BOOK I: The First Reflection

THE PROBLEM OF UNIFICATION
First Addition
Table of Contents

Chapter 1 .......................................................................................................................... 6

*The Status of Unification* ................................................................................................. 6

1.0 Introduction .................................................................................................................. 6

1.1 Physics: The Struggle for Unification ........................................................................ 10

  1.1.2 A Rigorous Approach to Sausage Making ......................................................... 12

  1.1.3 A “Theory of Everything”? .................................................................................. 13

  1.1.4 A *Theory of Everything*: Is it Physics or Speculation? .................................... 13

1.2 The Destiny: Will There be Unification or Not? ...................................................... 18

  1.2.1 The First Future - Plurality: the Death of Reductionism ..................................... 18

  1.2.2 The Second Future: A Unified Field Theory ..................................................... 19

    1.2.2.1 The Rise of String Theory ............................................................................ 25

    1.2.2.1.1 The Theoretical Advantages of String Theory .................................. 25

    1.2.2.1.2 Lingering Doubts about String Theory .................................................. 26

  1.2.3 The Third Future: A New, More Fundamental Theory ...................................... 28

Chapter 2 .......................................................................................................................... 31

*The Struggle for Unification and the Emergence of Quantum Mechanics* .................. 31

2.0 Introduction .................................................................................................................. 31

2.1 Einstein’s Attempt at Unification .............................................................................. 31

2.2 The Emergence of the Quantum Theories ................................................................. 35

  2.2.1 The Photoelectric Effect ..................................................................................... 36

  2.2.2 The Discovery of Matter Waves .......................................................................... 36

  2.2.3 Signs of a Brewing Controversy ......................................................................... 38

  2.2.4 The Development of the Atomic Model ............................................................. 39

    2.2.4.1 The “Plum Pudding” Model of the Atom ...................................................... 39

    2.2.4.2 The Rutherford Model of the Atom .............................................................. 40

    2.2.4.3 The Bohr Model of the Atom ....................................................................... 40

  2.2.5 The New Physics .................................................................................................. 41

    2.2.5.1 Non-Relativistic Quantum Mechanics ....................................................... 42

      2.2.5.1.1 The Wave Function ............................................................................... 42

      2.2.5.1.2 The Role of the Operator in Quantum Mechanics ............................... 43

      2.2.5.1.3 The Wave Equation .............................................................................. 45
Chapter 3.............................................................................................................. 59

The Standard Model and Beyond ........................................................................ 59

3.0 Introduction ...................................................................................................... 59

3.1 Group Theory: The Road to Understanding Elementary Particles .............. 60
   3.1.1 The Symmetries of Nature ........................................................................ 66

3.2 The Standard Model of Particle Physics .......................................................... 68
   3.2.1 The Characteristics of the Standard Model ................................................ 71
   3.2.2 Experimental Inputs to the Standard Model ............................................. 72
   3.2.3 The Theoretical Part of the Standard Model ............................................. 75
   3.2.4 Summary of the Standard Model ............................................................... 78

3.3 Beyond the Standard Model ............................................................................ 79
   3.3.1. Einstein’s Failure ..................................................................................... 79
   3.3.2 Supersymmetry ......................................................................................... 80
   3.3.3 Technicolor ............................................................................................... 81
   3.3.4 Grand Unified Theories (GUTS) ............................................................. 82
   3.3.5 Preons ...................................................................................................... 82
   3.3.6 String Theory .......................................................................................... 83

Chapter 4 ................................................................................................................ 91

The Practical Challenges of Unification ............................................................... 91

4.0 Introduction ...................................................................................................... 91

4.1 Quantum vs. Classical Theories ...................................................................... 92
   4.1.1 The Indivisibility of Quantum Processes ................................................ 92
   4.1.2 Incomplete Determinism in Quantum Laws ............................................ 92
   4.1.3 The Holistic Nature of Quantum Processes ............................................. 93
      4.1.3.1 Wave-Particle Duality ....................................................................... 94
      4.1.3.2 Schrödinger’s Cat ............................................................................. 98
   4.1.4 The Correspondence Principle ................................................................. 99
4.2 The Copenhagen Interpretation ................................................................. 100

4.3 The Measurement Problem in Quantum Mechanics .................................. 101
   4.3.1 The Many Worlds Interpretation ....................................................... 104
   4.3.2 Environmental Decoherence ............................................................ 104

4.4 The EPR Paradox and Bell’s Theorem ................................................. 108
   4.4.1 The EPR Paradox ............................................................................. 108
      4.4.1.1 The Stern-Gerlach Experiments ............................................... 108
   4.4.2 The EPR Argument ......................................................................... 111
   4.4.3 Bell’s Inequality .............................................................................. 113
   4.4.4 Bell’s Theorem .............................................................................. 116

Chapter 5 .............................................................................................................. 121

The Philosophical Challenges of Unification ..................................................... 121

5.0 Introduction ................................................................................................. 121

5.1 Two Philosophical Questions .................................................................... 121
   5.1.1 René Descartes (1596 – 1650) ............................................................. 123
   5.1.2 Benedict DeSpinoza (1632 – 1677) ..................................................... 124
   5.1.3 David Hume (1711 – 1776) ................................................................. 128
   5.1.4 John Locke (1632 – 1704) ................................................................. 128
   5.1.5 George Berkeley (1685 – 1753) ........................................................... 130
   5.1.6 Immanuel Kant (1724–1804) .............................................................. 132
   5.1.7 Monistic Idealism ............................................................................. 136

5.2 Conclusion .................................................................................................... 137
Chapter 1

The Status of Unification

Today scientists describe the universe in terms of two basic partial theories—the general theory of relativity and quantum mechanics. They are the great intellectual achievements of the first half of this century. The general theory of relativity describes the force of gravity and the large-scale structure of the universe, that is, the structure on scales from only a few miles to as large as a million million million million (1 with twenty-four zeros after it) miles, the size of the observable universe. Quantum mechanics, on the other hand, deals with phenomena on extremely small scales, such as a millionth of a millionth of an inch. Unfortunately, however, these two theories are known to be inconsistent with each other—they cannot both be correct.

— Stephen W. Hawking

1.0 Introduction

Since the theory of relativity and the quantum theory cannot both be correct, solving the dilemma of two incompatible competing theories that describe the same Universe should boil down to uncovering which of the two theories is correct. But, it is hardly that simple, since both theories make spectacularly successful predictions in their respective domain of applicability—relativity on the large scale, quantum mechanics on the small scale. Replacing one, the other or both theories requires that any replacement theory make just as many accurate predictions as the combined predictions of its two predecessors. This is why most attempts at developing a unified theory involve an approach that combines the two theories into one. The name for this “yet to be discovered” unified theory is ‘quantum gravity’.

A successful theory of quantum gravity requires that the apparent incompatibilities between the theory of relativity and quantum mechanics be reconciled in some fashion. Unfortunately, all the current approaches expose a myriad of difficult to resolve issues, some of which reach beyond the domain of traditional physics. Having access to an abundance of empirical data, physicists have become adept at recognizing patterns in the data and, from those patterns, creating fruitful conjectures for use as a basis for a theory. Comparing the conjectures to the physical evidence gauges how well the theoretical suppositions fit the observed facts. But pursuing a quantum theory of gravity carries physics into territories where little or no data is available. The absence of available data provides scanty few methods of verification. The question is: if not on empirical grounds, then what justifies accepting one case for unification over that of another [1]?

While the lack of data has practical implications, more problematic are the conceptual issues. For instance, reconciling the quantum concept of ‘quantizing’ fields with the
relativistic idea of ‘general covariance’ has found no generally accepted solution. In
general relativity, space-time is the entire substance and is described as a dynamically
smooth fabric; it can be molded, folded and stretched, but not torn. The quantization of
space-time involves turning the space-time smoothness into discrete quantities; put in
simpler terms, turning space-time into particles. In quantum field theory, each field is
associated with a different particle, but these fields are entirely dissimilar to the
gravitational field. Quantum particles propagate in a fixed space-time and are not
considered extensions of the dynamical fabric of space-time. On the other hand, the
equivalence of gravitational and inertial mass is the fundamental principle within general
relativity that allows gravity to be represented as a property of space-time itself, rather
than as a field propagating within space-time. Quantizing gravity subjects some of the
properties of space-time to quantum fluctuations, which is the same thing as saying that
space-time oscillates \[2\]. The best ways of describing the quantization of fields is by
specifying a specific and fixed space-time background (one that does not oscillate) and
then to describe quantization against that background. This makes it easier to formulate
the concepts related to the fields such as its energy, momentum, and angular
momentum \[3\]. But the space-time metric is dynamic, determines the space-time
genometry and dictates the propagation of all physical fields (including itself). Because
there is no background geometry, the concepts of energy and momentum are difficult to
define. It’s a “chicken and egg” problem. Since there is no fixed space-time, defining
basic notions, such as causality, time and evolution, become problematic without first
specifying a space-time geometry \[3\]. According to general relativity, the laws of
physics remain unchanged with a change in space-time geometry. This is the essence
behind the idea of ‘general covariance’. But in quantum physics a change in space-time
geometry would mean introducing an entirely new set of physical laws, since each
particle requires its own field. And it is hard to see how these two diametrically opposed
field concepts can be reconciled.

Additionally, the theory of relativity and the quantum theory treat time very differently. In
quantum mechanics, mathematically, time is a parameter; every quantum observation is
a function of time. Observing something means observing it at a particular moment.
While the things that can be observed change with time – a kind of quantum evolution,
the evolution itself is not observable. For those familiar with the mathematics of
quantum mechanics, this “quantum evolution” is described by a wave function. But the
evolution of the wave function does not describe an actual physical event i.e. it is not
observable. Once a measurement or observation is taken, the wave function ceases to
be a factor and only a real outcome is observed. In a sense, when a measurement is
taken, time ceases and what’s left is whatever was measured. By contrast, general
relativity’s ‘time’ must be measurable, since it describes, in part, the warping of space-
time. In this case, time is not a parameter, but a dimension with a direction just like the
other three spatial dimensions.

