. It might seem therefore that my more recent work had completely undone the results of my earlier work on singularities. But, as indicated above, the real importance of the singularity theorems was that they showed that the gravitational field must become so strong that quantum gravitational effects could not be ignored. This in turn led to the idea that the universe could be finite in imaginary time but without boundaries or singularities. When one goes back to the real time in which we live, however, there will still appear to be singularities. The poor astronaut who falls into a black hole will still come to a sticky end; only if he lived in imaginary time would he encounter no singularities.

This might suggest that the so-called imaginary time is really the real time, and that what we call real time is just a figment of our imaginations. In real time, the universe has a beginning and an end at singularities that form a boundary to space-time and at which the laws of science break down. But in imaginary time, there are no singularities or boundaries. So maybe what we call imaginary time is really more basic, and what we call real is just an idea that we invent to help us describe what we think the universe is like. But according to the approach I described in [a previous] chapter, a scientific theory is just a mathematical model we make to describe our observations: it exists only in our minds. So it is meaningless to ask: which is real, ‘real’ or ‘imaginary’ time? It is simply a matter of which is the more useful description.

One can also use the sum over histories, along with the no boundary proposal, to find which properties of the universe are likely to occur together. For example, one can calculate the probability that the universe is expanding at nearly the same rate in all different directions at a time when the density of the universe has its present value. In the simplified models that have been examined so far, this probability turns out to be high; that is, the proposed no boundary condition leads to the prediction that it is extremely probable that the present rate of expansion of the universe is almost the same in each direction. This is consistent with the observations of the microwave background radiation, which show that it has almost exactly the same intensity in any direction. If the universe were expanding faster in some directions than in others, the intensity of the radiation in those directions would be reduced by an additional red shift.

Further predictions of the no boundary condition are currently being worked out. A particularly interesting problem is the size of the small departures from uniform density in the early universe that caused the formation first of the galaxies, then of stars, and finally of us. The uncertainty principle implies that the early universe cannot have been completely uniform because there must have been some uncertainties or fluctuations in the positions and velocities of the particles. Using the no boundary condition, we find that the universe must in fact have started off with just the minimum possible non-uniformity allowed by the uncertainty principle. The universe would have then undergone a period of rapid expansion, as in the inflationary models. During this period, the initial non-uniformities would have been amplified until they were big enough to explain the origin of the structures we observe around us. In 1992 the Cosmic Background Explorer satellite (COBE) first detected very slight variations in the intensity of the microwave background with direction. The way these non-uniformities depend on direction seems to agree with the predictions of the inflationary model and the no boundary proposal. Thus the no boundary proposal is a good scientific theory in the sense of Karl Popper: it could have been falsified by observations but instead its predictions have been confirmed. In an expanding universe in which the density of matter varied slightly from place to place, gravity would have caused the denser regions to slow down their expansion and start contracting. This would lead to the formation of galaxies, stars, and eventually even insignificant creatures like ourselves. Thus all the complicated structures that we see in the universe might be explained by the no boundary condition for the universe together with the uncertainty principle of quantum mechanics.

The idea that space and time may form a closed surface without boundary also has profound implications for the role of God in the affairs of the universe. With the success of scientific theories in describing events, most people have come to believe that God allows the universe to evolve according to a set of laws and does not intervene in the universe to break these laws. However, the laws do not tell us what the universe should have looked like when it started – it would still be up to God to wind up the clockwork and choose how to start if off. So long as the universe had a beginning, we could suppose it had a creator. But if the universe is really completely self-contained, having no boundary or edge, it would have neither beginning nor end: it would simply be. What place, then, for a creator? ”

10.5 The Bible

We selected some verses from the Bible (new revised standard version) where the issue of mathematical wisdom is treated. In the biblical Apocrypha we find for example:

Wisdom 11,20

Punishment of the Wicked

20 Even apart from these, people could fall at a single breath when pursued by justice and scattered by the breath of your power. But you have arranged all things by measure and

number and weight.

Sirach 1,1-10

In Praise of Wisdom

1 All wisdom is from the Lord, and with him it remains forever.

2 The sand of the sea, the drops of rain, and the days of eternity – who can count them?

3 The height of heaven, the breadth of the earth, the abyss, and wisdom – who can search them out?

4 Wisdom was created before all other things, and prudent understanding from eternity.

5 The root of wisdom – to whom has it been revealed? Her subtleties – who knows them?

8 There is but one who is wise, greatly to be feared, seated upon his throne – the Lord.

9 It is he who created her; he saw her and took her measure; he poured her out upon all his works,

10 upon all the living according to his gift; he lavished her upon those who love him.

In the old testament we find the not explicit statement that $\pi = 3$:

Chronicles second book 4,2

2 Then he made the molten sea; it was round, ten cubits from rim to rim, and five cubits high. A line of thirty cubits would encircle it completely.

In the old testament we find God as the geometer drawing the Universe:

Proverbs 8,27

Wisdom's Part in Creation

27 When he established the heavens, I was there, when he drew a circle on the face of the deep

In philosophy, the problem of the creator of God is the controversy regarding the hypothetical cause responsible for the existence of God, on the assumption that God exists. It contests the proposition that the universe cannot exist without a creator by asserting that the creator of the Universe must have the same restrictions. This, in turn, may lead to a problem of infinite regress wherein each new presumed creator of a creator is itself presumed to have its own creator. A common challenge to theistic propositions of a creator deity as a necessary first-cause explanation for the universe is the question: "Who created God?"

Some faith traditions have such an element as part of their doctrine. Jainism posits that the universe is eternal and has always existed. Isma'ilism rejects the idea of God as the first cause, due to the doctrine of God's incomparability and source of any existence including abstract objects.

In this respect the belief that God became the Universe is a theological doctrine that has been developed several times historically, and holds that the creator of the universe actually became the universe. Historically, for versions of this theory where God has ceased to exist or to act as a separate and conscious entity, some have used the term pandeism, which combines aspects of pantheism and deism [2], to refer to such a theology. A similar concept is panentheism, which has the creator become the universe only in part, but remain in some other part transcendent to it, as well. Hindu texts like the "Mandukya Upanishad" speak of the undivided one which became the universe.

From Section 10.4 we can say that Hawking's discoveries speak only to the nature of God, not to its existence.

The "Book of Wisdom", or the "Wisdom of Solomon", is a Jewish work written in Greek and most likely composed in Alexandria, Egypt. Generally dated to the mid-first century BC, the central theme of the work is 'Wisdom' itself, appearing under two principal aspects. In its relation to man, Wisdom is the perfection of knowledge of the righteous as a gift from God showing itself in action. In direct relation to God, Wisdom is with God from all eternity. It is one of the seven Sapiential or wisdom books in the Septuagint, the others being Psalms, Proverbs, Ecclesiastes, Song of Songs (Song of Solomon), Job, and Sirach. It is included in the deuterocanonical books by the Catholic Church and the anagignoskomena (meaning "those which are to be read") of the Eastern Orthodox Church. Most Protestants consider it part of the Apocrypha. In the "Book of Wisdom" we find for example:

Wisdom 11,21-12,2

God Is Powerful and Merciful

21 For it is always in your power to show great strength, and who can withstand the might of your arm?

22 Because the whole world before you is like a speck that tips the scales, and like a drop of morning dew that falls on the ground.

23 But you are merciful to all, for you can do all things, and you overlook people's sins, so that they may repent.

24 For you love all things that exist, and detest none of the things that you have made, for you would not have made anything if you had hated it.

25 How would anything have endured if you had not willed it? Or how would anything not called forth by you have been preserved?

[2] Pandeism: This is the belief that God created the universe, is now one with it, and so, is no longer a separate conscious entity. This is a combination of pantheism (God is identical to the universe) and deism (God created the universe and then withdrew Himself).

26 You spare all things, for they are yours, O Lord, you who love the living. For your immortal spirit is in all things.

2 Therefore you correct little by little those who trespass, and you remind and warn them of the things through which they sin, so that they may be freed from wickedness and put their trust in you, O Lord.