There are other conceptual issues. Quantum theories are partly indeterministic while
general relativity is entirely deterministic. This fact, along with the issues just discussed,
makes it difficult to see how a quantum theory of gravity could be formulated.
Nevertheless, there have been many attempts at rectifying the incompatibilities between the two theories. In a nutshell, the problem boils down to this: the theory of relativity is described in terms of a four-dimensional space-time. On the other hand, the atomic theory of weak interactions is described by a two dimensional spinor space, while the atomic theory of strong interactions is described by a three dimensional color space. From a unification standpoint, the difference in the number of dimensions in each space is problematic. How can three theories be combined into one given that the number of dimensions in each theory is different?

A most promising approach was presented by Yang and Mills in 1954 involving the concept of a ‘gauge connection’. They, among others, put forth the idea of combining seemingly disparate spaces to form a new space. The formulation normally involves creating a base space, call it ‘\( M \)’ (this is usually space-time in most applications) then bundling that base space together with another space, call it ‘\( V \)’ (in current terminology this space is called a ‘fiber’) to form a new space (\( M \times V \)) called the ‘total space’. The bundling can be accomplished in a local manner. The best way to think about this as a local concept is to imagine that there is a copy of \( V \) (the fiber) that sits above each point of \( M \) (base space). The advantage of bundling is that the total space can exhibit very different characteristics than either \( M \) or \( V \) considered independently. Bundling can be accomplished in innumerable ways, which leads to an abundance of theories that can be created using this technique. What Yang and Mills, among others, recognized was that requiring the fiber to possess a specific symmetry gave the bundled space special properties. It also helped in reducing the number of possible theories that could be created [5].

Bundling became wildly successful and is at the heart of the mathematics (quantum field theory) that supports the standard model of particle physics. Nevertheless, Yang-Mills theory was not an immediate success, since the particles described by the theory all had zero mass and it was known that some elementary particles carried mass. Also, as originally conceived, it was a classical theory. But in 1964, Peter Higgs proposed a mechanism that described how the symmetries in electroweak interactions were broken. This explained the origin of the masses of elementary particles in general and of the ‘\( W \)’ and ‘\( Z \)’ bosons in particular. The Higgs mechanism, which had several inventors besides Higgs, predicts the existence of a new particle: the Higgs boson [6].

Solving the mass problem breathed new life into Yang-Mills theory. But, there was one additional hurdle. When some of the numbers were totaled up, the sums diverged, which meant that their summations added to infinity. Sensibly, applying mathematical techniques to explain how the world works should result in finite answers. And when calculations result in infinities, this is paramount to a disaster. However, techniques were developed where, if certain conditions existed, the infinities could be tamed or ignored by adjusting the parameters of the theory. The process of removing or ignoring infinities is called ‘renormalization’. In 1971, Gerard ‘t Hooft demonstrated that theories with a “spontaneously broken” symmetry were renormalizable [5]. This discovery guaranteed Yang-Mills theory a place in physics for the foreseeable future.
The enthusiasm for Yang-Mills theory led to attempts at constructing a quantum theory of gravity by employing bundling techniques. Unfortunately, the theory created using the approach was shown to be non-renormalizable. As per the Yang-Mills general approach, space-time was broken up into two spaces; one that represented a flat space-time, the other constructed so that perturbation expansion techniques so successful in the quantum theory could be applied. But introducing a background metric was incompatible with the concept of ‘general covariance’ fundamental to general relativity. To appreciate the significance of this, consider the electroweak theory. This theory is perturbatively renormalizable. This means that, although individual terms in the perturbation expansion (summations) may diverge, the infinities are of a specific type; they can be systematically absorbed in the values of the free parameters of the theory. Thus, by renormalizing these parameters, individual terms in the perturbation series can be systematically rendered finite. In general relativity, such a systematic procedure fails. The infinities that arise are genuinely troublesome. Put differently, a quantum theory of gravity approached in this manner acquires a number of undetermined parameters that cannot be absorbed [3].

Further attempts were made, specifically a theory called ‘$N = 8$ supergravity’. It was developed in hopes of taming the ultraviolet infinities. For a number of years, there was a great deal of confidence that supergravity was on the threshold of providing a complete quantum gravity theory. The theory did show a better ultraviolet behavior. But ultimately, it, too, was shown to be non-renormalizable [3].

There were other approaches, such as affixing additional dimensions to a given space (this is usually space-time). In 1921, Theodor Kaluza extended general relativity to a five, rather than a four dimensional space-time. His equations could be separated into further sets of equations, one of which was equivalent to Einstein’s gravitational field equations. The other set was equivalent to Maxwell's equations for the electromagnetic field. The theory could not be successfully extended into the quantum realm, but the idea of extra dimensions persevered.

Another idea was to replace the point particles of the standard model by 1-dimensional extended objects called ‘strings’ and associate particle-like states with the various “modes of excitations” of the string. The theory generated excitement and for the last twenty to thirty years has been the dominate approach for creating a quantum theory of gravity. It includes the spin-1 modes characteristic of the gauge theories associated with atomic interactions and a spin-2 massless excitation, which automatically incorporates the graviton (the fundamental particle thought to carry the gravitational force). In this sense, gravity was already built into the theory! Greene and Schwarz showed that perturbative string theory could be consistent in a certain number of space-time dimensions - 26 for a purely bosonic string and 10 for a superstring [7,8].

Strings are assumed to live in a flat space-time background, where perturbative estimation techniques apply. However, since it includes extended objects rather than point particles, all singularities (the cause of the infinite sums) are eliminated, which dramatically improves the ultraviolet behavior of the theory. It is widely believed that the
perturbation techniques applied to the theory are all finite. In short, the theory has all the ingredients of a ‘theory of everything’, which incorporates a quantum theory of gravity [3].

However, string theory faces important challenges: 1) to date, none of the string theories created so far corresponds to the observed world; 2) it is not certain that the perturbative expansions associated with string theory are, in fact, finite. Gross and Periwal have shown that in the case of bosonic strings, when summed, the series diverges (sums to infinity) and does so uncontrollably. They also gave arguments that the conclusion would not change if superstrings are used instead.

Why is the divergence of a sum regarded as a serious failure of string theory, but accepted in other quantum theories such as quantum electrodynamics (QED)? ‘QED’ is not a complete theory and has the advantage of ignoring many processes that come into play at high energies. It ignores the microstructure of space-time by assuming that space-time can be approximated by a smooth continuum below the Planck scale. On the other hand, a ‘theory of everything’ is not at liberty to ignore anything and must face the Planck problem head on non-perturbatively. It has not been experimentally verified that the extra dimensions and the supersymmetry associated with superstring theories are properties of Nature. Nonetheless, string theory has provided a glimpse of an entirely new vista, the concrete possibility that unification could be brought about by a tightly woven, non-local theory [3].

1.1 Physics: The Struggle for Unification

Modern physics dates from the 17th and 18th centuries beginning with the work of Isaac Newton followed by James Clerk Maxwell’s theory and Michael Faraday’s experimental description of electromagnetism. Extensions and refinements to the work of Newton and Maxwell resulted in non-relativistic classical physics, which succeeded in unifying most of the physics experienced on an everyday basis, everything from how things move, to jet and rocket propulsion, stresses on bridges, buildings and towers, electricity, sound, light and many other phenomena too numerous to mention [9].

Simple Natural laws expressed mathematically more or less describe classical physics. The concept of a particle represents one of the two major modes of describing physical phenomena; the other mode being the wave or vibration. Massive bodies are generally described as particle-like, while light, for example, is described as a wave or, more generally, by the “field” concept [9].

Maxwell’s theory united all previously unrelated observations, experiments and equations of electricity, magnetism and optics into a consistent theory of electromagnetism. His equations demonstrated that electricity, magnetism and light are all manifestations of the electromagnetic field. Maxwell’s laws, like Newton’s, expressed almost entirely in mathematical language, described electromagnetism as a wave-like (field) phenomenon [10,11].
The twentieth century witnessed the discovery of the theory of relativity (our current theory of gravity) along with quantum mechanics. In the process, two new forces (the strong and weak nuclear forces) were discovered. It was also a time that realized the crowning achievement of the era - the standard model of particle physics. The standard model succeeded in bringing together the theories of electromagnetism and the two nuclear forces under one theoretical umbrella. The last remaining step in the unification of physics is to merge gravity, the fourth of the four fundamental forces, with the standard model to create one all-encompassing theory.

During the latter half of the Twentieth Century, optimism for total unification grew with the development of string theory. In the search for a successful theory of quantum gravity, string theory is considered the most likely candidate. In the standard model, particles are described as infinitely small points in space. However, at a specific point, the equations of general relativity produce infinities. String theory, at the fundamental level, overcomes the ‘infinity problem’ by proclaiming that matter consists not of point particles, but rather tiny loops of string. Strings are two dimensional objects that do not collapse to single point. In a sense, strings make space-time fuzzy, removing the need to explain what happens at a specific space-time point. From this beginning, the laws of physics emerge. General relativity, electromagnetism and Yang-Mills gauge theories all appear in a surprising fashion. But string theory comes with unexpected surprises in the form of other ingredients, most strikingly, extra spatial dimensions of the Universe beyond the three that are observed [12]. Moreover, instead of one theory, there seems to be a vast landscape of possible theories — perhaps $10^{500}$ or more.

Interestingly, the emergence of string theory has fractured the professional physics community into a number of feuding ideological factions. There are those who claim that string theory has failed, that it currently makes no testable predictions and unlikely to do so. They criticize the theory as “not really science” arguing that the vast landscape of theories makes it impossible to truly evaluate the theory [13,14]. Conversely, some have adopted a “wait and see” attitude. They openly admit that the evidence for string theory is circumstantial, but harbor hope that the circumstantial evidence will one day bear the fruits of its potential [14]. And finally, a different set of physicists embrace the landscape of theories. But this raises the question: which string theory describes our Universe? Many physicists who endorse the idea of a large landscape of string theories argue that the key to finding the particular theory that governs our Universe can be found by applying the anthropic principle [15].