In his autobiography [4] Nikola Tesla writes:

“Peace can only come as a natural consequence of universal enlightenment and merging of races, and we are still far from this blissful realisation, because few indeed, will admit the reality – that God made man in His image – in which case all earth men are alike. There is in fact but one race, of many colours. Christ is but one person, yet he is of all people, so why do some people think themselves better than some other people?

[...] I have never had the faintest reason since to change my views on psychical and spiritual phenomena, for which there is no foundation. The belief in these is the natural outgrowth of intellectual development. Religious dogmas are no longer accepted in their orthodox meaning, but every individual clings to faith in a supreme power of some kind. We all must have an ideal to govern our conduct and insure contentment, but it is immaterial whether it be one of creed, art, science, or anything else, so long as it fulfils the function of a dematerialising force. It is essential to the peaceful existence of humanity as a whole that one common conception should prevail. [...] I have failed to obtain any evidence in support of the contentions of psychologists and spiritualists, [...] ”

List of Figures

Bibliography

[1] Giorgio Parisi. *In un volo di storni. Le meraviglie dei sistemi complessi.* Rizzoli, 2021.

[2] Galileo Galilei. *Dialogo.* 1632.

[3] Lucio Russo. *La rivoluzione dimenticata: il pensiero scientifico greco e la scienza moderna.* Feltrinelli, 2001.

[4] Nikola Tesla. *The Strange Life Of Nikola Tesla.* John R. H. Penner, 1995.

[5] Thomas S. Kuhn. *The structure of scientific revolutions.* University of Chicago Press, 1962.

[6] Nicolaus Copernicus. *De revolutionibus orbium coelestium.* 1543.

[7] Isaac Newton. *Philosophiae Naturalis Principia Mathematica.* 1687.

[8] Paolo Ruffini. *Teoria generale delle equazioni, in cui si dimostra impossibile la soluzione algebraica delle equazioni generali di grado superiore al quarto.* Stamperia di S. Tommaso d'Aquino, 1799.

[9] Dante Alighieri. *La Divina Commedia.* 1320.

[10] Eugene Wigner. The Unreasonable Effectiveness of Mathematics in the Natural Sciences. *Communications on Pure and Applied Mathematics*, 13:1, 1960.

[11] M. H. Kalos and P. A. Whitlock. *Monte Carlo Methods.* Wiley-Vch Verlag GmbH & Co., Germany, 2008.

[12] Paul Dirac. A Theory of Electrons and Protons. *Proceedings of the Royal Society A.*, 126:360, 1930.

[13] Emilio Segrè and Owen Chamberlain, 1959. The Nobel Prize in Physics 1959.

[14] Gerald Feinberg. Possibility of faster-than-light particles. *Phys. Rev.*, 159:1089, 1967.

[15] G. Benford, D. Book, and W. Newcomb. The tachyonic antitelephone. *Phys. Rev. D*, 2:263, 1970.

[16] P. C. W. Davies. Thermodynamics of Black Holes. *Reports on Progress in Physics*, 41:1313, 1978.

[17] Colin Montgomery, Wayne Orchiston, and Ian Whittingham. Michell, Laplace and the origin of the black hole concept. *Journal of Astronomical History and Heritage*, 12:90, 2009.

[18] B. Louise Webster and Paul Murdin. Cygnus X-1—a Spectroscopic Binary with a Heavy Companion? *Nature*, 235:37, 1972.

[19] C. T. Bolton. Identification of Cygnus X-1 with HDE 226868. *Nature*, 235:271, 1972.

[20] D. Clery. Black holes caught in the act of swallowing stars. *Science*, 367:495, 2020.

[21] B. P. Abbott et al. Observation of Gravitational Waves from a Binary Black Hole Merger. *Phys. Rev. Lett.*, 116:061102, 2016.

[22] Event Horizon Telescope. First M87 Event Horizon Telescope Results. I. The Shadow of the Supermassive Black Hole. *The Astrophysical Journal*, 875:L1, 2019.

[23] Katherine L. Bouman, Michael D. Johnson, Daniel Zoran, Vincent L. Fish, Sheperd S. Doeleman, and William T. Freeman. Computational Imaging for VLBI Image Reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 913, 2016.

[24] L. J. Oldham and M. W. Auger. Galaxy structure from multiple tracers – II. M87 from parsec to megaparsec scales. *Monthly Notices of the Royal Astronomical Society*, 457:421, 2016.

[25] Stuart L. Shapiro and Saul A. Teukolsky. *Black Holes, White Dwarfs and Neutron Stars, the physics of compact objects*. John Wiley & Sons, Inc., 1983.

[26] Charles W. Misner, Kip. S. Thorne, and John Archibald Wheeler. *Gravitation*. W. H. Freeman and Company, 1970.

[27] Alan P. Lightman, William H. Press, Richard H. Price, and Saul A. Teukolsky. *Problem book in relativity and gravitation*. Princeton University Press, 1975.

[28] J. R. Oppenheimer and G. M. Volkoff. On Massive Neutron Cores. *Phys. Rev.*, 55:374, 1939.

[29] R. Penrose. Gravitational Collapse: The Role of General Relativity. *General Relativity and Gravitation*, 34:1141, 2002.

[30] B. J. Carr. Primordial Black Holes: Do They Exist and Are They Useful? In H. Suzuki, J. Yokoyama, Y. Suto, and K. Sato, editors, *Inflating Horizon of Particle Astrophysics and Cosmology*. Universal Academy Press, 2005.

[31] S. B. Giddings and S. Thomas. High energy colliders as black hole factories: The end of short distance physics. *Phys. Rev. D*, 65:056010, 2002.

[32] T. Harada. Is there a black hole minimum mass? *Phys. Rev. D*, 74:084004, 2006.

[33] N. Arkani-Hamed, S. Dimopoulos, and G. Dvali. The hierarchy problem and new dimensions at a millimeter. *Physics Letters B*, 429:263, 1998.

[34] Heino Falcke, Fulvio Melia, and Eric Agol. Viewing the Shadow of the Black Hole at the Galactic Center. *The Astrophysical Journal*, 528:L13, 2000.

[35] Event Horizon Telescope Collaboration. First M87 Event Horizon Telescope Results. VII. Polarization of the Ring. *The Astrophysical Journal*, 910:L12, 2021.

[36] M. D. Johnson, V. L. Fish, S. S. Doeleman, D. P. Marrone, R. L. Plambeck, J. F. C. Wardle, K. Akiyama, K. Asada, and C. Beaudoin. Resolved magnetic-field structure and variability near the event horizon of Sagittarius A*. *Science*, 350:1242, 2015.

[37] B. P. Abbott et al. Properties of the binary black hole merger GW150914. *Phys. Rev. Lett.*, 116:241102, 2016.

[38] V. Cardoso, E. Franzin, and P. Pani. Is the gravitational-wave ringdown a probe of the event horizon? *Phys. Rev. Lett.*, 116:171101, 2016.

[39] B. P. Abbott et al. Tests of general relativity with GW150914. *Phys. Rev. Lett.*, 116:221101, 2016.

[40] B. P. Abbott et al. Astrophysical Implications of the Binary Black Hole Merger GW150914. *Astrophys. J. Lett.*, 818:L22, 2016.

[41] A. Einstein. Näherungsweise Integration der Feldgleichungen der Gravitation. *Sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften Berlin*, 1:688, 1916.

[42] A. Einstein. Über Gravitationswellen. *Sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften Berlin*, 1:154, 1918.

[43] A. Einstein and N. Rosen. On gravitational waves. *Journal of the Franklin Institute*, 223:43, 1937.

[44] Thomas W. Baumgarte and Stuart L. Shapiro. *Numerical Relativity Solving Einstein's Equations on the Computer.* Cambridge University Press, 2010.