Roughly speaking, the anthropic principle subscribes to the notion that finely tuned connections exist between the physical parameters used as modeling inputs to a theory and the conditions necessary for life to subsist. This means that the physical parameters are confined to a narrow range of values. As the anthropic argument goes, these connections allow choosing among potential string theories only those that have the parameter values that allow life to exist [16].
1.1.2 A Rigorous Approach to Sausage Making

Theoretical physics entails writing down a number of Natural laws in the form of a set of mathematical equations, then arguing that the equations faithfully represent the observable phenomena under consideration. The term ‘Natural law’ is often misunderstood and it will be useful to review what it generally connotes in a scientific context.

Repeatedly observing similar results under similar observational or experimental conditions implies that Natural laws are eternal truths. Several Natural laws have never been violated in any experiment performed so far. Yet, this is no guarantee that in the future these Natural laws might not be subject to modification. This has been, roughly speaking, the general approach to physical science. Natural laws are deemed the most probable explanations [9].

Science, approached in this manner, more or less embodies the idea of ‘falsification’. Falsification involves determining whether the Natural laws written down in support of a theory faithfully represent the physical phenomena under consideration. A theory, to be considered scientifically acceptable, should make testable predictions. There should be a way through observation and experimentation of showing that the predictions of a theory do not contradict known facts. If there is no way of empirically testing a theory, it is generally considered a speculative conjecture. The upshot of the idea of ‘falsification’ is that a theory can never be proven, but should be capable of being disproven.

To a certain extent, the concept of a Natural law is confounded by the uncertainty principle associated with quantum mechanics. In much of the current theoretical work, conservation laws, the backbone of physics, are routinely violated. However, these violations occur only within the dictates of the uncertainty principle. Ironically, the uncertainty principle can be thought of as the supreme Natural law. All other Natural laws are subject to it. Put differently, no Natural law can be written more exactly than permitted by the uncertainty principle.

This kind of positivist outlook suggests that physicists work in the ultimate of structured cookbook environments, where rigorous methods produce irreproachable results. But physicists are, above all, human beings, who are prone to human frailties and prejudices. The history of physics is replete with instances where debates can, particularly if there is scant observational evidence in support of an argument, evolve into bitter disputes. Personalities and individual prejudices can play large roles and personal clashes have often led to mistakes, dead ends and created impediments to new discoveries. Obstinately refusing to accept new but ultimately correct views have resulted in embarrassing personal confrontations, sometimes contributing to the slow advancement of the subject. Yet through all the sausage making that creating theories sometimes displays, the end product has, for the most part, reflected something fairly close to the truth.
1.1.3 A “Theory of Everything”? 

The goal of complete unification is for physics to be expressed by a single set of consistent and simple Natural laws. In just over a century, remarkable progress towards unification has been achieved in the face of huge challenges. Advancements in sophisticated technologies have made it possible to collect mounds of critical, if disparate, data. Analysis of the data led to new discoveries and generated the necessary enthusiasm and motivation required to explain the mysteries lurking in the data. A century ago discussions and prognostications were limited to explaining how our solar system worked. Today, theories are much bolder, sporting complex sometimes esoteric ideas that portend to explain how the entire Universe works.

Some physicists believe that a “theory of everything” is at hand. In fact, unifying the strong and weak nuclear forces, the electromagnetic force and gravity has been at the forefront of physics for about four decades. Of course, the ‘theory of everything’ is not remotely a theory of everything as the phase is usually understood to mean. The term is largely an invention of the Press. It simply denotes the single theoretical framework that would explain the fundamental interactions of Nature [17].

1.1.4 A ‘Theory of Everything’: Is it Physics or Speculation?

Ironically, the race toward unification gave birth to a number of theories sporting a wide range of speculative ideas. While interesting in their own right, the new theories arrive donning a host of mysterious ingredients [12], are often highly abstract and frequently accompanied by a plethora of meritorious declarations - claims that are often difficult to substantiate, outright dubious or, in some cases, just plain ridiculous. As an example, the current proposed solutions to unifying the quantum theory with the theory of relativity describe physics at energy levels that are experimentally out of reach even for future technological developments. Some physicists find it patently unprofessional to prognosticate in realms where experimental observations are relegated to the sidelines, replaced by conjectures whose advocates point to their esthetic beauty as proof that the conjecture is correct.

The conservative view has always been, to be taken seriously, a theory should contain ideas that are, at least potentially, falsifiable. Interestingly, in the race to find the ultimate theory, the “strictly falsifiable” criterion has increasingly been ignored. Ideally, a good theory should be subjected to experimental verification and obliged to provide new insights into the way Nature actually works [14]. Relatively strict adherence to these tenants assured that prognostications treaded on comparatively safe ground. There was less risk of error and theoretical predictions were based on something more than philosophical speculation.

But physicists complain that much of the current work harbors little resemblance to the characteristics just mentioned. It is common today to find works that are patently unfalsifiable and make no truly insightful predictions, and in some cases, no predictions. While it is debatable whether the “strictly falsifiable” evaluation criterion should be levied on current theoretical work, it begs the question: how are the merits of various unifying
proposals to be determined? On what grounds is a theory to be found plausible, its arguments sound?

Lee Smolin, a brilliant if somewhat disgruntled and a bit of a rebel theoretical physicist and author of “The Trouble with Physics” is greatly concerned with how his profession currently supports research projects in physics. Smolin argues that the current method of funding projects demotes the pursuit of ideas to an exercise in democracy. He contends that the current environment has split the physics community into factions of similarly thinking groups. Part of the job of a physicist as a member of a group is to attract other talented thinkers to join. More group members equate to a higher number of published research papers that reflect the commonly held beliefs of the group. There follows naturally a rapid advancement of the popular idea at the expense of those ideas that lack sufficient representation [14].

For many physicists, including Smolin, the environment is distasteful. From their perspective, there is a danger that the theoretical direction of physics becomes guided by authority, rather than by experiment and observation. For example, the theoretical physics’ departments at most major universities are dominated by string theory supporters, and this, it is argued, creates a “string theory cult” that inherently marginalizes other approaches. Science, like any other field of endeavor, needs criticism. Psychologists have shown that the phenomenon of “groupthink” takes hold in situations where the only people who are allowed a seat at the table are those who think alike. To have a robust intellectual exchange, it’s important to include others who will challenge the popular viewpoint and not just agree with it [16].

The primary method for determining the merits of a proposal is peer review and this, as mentioned above, comes with its own set of issues. Modern theories addressing unification are highly technical, some uniquely so. It is difficult in many cases to find experts with the requisite knowledge necessary to adequately judge the merits of an idea. Physicists often complain that many ideas judged to be worthy by peer reviewers oftentimes merit no such worthiness. On the other hand, some ideas that might be worthy are sometimes rejected or simply ignored.

The problem is exacerbated by the fact that modern theories in physics are temples of extreme abstraction, where few individuals are allowed in. Gone are the days when theories involved forces and masses, rods and pulleys. Concrete mechanical concepts have been replaced by a baffling array of logical, mathematical structures. In a certain sense, modern particle physics is completely incomprehensible. The points, pushes and pulls of the old physics are there somewhere if hidden behind the mathematical concepts of fields, interactions, wave functions and symmetry groups. The modern theoretical physicist has much in common with the post-modern deconstructionist. Both wallow in their own kind of special gobbledygook understandable only to those who are specially trained, but inaccessible to the uninitiated. To many outsiders, modern theoretical physics can seem like a capricious subject where aesthetic beauty trumps scientific fact and increasing abstraction threatens to turn theoretical physics into recreational mathematics [18].
To be fair, while the current situation is not ideal, it is not unlike the epic periods of past
generations, where similar doubts have stunted progress. Even Maxwell’s theory was
slow to be accepted. He tried postulating that little mechanical wheels were responsible
for the swirling effect of electromagnetic phenomena. But ultimately, in the final
publication of his work, the wheels vanished, replaced only by the concept of the
‘electromagnetic field’ and leaving no explanation as to what the source of the field was
or how the force field moved through space [18]. Traditional thinkers tried desperately
to hang on to the idea of the mechanical nature of the field by postulating that it moved
through space in a luminous aether. They clung to the idea that waves of light like
sound or water waves required a medium through which to travel. When carefully
designed experiments failed to uncover any indication that the luminous aether existed,
theories became far less mechanical and more insubstantial. It was the first step away
from advocating a totally mechanized Universe and one step toward a Universe
explained through mathematical abstractions.

The current unification proposals have become so broad in scope suggesting that the
only feasible experimental laboratory available, realistically speaking, is the entire
Universe. This places an ever increasing importance on the science of cosmology.
Cosmology is the discipline that deals with the nature of the Universe as a whole.
Cosmologists seek to understand the origin, evolution, structure, and ultimate fate of the
Universe at large, as well as the natural laws that keep it in order. Modern cosmology is
dominated by the ‘big bang theory’, which brings together observational astronomy and
particle physics [19].

Of all the disciplines within physics, cosmology is the newest and by far the most
tenuous. It is exceedingly difficult, given the enormous distances from the Earth of most
celestial objects, to estimate how far away, how bright, or even identify what the objects
are. As recently as the 1920’s astronomers were unsure whether the Universe was
confined to the Milky Way galaxy or whether it was much larger. Today, it is known to
be much larger, but how much larger remains an open question. The current practice of
astronomy requires exacting technology coupled with ingenious methods of calculation,
where gross statistical estimations are common. The science was dynamically enriched
by the discoveries of the optical and radio telescopes and the photoelectric plate. The
photoelectric plate replaced the notoriously unreliable human eyeball greatly enhancing
the consistency of celestial observations. But comparatively speaking, cosmological
theories, even aided by the astronomer’s propensity for exacting detail, are by far the
most precarious. During the last fifty or so years, within the big bang theory, the
estimated age of the Universe has been modified (by billions of years) no less than
three times, necessitated primarily because new and better methods for estimating
celestial distances were developed [20]. It is indeed remarkable how the slightest
astronomical insight can dramatically affect how the nature of the Universe is viewed.