[45] F. Pretorius. Evolution of Binary Black-Hole Spacetimes. *Phys. Rev. Lett.*, 95:121101, 2005.

[46] M. Campanelli, C. O. Lousto, P. Marronetti, and Y. Zlochower. Accurate Evolutions of Orbiting Black-Hole Binaries without Excision. *Phys. Rev. Lett.*, 96:111101, 2006.

[47] J. G. Baker, J. Centrella, D.-I. Choi, M. Koppitz, and J. van Meter. Gravitational-Wave Extraction from an Inspiraling Configuration of Merging Black Holes. *Phys. Rev. Lett.*, 96:111102, 2006.

[48] R. Abbott et al. Observation of Gravitational Waves from Two Neutron Star–Black Hole Coalescences. *The Astrophysical Journal Letters*, 915:L5, 2021.

[49] Michele Maggiore. *A Modern Introduction to Quantum Field Theory.* Oxford University Press, 2005.

[50] N. H. March and M. P. Tosi. *Coulomb Liquids.* Academic Press, 1984.

[51] CMS Collaboration. Constraints on anomalous Higgs boson couplings using production and decay information in the four-lepton final state. *Physics Letters B*, 775:1, 2017.

[52] S. K. Lamoreaux. Demonstration of the Casimir Force in the 0.6 to 6 μm Range. *Phys. Rev. Lett.*, 78:5, 1997.

[53] C. Genet, F. Intravaia, A. Lambrecht, and S. Reynaud. Electromagnetic vacuum fluctuations, Casimir and Van der Waals forces. *Annales de la Fondation Louis de Broglie*, 29:311, 2004.

[54] A. Lambrecht. The Casimir effect: a force from nothing. *Physics World*, 2002.

[55] R. Jaffe. Casimir effect and the quantum vacuum. *Phys. Rev. D*, 72:021301, 2005.

[56] I. E. Dzyaloshinskii, E. M. Lifshitz, and L. P. Pitaevskii. The general theory of van der Waals forces. *Advances in Physics*, 1961.

[57] T. H. Boyer. Quantum Electromagnetic Zero-Point Energy of a Conducting Spherical Shell and the Casimir Model for a Charged Particle. *Phys. Rev.*, 174:1764, 1968.

[58] K. Sanderson. Quantum force gets repulsive. *Nature*, 2009.

[59] Z. Z. Du, H. M. Liu, Y. L. Xie, Q. H. Wang, and J.-M. Liu. Spin Casimir effect in noncollinear quantum antiferromagnets: Torque equilibrium spin wave approach. *Phys. Rev. B*, 92:214409, 2015.

[60] S. E. Rugh and H. Zinkernagel. The quantum vacuum and the cosmological constant problem. In *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, volume 33, page 663. 2002.

[61] H. B. G. Casimir. On the attraction between two perfectly conducting plates. *Proc. Kon. Ned. Akad. Wet.*, 51:793, 1948.

[62] M. J. Sparnaay. Attractive Forces between Flat Plates. *Nature*, 180:334, 1957.

[63] M. J. Sparnaay. Measurements of attractive forces between flat plates. *Physica*, 24:751, 1958.

[64] G. Bressi, G. Carugno, R. Onofrio, and G. Ruoso. Measurement of the Casimir Force between Parallel Metallic Surfaces. *Phys. Rev. Lett.*, 88, 2002.

[65] J. Zao et al. Casimir forces on a silicon micromechanical chip. *Nature Communications*, 4:1845, 2013.

[66] T. Lu et al. Measurement of non-monotonic Casimir forces between silicon nanostructures. *Nature Photonics*, 11:97, 2017.

[67] M. Wang et al. Strong geometry dependence of the Casimir force between interpenetrated rectangular gratings. *Nature Communications*, 12:600, 2021.

[68] Y. Aharonov and D. Bohm. Significance of electromagnetic potentials in quantum theory. *Phys. Rev.*, 115:485, 1959.

[69] H. Batelaan and A. Tonomura. The Aharonov–Bohm effects: Variations on a Subtle Theme. *Physics Today*, 62:38, 2009.