There is no agreed upon list of criteria or even general guidelines detailing how potential
unification proposals should be evaluated. But there are boundaries and constraints
that potential solutions must work within. There is the reality of the impenetrable.
Technological limitations constrain the extent to which man made laboratory
experiments can act as judge and jury over competing theories. They can help, but there is little hope that new technologies will penetrate and explore the energy scales where the current theories tell at least part of their stories. To expect modern unification theories to be completely or even mostly falsifiable is, for all practical purposes, unreasonable. On the other hand, modern physics has uncovered fundamental laws covering a wide range of physical phenomena through the quantum theory of matter, Einstein’s theory of space-time and the big bang theory of cosmology. And it is imperative that the knowledge gained from these theories be incorporated into any potential unifying proposal.

This offers a big enough challenge. But, there are other major hurdles to overcome. Beyond finding a way of reconciling the general theory of relativity with the quantum theory, the big bang theory implies that the objects in the Universe are accelerating away independently of vantage point—presumably in a manner consistent with the second law of thermodynamics. In a nutshell, the second law implies that the Universe evolves from a state of order to an increasing state of disorder as time moves forward. This process, called ‘entropy’, is thought to be irreversible. Perhaps no law of Nature commands a greater sense of credibility than the second law, which can be summed up in the following quotation:

“The law that entropy always increases holds, I think, the supreme position among the laws of Nature. If someone points out to you that your pet theory of the Universe is in disagreement with Maxwell’s equations — then so much the worse for Maxwell’s equations. If it is found to be contradicted by observation — well, experimentalists do bungle things sometimes. But if your theory is found to be against the second law of thermodynamics, I can give you no hope; there is nothing for it but to collapse in deepest humiliation.”

- Sir Arthur Stanley Eddington

Imagine rolling back the clock to the time of the big bang. According to second law of thermodynamics, at that moment, the Universe must have been incredibly orderly. So orderly, in fact, that any model universe adequate enough to describe the “real” Universe would require input parameters so finely tuned as to be, for all practical purposes, exact. The tiniest variation in the description of those initial conditions would create a model universe vastly different than the one observed. This presents a challenge when describing the role of quantum mechanics in the early Universe. The fundamental principles of quantum mechanics don’t allow the clock to be turned back in the manner suggested by the second law. If the clock was turned back to that moment, quantum mechanics suggests that the state of the Universe at any previous epic would be unpredictable. According to the quantum theory, such exacting conditions at the moment of the big bang would not be possible. It seems unlikely that the amount of exactness required of the big bang model would be consistent with the uncertainty principle. A major outstanding problem in physics, often referred as the ‘vacuum catastrophe’, is that most quantum field theories predict a huge value for the quantum vacuum. It is reasonable to think that the big bang is somehow related to this quantum
vacuum state. A common assumption is that the quantum vacuum is equivalent to the cosmological constant fundamental to modern cosmology [21]. But, if the Universe is described by an effective local quantum field theory down to the Planck scale, the theoretical value of the cosmological constant is of the order $10^{120}$ greater than its measured value. This discrepancy has been called "the worst theoretical prediction in the history of physics!" Making things worse, there is no known natural way of deriving a tiny cosmological constant from particle physics [22].

Moreover, the objects in the Universe appear to be accelerating away from all vantage points. This is happening in a Universe where gravity is the dominate force. Common sense would dictate that gravity should have the opposite effect. How can the Universe be expanding if the physics at these large scales is dominated by gravitational forces, which should act to counter such an expansion? Einstein's general theory of relativity helps somewhat, in this regard, because it is dynamic and does not require the Universe be static. Still, it leaves an open ended description regarding the dynamics of the Universe. It is not clear whether the Universe will expand forever, eventually contract into a big crunch or remain pretty much as it is. It would seem that any serious unification theory should provide a reasonable explanation of this expansion/gravitational effect.

Finally, there is a host of unsolved problems in physics; finding an explanation for dark matter/energy, the black hole information paradox, supersymmetry and many others. Any unification proposal will be judged partly on how reasonably it addresses these issues.

At this point, it appears the recipe for creating an adequate unification theory will involve some confirmation by observation and experiment, consist of clever reasoning and be acceptable to peers. This is about the best that can be expected of any unification proposal. It is certain that proposals will arrive laced with healthy doses of speculation. Hopefully, methods will arise for determining, other than on purely authoritarian grounds, whether or not the ideas contained within those proposals are acceptable.

With any luck, the final unifying theory, if one emerges, will become something more than a myth. A myth is a self-contained belief that offers explanations for everything, but can be neither proven nor disproven. And therein lays a danger. A myth is something everyone agrees upon because it is convenient, not because it is true [18].
1.2 The Destiny: Will There be Unification or Not?

At a crossroads, the quest for a complete unification of physics appears destined for one of three possible future states:

1. The thrust toward unification will be replaced by a plurality of theories. No single underlying concept will be uncovered; there will be no unified theory.

2. The current unification proposals are incomplete, but further work will lead to a satisfactory unified theory of physics (e.g. string theory).

3. Unification will eventually be achieved, but an entirely different approach is required from those currently being pursued.

1.2.1 The First Future - Plurality: the Death of Reductionism

Those who pursue a unified theory of physics are often called ‘reductionists’. A reductionist believes that the Universe is made from the same “stuff”. In this sense, the reductionist is like the monist, who subscribes to the philosophical opinion that reality is a unified whole and is grounded in a single basic substance or principle. Accordingly, there should be one theory that explains how that ‘one substance’ works. The fact there are currently two seemingly incompatible theories that explain how the Universe works is, to a reductionist, unacceptable.

But to a non-reductionist, that there are two incompatible theories is not at all disingenuous. The non-reductionist argues that the essential role played by higher organizing principles in determining new emergent physics is generally overlooked [23]. Furthermore, any single unifying theory is unlikely to provide a complete description of Nature. Those who think along these lines find no compelling reason why the laws of Nature should require the same essential structure and do not share the reductionist view that all phenomena should emerge from a single underlying principle.

Non-reductionists often argue that the essential role played by higher organizing principles in determining emergent behavior “continues to be disavowed by so many physical scientists”, that the situation equates to a poignant comment on the nature of modern science. They point to unpredictable electronic phenomena such as organogels, Kondo insulators and cuprate superconductivity as obvious examples “not discussed in polite company” because these ideas are fundamentally at odds with the reductionist beliefs so central to physics [23]. They accuse their colleagues of acknowledging only the facts that support the reductionist case and argue that this is fundamentally incompatible with science.
Non-reductionist sentiments can be summed up in the following quotation:

“Cosmologists think that cold dark matter is a safe bet because particle physicists can provide it; particle physicists invent theories containing cold dark matter because cosmologists seem to want it. Cosmologists think the Universe must have exactly the critical density, despite observational evidence, because inflation demands it; particle physicists think inflation is a marvelous idea because it predicts a Universe of precisely the critical density, which is what the cosmologists want. The hopes of the cosmologists and the particle physicists have become the same hopes, and their methods have become the same method: on both sides of the joint effort there is absolute reliance on the notion that a single theory will explain everything, and when such a theory comes along it will be instantly recognizable by all. This might be called the messianic movement in fundamental science.”

- David Lindley

In addition to those physicists who flatly reject the reductionist view, there are the ‘instrumentalists’ who adopt an agnostic attitude toward reductionism. From their perspective physical theories should primarily make credible predictions. Theories should not be expected to uncover a reality in any deeper sense. Those ‘deeper’ questions should be left to the philosophers and the churches [1]. Since the theory of relativity and quantum mechanics make accurate predictions in the realms of their applicability, instrumentalists see no reason to pursue a quantum theory of gravity and regard such an endeavor as needless speculation.

But the quantum theory demands that matter arises from matter fields. The theory of relativity says the matter causes space-time curvature and, hence, at high energies quantum matter should create a space-time curvature. At present, neither the theory of relativity nor the quantum theory addresses this situation adequately. Even an instrumentalist would not likely argue that the two theories should not be modified at least to the extent that the correct physics is predicted [1].

1.2.2 The Second Future: A Unified Field Theory

In spite of its predictive successes, the standard model is not the final answer. This fact has motivated the search for a higher level of unification. An enhanced standard model should:

1. Include gravity and explain dark energy/matter
2. Explain why the 19 + free input parameters have the values they do
3. Give the correct prediction for the masses of neutrinos
4. Develop a truly finite theory
5. Explain why there are three generations of leptons and quarks
That the standard model does not include gravity has been a known shortcoming since its inception. But beyond that, the model has the rather contradictory feature of requiring a list of 19 or so adjustable constants. The laws of the theory require specifying the constants. Their values are found by experiment. These constants specify the properties of the particles. But there is no known explanation for why the constants have the values they do. Every time the model is run, the values of the constants must be specified [14]. Each adjustment represents a basic physical fact of which current theory provides no explanation.

Both the theory of relativity and the quantum theory suffer from the “problem of infinities”. To the many sensible questions asked of these theories, the answer is often “infinity”, when observations or experiments suggest otherwise. This implies there are fundamental flaws or at least significant shortcomings in both theories. While the standard model does not suffer directly from this problem, it handles the “problem of infinities” through a process called ‘renormalization’. Renormalization involves methods where infinites, when they occur, are cancelled or ignored. In quantum field theory, renormalization refers to any collection of techniques for treating infinities arising in calculated quantities that determine the relationship between the parameters describing large distance scales (low energy scales) in a theory and the parameters describing small distances (high energy scales).

Understanding renormalization will require a bit of a detour. First of all, what is it that is being renormalized? The answer: the Lagrangian. More specifically, it is the parameters of the Lagrangian that are renormalized. The purpose of the Lagrangian is to summarize the dynamics of a system. Named for Joseph Louis Lagrange (25 January 1736 – 10 April 1813), it was the Irish mathematician William Rowan Hamilton (4 August 1805 – 2 September 1865) who introduced the Lagrangian as a reformulation of classical mechanics. The Lagrangian is defined as the kinetic energy ‘$T$’ of the system minus its potential energy ‘$V$’. In symbols, $L = T - V$.

Finding a serviceable Lagrangian is crucially important from a unification perspective. This is because energy can be represented in a way that does not require that the full status of a system be known at every moment in time. One of the major characteristics of classical physics is the prospect of knowing the complete status of a system at every moment. In fact, classical physics depends on this assumption. But, quantum processes are uncontrollable. It is not possible to know the complete status of system at each moment. Classical equations that require specifying the exact position and momentum of a system at each moment cannot be employed at the quantum level [24].

Although the Lagrangian was originally developed to describe classical mechanics, it is applicable to quantum mechanics (more on this later). In fact, if the Lagrangian of a system is known, then the equations of motion of the system can be obtained.

In the latter half of the twentieth century, theorists began to recognize the increasing importance of ‘symmetry’ as a fundamental principle and began incorporating it into their work. The idea is that all the observable features of a physical system exhibiting a
kind of symmetry remain "unchanged" even though other facets of the physical system change [25]. The symmetry of a system is a physical or mathematical feature (observed or intrinsic) that is "preserved" under some change. An important example of symmetry is the invariance of physical laws under arbitrary differentiable coordinate transformations in general relativity.

Symmetry took on greater importance when in 1919 the German mathematician Emmy Noether (23 March 1882 – 14 April 1935) showed that every continuous conservation law was associated with a given symmetry. And because conservation laws are at the heart of theoretical physics, this was a celebrated discovery. There are many different kinds of symmetries. The standard model, for example, is characterized by three different symmetries denoted ‘$U(1) \times SU(2) \times SU(3)$’. The $U(1)$ symmetry is associated with the electromagnetic interaction, $SU(2)$ with the weak nuclear interaction and $SU(3)$ the strong interaction.

Lagrange formalism is tied closely to Noether's theorem. If the Lagrangian is invariant (does not change) under a given symmetry, then the accompanying equations of motion are also invariant under that symmetry. This is called 'gauge symmetry' and is helpful in showing that theories are consistent with either special or general relativity.

Moreover, without symmetry, the process of renormalization would not be possible. There must be a way of cancelling the left-handed infinities with the right-handed ones, the up infinities with the down infinities. There must be just as many lefts as rights or ups as downs to make it all work. This is why symmetries play such a crucial role in the renormalization process.

Notably, however, Nature is not symmetrical. The process by which a system, described in a theoretically symmetrical way, ends up in an apparently asymmetric state is called 'spontaneous symmetry breaking'. Interestingly, quantum mechanics seems to demand that symmetry breaking take place at some point. For spontaneous symmetry breaking to occur, there must be a system in which there are several equally likely outcomes. The system as a whole is, therefore, symmetric with respect to those outcomes. However, if a measurement of the system is taken, i.e. if the system is interacted with in any way, the result is one specific outcome. While knowing the system as a whole is symmetric, it never displays this symmetry, but ends up in one specific state [26]. This is basically a statement of the 'measurement problem' in quantum mechanics which I will say more about later.

Spontaneous symmetry breaking plays an important role in elementary particle interactions. The standard model predicts the existence of a number of particles. However, within the normal Yang-Mills approach, particles are massless, when, in reality, some of them have mass. This is a shortcoming of the theory. The idea of spontaneous symmetry breaking is used in conjunction with the Higgs mechanism to explain how particles gain mass. It also predicts the existence of the recently detected Higgs boson [27,28,29]. The fact that Nature appears to be in part asymmetrical complicates the job of creating a unified theory because it establishes a requirement for finding an explanation for broken symmetries.
To gain a relatively oversimplified idea of renormalization, consider the following problem: in physics, the notion of velocity is described in terms of distance and time i.e. \( s \) (distance) = \( v \) (velocity) \( \times t \) (time). Suppose I am driving a car and want to know how fast I’m going right now. How is this question answered in terms of the distance formula? If it is ‘right now’, then no time ticks off the clock \( t = 0 \), and, in that time, I go nowhere, so \( s = 0 \). Hence, \( v = 0/0 \). But since the speedometer reads, say ‘69.3 miles per hour’ right now, the answer ‘\( v = 0/0 \)’ is nonsense. How is this discrepancy overcome? Suppose, instead of this very instant, I measure how far I travel in, say \( t = .001 \) sec. That’s not an exact instant, but it is close. Further, suppose in \( t = .001 \) sec, I travel .0692 miles i.e. \( s = .0692 \), then \( v = 69.2 \) (.0692/.001) miles per hour. This is a better answer. On the other hand, the measurement ‘\( t = .001 \) sec’ is not instantaneous. To get a better estimate, suppose I measure the distance traveled in \( t = .0001 \) sec. This is closer to instantaneous than \( t = .001 \) and if the distance traveled during that time is ‘\( s = .006927 \)’, then \( v = .006927/.0001 = 69.27 \) miles per hr. This measurement is closer to instantaneous, but still, not exact. I could repeat this process again by choosing \( t = .00001 \), where say \( s = .0006929 \), then \( v = 69.29 \). Notice a pattern? As the time intervals get shorter and shorter, my speed approaches 69.3 miles per hr. This is an assumption on my part because I didn’t measure \( v \) at \( t = 0 \). Instead, I extrapolated this result from the measurements I took previously. Theories that are amenable to this kind of procedure are said to be ‘renormalizable’. While renormalization procedures are, in general, much more complicated, this somewhat pedestrian example gives a feel for what the renormalization process entails [56, 90].

Renormalization has become a necessary feature of quantum field theory because of the fluctuations that occur at short distances in quantum physics. Observables in quantum theories ‘\( E(x,t) \)’, for example, which designate the electric field strength, fluctuate from measurement to measurement and from point to point. An exact value for \( E(x,t) \) cannot be calculated. The best that can be done is to measure the average value of \( E(x,t) \), say, within the area of a circle (or sphere) of radius ‘\( a \)’. Mathematically this can be represented by

\[
\langle 0 | (E(x,t) - E(x + a, t))^2 | 0 \rangle \rightarrow \frac{1}{a^4}
\]

And as \( a \rightarrow 0 \), \( 1/a^4 \rightarrow \infty \). The quantum electric field cannot be defined under these circumstances. For example, derivatives of a field ‘\( E(x,t) \)’ cannot be defined when the difference ‘\( E(x + a) - E(x) \)’ diverges as the separation (\( a \)) vanishes. This is a fundamental problem and is a consequence of the uncertainty principle in quantum mechanics. The question becomes: what to do about the infinities [37]?

To give any meaning to a quantum field theory, it is necessary to complete a process called ‘regularization’. Regularization is a sub-process within renormalization. It is accomplished by removing from the theory all states having energies much larger than some cutoff value, say ‘\( \epsilon_0 \)’. This effectively removes the infinities [37]. This is the same
thing as replacing \( t = 0 \) with \( t = .001 \) in the ‘car velocity’ example. Making that replacement got rid of the nonsensical answer.

However, regularization comes at a price. There is no way of guaranteeing that cutting off part of the short-distance scale structure will have little or no impact on the long-distance scale behavior of the theory. If the short distance structure were relevant to long distances calculations in any given problem, it would introduce, and often does introduce, unacceptable errors in the quantum calculations.

The basic idea behind renormalization is that all effects of the very high-energy states on the low-energy behavior of the theory can be discarded provided the theory’s Lagrangian is modified to account for the effects that result from the discarded states. To accomplish this, a Lagrangian \( (L_0) \) density function is defined which describes the energy states of the system together with a regulator \( \epsilon_0 \) (cutoff) that truncates the theory’s state space at some very large value. The other parameters specified by the Lagrangian remain well defined so long as \( \epsilon_0 \) is kept finite. These parameters are referred to as ‘bare parameters’. To compute the bare parameters for a particular \( \epsilon_0 \), choose two convenient processes or quantities, compute them in terms of the bare parameters using \( L_0 \), then adjust the bare parameters until theory and experiment agree [31]. Again, this is similar to the ‘car velocity’ example where, as shorter and shorter time intervals were chosen, the answer got closer and closer to the actual answer.

To understand the role of the cutoff in defining the theory, start with \( 'L_0' \) and a cutoff \( \epsilon_0 \). Remove from this theory all states having energies or momenta larger than some new cutoff \( \epsilon (\epsilon \ll \epsilon_0) \). Repeat the process described above to determine if the energies discarded in the first iteration would have any long-range impact on the original calculations of interest. If significant errors are found, \( L_0 \) must be modified to account for the previously discarded energies. This produces a new Lagrangian, say \( 'L = L_0 + \delta L_0' \), where \( \delta L_0 \) represents an additional adjustment producing results that more closely agree with experiment. When a theory is renormalizable, only a finite number of adjustments are required to closely agree with experiments. The result is a theory without infinites [31].

The infinites that appear in the series expansion calculations associated with quantum field theories result from the mathematical approaches used in the calculations. Renormalization is driven by the physics. Terms within the mathematical expansions are discarded or adjusted only when it makes “good physics sense” to do so. Moreover, formalized methods have been developed for determining if the Lagrangian associated with a field theory is renormalizable. In fact, renormalization has become quite useful in determining if new physics is involved beyond the cutoff energies associated with a given theory [32].

Renormalization is only applicable to certain types of field theories having specific characteristics. Not all theories are renormalizable. At times no matter the degree of compensation applied to the Lagrangian, all manner of adjustment fails to tame the unruly behavior. Such theories are non-renormalizable. The general theory of relativity,
for example, is a non-renormalizable theory. There is no known way of modifying general relativity so that it can explain the physics at extremely short distances.

If the Lagrangian associated with a given theory is non-renormalizable, it serves to indicate that beyond a cutoff value, the calculations of the theory cannot be trusted and signifies that new physics is involved past the cutoff. As it relates to unification, renormalization is crucial. The fact that the theory of general relativity is non-renormalizable indicates that the physics beyond a cutoff is entirely different than what general relativity says it is, providing fairly good evidence that general relativity cannot describe physics at very small distances. This is the primary reason physicists scurry to the safer harbingers of string theory [32].