[70] E. Sjöqvist. Locality and topology in the molecular Aharonov–Bohm effect. *Phys. Rev. Lett.*, 89:210401, 2014.

[71] W. Ehrenberg and R. E. Siday. The Refractive Index in Electron Optics and the Principles of Dynamics. *Proceedings of the Physical Society B*, 62:8, 1949.

[72] Y. Aharonov and D. Bohm. Further Considerations on Electromagnetic Potentials in the Quantum Theory. *Phys. Rev.*, 123:1511, 1961.

[73] M. Peshkin and A. Tonomura. *The Aharonov–Bohm effect.* Springer-Verlag, 1989.

[74] F. London. On the Problem of the Molecular Theory of Superconductivity. *Phys. Rev.*, 74:562, 1948.

[75] Richard Feynman, Robert Leighton, and Matthew Sands. *The Feynman Lectures on Physics*. California Institute of Technology, 1963.

[76] R. G. Chambers. Shift of an Electron Interference Pattern by Enclosed Magnetic Flux. *Phys. Rev. Lett.*, 5:3, 1960.

[77] S. Popescu. Dynamical quantum non-locality. *Nature Physics*, 6:151, 2010.

[78] S. Olariu and S. Popescu. The quantum effects of electromagnetic fluxes. *Rev. Mod. Phys.*, 57:339, 1985.

[79] S. M. Roy. Condition for Nonexistence of Aharonov-Bohm Effect. *Phys. Rev. Lett.*, 44:111, 1980.

[80] A. Tonomura et al. Evidence for Aharonov-Bohm effect with magnetic field completely shielded from electron wave. *Phys. Rev. Lett.*, 56:792, 1986.

[81] R. A. Webb, S. Washburn, C. P. Umbach, and R. B. Laibowitz. Observation of h/e Aharonov–Bohm Oscillations in Normal-Metal Rings. *1985*, 54:2696, Phys. Rev. Lett.

[82] C. Schönenberger et al. Aharonov–Bohm oscillations in carbon nanotubes. *Nature*, 397:673, 1999.

[83] D. Levine and P. Steinhardt. Quasicrystals: A New Class of Ordered Structures. *Phys. Rev. Lett.*, 53:2477, 1984.

[84] A. L. Mackay. De Nive Quinquangula. *Krystallografiya*, 26:910, 1981.

[85] A. L. Mackay. Crystallography and the Penrose Pattern. *Physica A*, 114:609, 1982.

[86] Erwin Schrödinger. *What Is Life? The Physical Aspect of the Living Cell*. Cambridge University Press, 1944.

[87] L. Bindi, P. J. Steinhardt, N. Yao, and P. J. Lu. Natural Quasicrystals. *Science*, 324:1306, 2009.

[88] D. Shechtman, I. Blech, D. Gratias, and J. Cahn. Metallic Phase with Long-Range Orientational Order and No Translational Symmetry. *Phys. Rev. Lett.*, 53:1951, 1984.

[89] Luca Pacioli. *Divina proportione*. Paganini (Venice), 1509.

[90] Douglas R. Hofstadter. *Gödel, Escher, Bach: An Eternal Golden Braid*. Penguin, 1980.

[91] Bertrand Russell. *The Study of Mathematics*. Mysticism and Logic: And Other Essays. Longman, 1919. pag. 60.

[92] Susantha Goonatilake. *Toward a Global Science*. Indiana University Press, 1998.

[93] Edwin Abbott Abbott. *Flatland: A Romance of Many Dimensions*. Seeley & Co., 1884.

[94] R. Fantoni, B. Jancovici, and G. Téllez. Pressures for a one-component plasma on a pseudosphere. *J. Stat. Phys.*, 112:27, 2003.

[95] Ern o Lendvai. *Béla Bartók: An Analysis of His Music*. Kahn and Averill, 1971.

[96] Peter F. Smith. *The Dynamics of Delight: Architecture and Aesthetics.* Routledge, 2003.

[97] Howat Roy. *Debussy in Proportion: A Musical Analysis.* Cambridge University Press, 1983.

[98] Sylvestre Loïc and Marco Costa. The Mathematical Architecture of Bach's The Art of Fugue. *Il Saggiatore musicale*, 17:175, 2011.

[99] Paolo Zellini. *La ribellione del numero.* Adelphi, 1997.

[100] Mario Livio. *The Golden Ratio: The Story of Phi, the World's Most Astonishing Number.* Broadway Books, 2003.

[101] Stephen Hawking. *God Created the Integers: The Mathematical Breakthroughs That Changed History.* Running Press, 2005.

[102] Leon M. Lederman. *The God Particle.* Bantam Doubleday Dell, 1993.

[103] Riccardo Fantoni. *The Janus Fluid.* SpringerBriefs in Physics. Springer, 2013.

[104] C. Casagrande and M. Veyssie. . *C. R. Acad. Sci. (Paris)*, II-306:1423, 1988.

[105] S. Jiang, Q. Chen, M. Tripathy, E. Luijten, K. S. Schweizer, and S. Granick. Janus particle synthesis and assembly. *Adv. Mater.*, 22:1060, 2010.

[106] S. C. Glotzer and M. J. Solomon. Anisotropy of building blocks and their assembly into complex structures. *Nature Materials*, 6:557, 2007.

[107] P.-G. de Gennes. Soft matter. *Rev. Mod. Phys.*, 64:645, 1992.

[108] Paolo Zellini. *La matematica degli dèi e gli algoritmi degli uomini.* Adelphi, 2016.

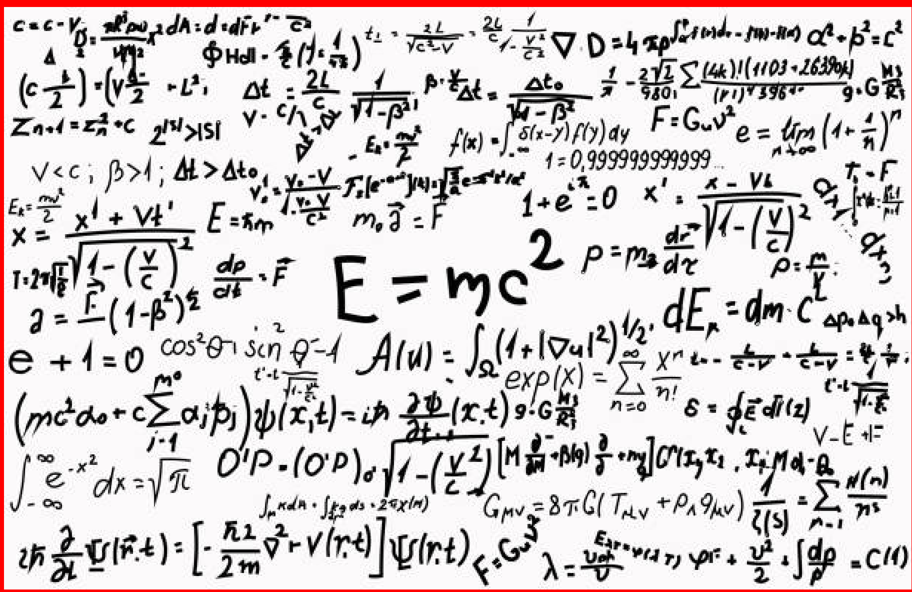
[109] Paolo Zellini. *Breve storia dell'infinito.* Adelphi, 1993.

[110] Benedikt Paul Göcke and Christian Tapp, editors. *The Infinity of God New Perspectives in Theology and Philosophy.* University of Notre Dame Pess, 2018.

[111] Ivar Ekeland. *The Broken Dice, and Other Mathematical Tales of Chance.* University of Chicago Press, 1993.

[112] Stephen Hawking. *A Brief History of Time: From the Big Bang to Black Holes.* Bantam Dell Publishing Group, 1988.

Index


$E = mc^2$