Steven Hawking in collaboration with the Oxford mathematician Roger Penrose showed that the Universe started with a ‘singularity’ – a mathematical point of infinite density. Suspicions had been swirling for some time that this might be the case, but it was never really clear whether the singularity was inevitable or a consequence of the assumptions that went into the theory. All doubt was removed when Penrose proved that the singularity conjecture was inescapable. Hawking’s and Penrose’s demonstration was completed entirely within general relativity. This discovery had two important outcomes. Firstly, it showed that the theory of relativity alone could not provide a theory of everything – a complete unification of physics. The idea that the Universe started at a singular point of infinite energy seemed to any reasonable intellect an absurd assertion. More likely, it was general relativity that broke down. Past a certain cutoff, general relativity gave absurd answers that cannot be trusted. Secondly, it put general relativity and the quantum theory on an unavoidable collision course. The quantum theory does not allow for the exact specification of an amount of energy at a point. The uncertainty principle within quantum mechanics forbids a singularity. This put general relativity and the quantum theory into direct contradiction and illustrated the importance of finding some way of uniting the two theories.

Physicists desire theories that are natural. They expect the macro aspects of a theory to follow naturally from its microscopic aspects and find it undesirable, indeed unlikely, that the microscopic theory contains various free parameters that are carefully adjusted to give the macro part of the theory its special properties. What they really want is an effective theory without fine-tuning the various parameters in the high-energy theory [33]. This is called the ‘hierarchy problem’. Theorists have yet to uncover a suitable theory that does not require fine-turning at least to some extent. It was thought that string theory or $M$-theory might provide an approach free of fine-tuning, but this hope has faded somewhat in the last few years.

For all intent and purposes, from the standpoint of naturalness, the quantum theory is a theoretical mess. Not only is there no obvious approach to uniting it with general relativity, but its logical foundation is incomprehensible. No one really understands it. The theory seems on its face to be completely unnatural. To get the theory to agree with experiments mathematical inventions have been devised. These inventions do not arise from the foundations of quantum mechanics. They are instead simply inserted to
get the theory to agree with experimental results. To argue that the process of renormalization or, for that matter, spontaneous symmetry breaking is fundamentally the way nature works borders on the ridiculous.

A good example is the Higgs mechanism. It is a completely unnatural invention. The hunt for the Higgs boson is not like your neighborhood Easter egg hunt. There is really nothing to find. The Higgs particle is a boson and bosons are not directly detectable in particle accelerators. The best that can be hoped is that by smashing particles together at near the speed of light, the numbers of particles directly detected emerge in the percentages predicted by the standard model. Then an argument can be made that, since the numbers of particles that can be seen, show up in percentages that agree with what the Higgs mechanism predicts, things must be on the right track. The argument is surely circular. The preference would be a more natural theory, where concepts like ‘spontaneous symmetry breaking’ emerge from the constraints of the theory and are free from inventions for which there is no explanation other than “it just works”. It is a safe bet that some physicists will continue searching for a theory that gives a better, more natural explanation than what the quantum theory currently provides, regardless of how perfect its experimental predictions are. To that end, the search continues for a unified field theory. The most common approach is to create a higher dimensional extension of quantum field theory.

1.2.2.1 The Rise of String Theory

The overwhelming favorite, some might argue the only candidate for a complete unification of physics, is string theory. With thousands of researchers working on the idea, it is where most physicists, at least those who think that a unified theory of physics is important, are placing their bets. The premise of string theory is that matter does not consist of point-like particles, but rather tiny loops of string. The strings vibrate like the strings of a musical instrument. Each of the different harmonics of a string represents a different particle. From this beginning, the laws of physics emerge: general relativity, electromagnetism and the Yang-Mills gauge theories that describe nuclear interactions [12]. For this reason, string theory can rightly claim to be a potential “theory of everything”.

1.2.2.1.1 The Theoretical Advantages of String Theory

From a unification perspective, string theory offers many alluring features. All the forces and particles of Nature can be explained as characteristic vibrations and actions of the string. The different particles arise from the different resonance vibrations. The vibrations of open strings are consistent with the gauge principle, which is responsible for the electromagnetic and the nuclear forces. It includes gravity, which is explained by the vibrations of closed strings. It unifies bosons with fermions – the two classifications of “particles/forces” found in Nature. In most theories, the laws of motion of the particles are separate from the laws that governed the forces that act on the particles. This is certainly true of classical physics. But in string theory once a description of how strings move is given, forces arise naturally by the breaking and joining of the strings. There are but two fundamental quantities: the ‘string tension’, which describes the energy in
the string, and the 'string coupling constant', which gives the probability of a string breaking into two strings or two combining into one. In addition, string theory is governed by one simple principle: as a string moves through space, it draws out a two dimensional surface in space-time. The string moves so as to minimize this surface area. This constitutes the whole of the Natural law governing strings [14].

String theory overcomes the road blocks that have plagued previous approaches to unifying the quantum theory with general relativity. Particles in quantum field theory are conceived as existing at a point. This means that they can potentially become arbitrarily small. But the smaller the scale gets the more turbulent quantum fluctuations become. At short distances it is simply impossible to reconcile the smooth space-time associated with general relativity with the violent fluctuations of the quantum theory. String theory overcomes this by placing a limit on how small the particles (strings in string theory) can become. The lengths of most strings are thought to be on the order of $10^{-33} \text{cm}$ about the size of a Planck length. Putting a limit on the size of a string tames the quantum fluctuations just enough to allow compatibility with general relativity [34].

Furthermore, the supersymmetric string theories seem to be free of the mathematical anomalies that have plagued previous attempts at unification. Supersymmetry greatly enhances the aesthetic beauty of a theory. "Unification physicists" want to believe that Nature uses all the symmetries that are mathematically possible because it creates a highly aesthetic and tightly woven theory. But beyond purely aesthetic considerations, there is the practical issue of explaining quantum fluctuations. At present, the parameters of the standard model must be finely turned to about one part in $10^{15}$ to adequately deal with these fluctuations. This is because each distinct particle makes its own contribution to the quantum frenzy. In order to cancel these effects, fine-turning is required. However, if each boson is paired with a fermion (supersymmetry), this greatly enhances the canceling effect and gives a much more natural explanation for what is actually experienced in Nature, removing the necessity for fine-tuning the standard model parameters. In addition, at high energies, the strengths of the three forces of the standard model almost, but do not quite agree. But when supersymmetry is included, the discrepancy in their strengths vanishes due to the quantum effects of the additional super-partners. It appears that at high temperatures, the three forces of the standard model unify into one force. Theories that predict this unification are called 'grand unified theories' or 'GUTs' for short [34]. I'll be saying more about them later.

With its apparent bonanza of irresistible elegance and simplicity that seems to explain so much, many theorists regard string theory as on the brink of providing a completely unified theory of physics. In addition, there is strong evidence that string theory is finite. A theory devoid of infinities escapes the necessity of using approximation formulas based on renormalization.

1.2.2.1.2 Lingering Doubts about String Theory

In spite of all its positive features, string theory has detractors. There are five ten dimensional superstring theories as opposed to one. And none of the five theories
describes our Universe. All five theories sport six extra spatial dimensions that are not directly experienced. The extra spatial dimensions introduce complexities into the theory. The only feasible manner of describing these extra dimensions is to imagine them curled up into small, difficult to detect, compactions. But, loosely speaking, the curling up can be accomplished in innumerable ways, which leads to thousands of variations in the number string theories and none of the variants seem compatible with the standard model [14].

However, in 1995, it was conjectured that the five superstring theories were in some sense part of a more fundamental theory. The more fundamental theory was given the name ‘\( M \)-theory’. But, \( M \)-theory has yet to be completely defined. It attracts criticism for lacking predictive power, for being incomplete and untestable. Likely, it will require the development of a new mathematical language, which makes further progress on string theory difficult. At some point the question “what is string theory?” will have to be addressed [37].

The \( M \)-theory conjecture had barely been digested before it was discovered that to make the conjecture feasible, newer higher dimensional objects called ‘branes’ would have to be introduced. Branes are entities to which the ends of strings can attach themselves. The inclusion of the ‘brane’ concept gave string theory added flexibility. Inspired by the holographic principle, it gave the ability to mathematically model black hole entropy. Unfortunately, the black holes modeled so far in string theory are hypothetical, not the ones found in Nature.

In 1997, a new conjecture appeared, inspired by the holographic principle, that there might be a way of interpreting the extra spatial dimensions in the context of real world physics. Referred to as the ‘\( AdS/CFT \) correspondence’, it showed that there was an equivalence relationship between a string theory and gravity defined on one space and a quantum field theory without gravity defined on the conformal boundary of that space. This suggested that the quantum gravitational theory embodied in string theory is equivalent to a quantum field theory with one less dimensional element, namely gravity. It was the first time string theory could be connected with another theory – one associated with real world physics [16].

But the \( AdS/CFT \) correspondence conjecture employs background geometries in anti-de Sitter spaces. Anti-de Sitter geometries have negative curvatures, which require a negative cosmological constant. But, so far, all observations indicate, fairly conclusively, that the cosmological constant associated with our Universe is positive. It remains to be seen whether or not the same conjecture will hold in model universes that have a positive cosmological constant [14].

Of the current string theory candidates that could describe our Universe, all of them include the notion of ‘supersymmetry’. Supersymmetry is the conjecture that all fermions have a bosonic partner. But none of these super-partners has, as yet, been experimentally detected, possibly because the super-partners are so massive it would take a large amount of energy to detect them. This suggests that either supersymmetry
does not exist or that the symmetry is broken in some manner. If supersymmetry is ruled out as a fundamental part of Nature, it could bring an end to string theory [36].

The black holes modeled so far in string theory are not the ones found in Nature, which brings up the broader question: can string theory model our Universe? While the introduction of the “brane” concept brought a welcome amount of flexibility to string theory, it added complexity. It greatly increased the number of background geometries upon which the string can live. The effect was to increase enormously the number of possible consistent string theories [17]. At this point it is not clear how to find or if there is a string theory that describes our Universe.

The critics of string theory argue that, with so many possible theories, there are no ways of falsifying the theory and never will be. They argue that string theory has a mythical character and will make no viable testable predictions and, in the final analysis, is not science. As a counter to this view, some string theorists have endorsed the concept of ‘the landscape’ of string theories - a vast number of mathematically consistent possible universes, some of which may actually exist. The landscape is their solution to the unfathomable number of possible string theories, where the anthropic principle is suggested as means of determining which string theories could describe our Universe.

The anthropic principle is controversial. It is often viewed as unscientific, in large part, because it is sometimes invoked as proof that there is a supernatural designer of the Universe or, in its weaker form, as an argument against the existence of a supernatural designer [17].

John Schwartz and Michael Green provided strong evidence that string theory is a finite theory, which means it should give finite answers to sensible questions and avoid the “infinity” problem [38]. But there does not seem to be general agreement on this point. Perturbation approximations, which are the primary tools employed by string theorists, generally have an infinite number of terms. String theory would have to be shown completely renormalizable or do away entirely with renormalization. In the final analysis, it will have to be convincingly proven that string theory is finite [14]. And so far not all physicists agree whether or not this has been achieved or even if it is necessarily required.

Whether or not string theory becomes an acceptable quantum theory of gravity, for now, remains an open question. It shows great promise. But, it must overcome a myriad of issues not the least of which is answering the question “what is string theory?” Recent progress has been slow, so watching and waiting has become a primary activity.

1.2.3 The Third Future: A New, More Fundamental Theory

An increasing number of physicists subscribe to the idea that a new, more fundamental theory is the only road to unifying physics. Frustrated with the quantum theory because “it doesn’t make sense”, many physicists search for an explanation that can make better sense of or can replace the quantum theory.
But, those who think along these lines generally have an “I'll know it when I see it” intuition rather than a well thought out approach that would lead to a new theory. At present, no really serious ‘new theory’ is capturing the imagination of at least a significant number of physicists. Quite the contrary, those who support this idea, for the most part, appear genuinely conflicted as to what approach is required. Lacking an acceptable alternative, the majority of physicists cast an eye towards string theory.

In Book IV, a new approach to unification, which I will name ‘Image Theory’, will be presented. It’s a radical departure from the current unification proposals. It is based on an entirely new approach to how the principles of logic are developed. Since the foundations of mathematics are traceable to the constructs of modern logic, a change in logical approach necessitates an alternate approach to developing the mathematics of the theory.

Of the important questions at the foundations of Image Theory, the first asks: given a physical system, presumably an assembly of Natural laws described mathematically, do the ideas associated with those Natural laws belong in the theory? Image theory will answer this question in the affirmative. Admittedly this is a controversial topic. The question boils down to determining what our ideas, if they exist, are made of. If ideas are material, then there should be little debate that they belong within the domain of physics. Presumably material ideas would be governed by the same Natural laws that underpin the physics of other material things, brain waves, for example, that could, in some manner, be detected either directly or indirectly through sense experience. Physics has not entirely ignored this question. The question normally presents itself as “what is consciousness?”

But traditionally, philosophers have regarded ideas as non-material. And physics has been reluctant to include anything within its domain that cannot be, at least potentially, verified through sense experience. From a traditional philosophical standpoint, things perceived though sense experiences are material – something made of matter/energy. Beyond that, man is endowed with a mind, usually regarded as non-material, as opposed to his ‘material’ body. A physicist would have a reasonable shot, at least potentially, at describing the nature of a material substance – something assessable to sense experience and potentially observationally verifiable or, in some sense, physically describable. But what would it mean to describe the non-materiality of the mind? And, moreover, why is it necessary to include a non-material mind, something unlikely to be assessable to measurement, within the description of the material world? Physicists would generally expect experimental verification of theoretical suppositions and it is not clear how there could be experimental verification of a non-material thing.

The second question is an old one thought to be a purely metaphysical inquiry. Does an external world exist independently of our perception of it? I think most physicists would answer in the affirmative. It would seem odd if physicists adopted the contention of Lord Berkeley, who denied that an external world existed. If it isn’t an external world, what is it that physicists are writing and prognosticating about?
Oddly, the prevailing attitude amongst physicists appears decidedly agnostic toward this question. Current interpretations of the quantum theory avoid the issue of whether or not an external reality exists by simply claiming that what’s pertinent is whether or not the theory predicts the correct outcome of a measurement. In fact, this is precisely what the quantum theory suggests: that the only reality is the one revealed after a measurement or an observation. Outside of that, knowledge of the unobserved world is out of reach; probabilities, but no reality, can be assigned to it.

If an external world does exist, are our mental perceptions of it faithful representations of that world? Image Theory will answer the first part of the question in the affirmative, but the second part in the negative. Yes, there exists an external world independent of our perception of it. But no, those perceptions are not faithful representations of that world. And I think that the majority of physicists would not be overly surprised by this statement and many might support it. But, if our perceptions are not faithful representations of the external world, wouldn’t the prognostications about that external world be factually wrong by default? Of course, this begs the additional question: if it is true that our perceptions are not faithful representations of an external world, then what are they representations of? The quantum theory gives probabilities for the possible realities that appear when an observation is made. Image theory will argue that it might be possible to do better.

These are the questions that this study will address. But, before focusing attention on these important questions, it will be necessary to understand more thoroughly the issues faced in attempting to unify all of physics. This will be topic of the remaining chapters.
Chapter 2

The Struggle for Unification and the Emergence of Quantum Mechanics

“The microwave oven is the consolation prize in our struggle to understand physics.”

~ Jason Love

2.0 Introduction

Chapter 1 reviewed the status of the attempts to unify physics into a single consistent theory. There, it was learned that physics is dominated by two remarkably successful theories – the theory of relativity and the quantum theory. The main obstacle to unifying the two theories is their seemingly fundamental incompatibility. It will be instructive to examine, from a historical perspective, how past events and discoveries shaped the direction of modern physics and its march toward unification. This will be the subject of the next two chapters.

2.1 Einstein’s Attempt at Unification

The first serious attempt towards a complete unification of physics came at the hands of Albert Einstein (14 March 1879 – 18 April 1955). The pursuit lasted the last 30 years of his life. Over about a ten year period, 1905 to 1915, he formulated the special theory of relativity, which corrected the anomalies within Newtonian mechanics and produced a theory of gravity embodied in the general theory of relativity. Earlier, Maxwell had shown that electricity, magnetism and light were all manifestations of the electromagnetic field. However, the general theory of relativity left Maxwell’s electromagnetic theory unaddressed (besides gravity, the electromagnetic force was the only other force known at the time). Einstein set his sights on remedying “this deficiency” by attempting to unify general relativity with electromagnetism.

The twentieth century saw a movement towards a more personalized physics, where the point of view of the individual observer became increasingly important. Physical systems of the prior generations, Newtonian physics and earlier attempts at describing the electromagnetic force, assumed the motions of objects could be described against an idealized stationary space, which acted as a reference frame for all observers. This simplified the task of describing how objects moved in space. At this point, physics was dominated by the law of Galilean relativity. Galileo postulated that any two observers moving at constant speed and direction with respect to one another would obtain the same results for all mechanical experiments. But Galilean relativity, when applied to electromagnetic experiments, appeared to fail leading to results that did not agree with observations. The arguments designed to justify the observational anomalies became increasingly indefensible and, in the end, a revolution in physics ensued.
The revolution saw two significant changes in perspective. Physics became both localized and individualized. Before the revolutionary ideas of Einstein and Maxwell, the gravitational force, as well as the electric force between two charged particles, was described by an ‘action-at-a-distance’ formulation. The force between two masses or two charges was described as proportional to the distance between them - acting instantaneously regardless of the distance between the two masses or two charges. But such an ‘action-at-a-distance’ formulation violated both common sense and modern theories. Relativity proclaimed that nothing traveled faster than light. Having a limit on how fast signals could propagate through space made it difficult to defend the idea of ‘instantaneous action-at-a-distance’.

The action-at-a-distance formulation, which maintains that physical influences are transmitted through empty space without any material or physical agency, was eventually replaced with a localized field concept in both the theories of electromagnetism and gravity. A field is region throughout space in which two bodies separated in space exert a force on each other. In a field description, rather than body ‘A’ directly exerting a force on body ‘B’, body ‘A’ creates a field in every direction around it and body ‘B’ experiences the field at its position in space. If a change occurs at the source, its effect propagates outward through the field at a constant speed and is felt only after a certain time delay. Each type of force (electric, magnetic, nuclear, or gravitational) has its own field; a body experiences the force only if it is also a source of that kind of field. If two bodies exert a mutual force, they possess potential energy that depends on their relative positions; energy is regarded as residing in the field and not in the bodies.

The individualization of physics greatly modified the job of the physicist. Before relativity was discovered, it was believed that Natural laws remained unaffected by the motions of observers relative to one another (Galilean relativity). But experiments with light showed that Galilean relativity failed, requiring a new set of Natural laws that remained invariant (did not change) while taking into account the motions of individual observers.

The theory of relativity, developed in 1905, was restricted to inertial reference frames, where observers move with a constant velocity. But it succeeded in removing the contradictions that arose between Newtonian mechanics and Maxwell's theory of electromagnetism. Newtonian mechanics tacitly assumed that light speed in free space was infinite, and hence, light signals, emitted from any position in space, reached all observers simultaneously. But Maxwell's theory had shown that light traveled at a finite speed. The theory of relativity postulated that the velocity of light (in a vacuum) was indeed finite, but that its speed would be measured by all inertial observers as constant, regardless of how fast the observers were moving relative to one another. This meant that light signals would not necessarily reach all inertial observers simultaneously. The previously accepted idea of event simultaneity for all inertial observers had to be replaced by the “principle of relativity”: “Physical laws must be the same in all inertial reference frames.”
This change in perspective produced consequences that seemed odd, but true. Time became local. Observers in two different inertial frames would not necessarily agree on the time that a given event occurred. Neither would they necessarily agree on where the event occurred. What they could agree on was the 'space-time interval' as expressed by the equation

\[ x^2 + y^2 + z^2 + (ict)^2 = x'^2 + y'^2 + z'^2 + (ict')^2, \]

where \( c \) is the speed of light in a vacuum and the spatial locations and times of an event are given by \((x, y, z, t)\) for one observer and \((x', y', z', t')\) for another observer respectfully. While there might not be agreement on the exact location or the time that an event occurred i.e. \( x \neq x' \) and \( y \neq y' \) etc., the two observers would agree on the space-time interval. A physicist would say that the space-time interval remains 'invariant' in all inertial frames. Often referred to as 'Lorentz invariance', named for the Dutch physicist H. A. Lorentz (1853 – 1928), the equation above is considered a law of Nature subject to the constraint that it only applies to inertial systems. The redefinition of simultaneity allowed Einstein to resolve the conflict between Newtonian mechanics and the principle of the constancy of the velocity of light by providing a conceptual justification for Lorentz invariance. He successfully integrated two major fields of physics in what amounts to a conceptual unification [40].

Special relativity demands an absolute and finite limit to the speed of light, which applies to any signal transmission. This violated the Natural law of Newtonian gravitational theory. A gravitational interaction between massive bodies, an instantaneous action-at-a-distance, posed a contradiction to the postulate of the non-existence of any signal transmission exceeding the speed of light [40]. This, coupled with the conceptual conflict between Newtonian gravitation, which conceptualizes gravitational interactions as “action-at-a-distance” and Maxwell’s electromagnetism, which conceptualizes the electromagnetic interactions in terms of a dynamic field concept, led Einstein to his principle of equivalence [40]:

“We arrive at a very satisfactory interpretation of this law of experience, if we assume that the systems \( K \) and \( K' \) are physically exactly equivalent, that is, if we assume that we may just as well regard the system \( K \) as being in a space free from gravitational fields, if we then regard \( K \) as uniformly accelerated. This assumption of exact physical equivalence makes it impossible for us to speak of the absolute acceleration of the system of reference, just as the usual theory of relativity forbids us to talk of the absolute velocity of a system; and it makes the equal falling of all bodies in a gravitational field seem a matter of course.

— Einstein, 1911
Einstein suggested that this should be elevated to the status of a general principle. He goes on:

“As long as we restrict ourselves to purely mechanical processes in the realm where Newton’s mechanics holds sway, we are certain of the equivalence of the systems $K$ and $K'$. But this view of ours will not have any deeper significance unless the systems $K'$ and $K'$ are equivalent with respect to all physical processes, that is, unless the laws of nature with respect to $K$ are in entire agreement with those with respect to $K'$. By assuming this to be so, we arrive at a principle which, if it is really true, has great heuristic importance. For by theoretical consideration of processes which take place relatively to a system of reference with uniform acceleration, we obtain information as to the career of processes in a homogeneous gravitational field.”

This observation culminated in the theory of general relativity. In Einstein’s version, the principle asserts that in free-fall the effect of gravity is totally abolished in all possible experiments and general relativity then reduces to special relativity, as in the inertial case. This insight allowed Einstein to apply his principle of relativity to non-inertial systems, and thus, to provide a local field theory of gravity – localized in the sense that gravitational effects would depend on where an observer was positioned within the gravitational field; and individualized in the sense that the time registered on the clocks of two observers would not necessarily agree if the gravitational potentials were not equivalent at the places where the observers were located within the gravitational field.

Still, the success of this unification came at a price. The gravitational interaction was conceptualized as purely geometrical. Einstein now believed that his geometrized relativistic theory of the gravitational field demanded unification with the concept of the electromagnetic field. He faced a huge challenge in attempting to unify gravity with electromagnetism. The equations of general relativity are non-linear, whereas those describing electromagnetism are linear. Gravity, as it turned out, was not a force, but the consequence of the warping of space and time brought about by the presence of matter/energy. Gravity was described geometrically by differential equations associated with the curvature of space-time. Applying differential equations to the problem of gravity required that space-time be a smooth continuous substance. In Benjamin Franklin’s day, electricity was conceived as a continuous fluid, but today, it is conceptualized as coming in discrete packets generally referred to as ‘quanta’. This meant that the smooth continuous non-linear geometrically based field equations of general relativity would require some kind of transformation into chunky quantized entities. And it was not at all clear how this ‘quantization’ could be accomplished. The problem boiled down to describing how discrete particles arose from the smooth fabric of space-time.

In 1921, Theodor Kaluza (November 9, 1885 – January 19, 1954) purposed adding a fifth dimension to space-time and was able to show that his theory contained four-dimensional general relativity in the presence of an electromagnetic field. But the extra dimension was bothersome. Where did it come from and why was it not observable? In
1926, the Swedish mathematician Oscar Klein (September 15, 1894 – February 5, 1977) proposed that the fifth dimension was curled up in a circle of very small radius that would not be observable, except at very high energies [41,42].

In a sense, Einstein became smitten with the Kaluza-Klein approach. As a minimal extension of his general theory, it was now possible to envision how gravity and electromagnetism (space-time and matter) could be unified into a purely geometrical theory. All he had to do was show that particles emerged in the theory as non-singular solutions of matter fields [17]. It is interesting to note, at the time, many of his colleagues considered Einstein's pursuit, along with the idea of extra dimensions, a silly waste of time. It is ironic and somewhat of a testament to Einstein's foresight that, of the current theories considered serious contenders for a total unification of physics, most sport numerous compactified spatial dimensions. Nevertheless, the quantization of matter fields from the gravitational field remains an unsolved problem.

2.2 The Emergence of the Quantum Theories

Meanwhile, physics was experiencing a revolution of a different sort. In 1894, Max Planck (April 23, 1858 – October 4, 1947) began studying the effects of black-body radiation. A black body is an idealized object capable of perfect emission and absorption of radiation. Planck was interested in discovering how the intensity of the electromagnetic radiation emitted by a black body depends on the frequency of the radiation and the temperature of the body [43]. Planck knew that the current theory, embodied in the classical 'Rayleigh-Jeans law', agreed with experimental results at low frequencies, but failed to do so at high frequencies. The law predicted that an infinite temperature increase would result in an infinite amount of energy emissions [44].

Planck came to a better solution by realizing that the interplay of emission and absorption in the black-body is what kept the energy from shooting to infinitely high levels. His equations employed an exponential damping function and statistical arguments, which dealt with the probability that the black-body would emit or absorb an amount of radiation. His results were in much better agreement with experiment. But in order to achieve an acceptable solution, he was forced to assume, evidently, much to his dismay, that electromagnetic energy was emitted only in discrete energy packets, which came in multiples of an elementary unit ‘\( E = h\nu \)’, where \( h \) is a constant (subsequently to become known as 'Planck's constant') and \( \nu \) is the frequency of the radiation [45].

Planck's discovery would lead to a revolution in physics. His 'discrete packets of energy' would become a fundamental concept within the new quantum physics. After Planck's discovery, more and more phenomena, not only radiant energy, but material energy as well, required an explanation involving the 'discrete energies' originally described by Planck in his research.
2.2.1 The Photoelectric Effect

Experiments had shown that electrons were emitted when light fell on a thin piece of foil. It was Einstein who explained this odd result by assuming that light acted like a stream of small particles, which he named ‘photons’. Eventually called the ‘photoelectric effect’, the phenomenon could only be explained if it was assumed that the photons came in Planck’s discrete chunks of energy \( E = h\nu \) [45]. But this was completely at odds with Maxwell’s theory that described light as a wave embodied in the electromagnetic field. In 1923, Arthur Compton (September 10, 1892 – March 15, 1962) published a paper confirming the photoelectric effect for which he won the 1927 Nobel Prize [46]. Unwittingly, Einstein had contributed to the irresistible force of the new quantum revolution - a movement he would later openly detest.

2.2.2 The Discovery of Matter Waves

Two decades later, in 1924, Louis de Broglie (15 August 1892 – 19 March 1987), a French physicist proposed that not only did radiation exhibit particle properties, but particles exhibited wave properties. In his doctoral thesis ‘Recherches sur la théorie des quanta’, based on the work of Planck and Einstein, he introduced a theory of electron waves. His research initiated the ‘wave-particle duality’ theory of matter in which the momentum of a particle depended on its wavelength. The doctoral thesis examiners, unsure of the material, passed his thesis to Einstein for evaluation. Einstein endorsed the wave-particle duality proposal wholeheartedly. De Broglie had created a new concept in physics, ‘matter waves’, uniting the physics of energy (wave) and matter (particle). For this he won the Nobel Prize in physics in 1929 [45].

With every particle of matter with mass \( m \) and velocity \( v \) a real wave must be “associated” related to the momentum by the equation

\[
\lambda = \frac{h}{p} = \frac{h}{mv} \sqrt{1 - \frac{v^2}{c^2}},
\]

where \( \lambda \) is the wavelength, \( h \) is Planck’s constant, \( p \) is the momentum and \( c \) is the speed of light in a vacuum. De Broglie’s equation was confirmed by the electron diffraction experiments of Davisson and Germer [45].

The equation above is interesting because when \( v = c, \lambda (\text{the wavelength}) = 0 \). But light, which travels at the velocity ‘\( c \)’, comes in all different wavelengths, not the zero wavelength the equation predicts. This suggested that de Broglie’s equation only applied to particles that have a non-zero mass. And it further supports the notion that particles with a non-zero mass cannot obtain the speed of light.

On the other hand, in empty space, the photon (light particle) moves at a speed ‘\( c \)’ and its energy and momentum are related by the relativistic relation \( (m = 0) \)
\[ E^2 = p^2 c^2 + m^2 c^4 \]

The energy and momentum of a photon (a massless particle) depend only on the frequency (\( \nu \)) or inversely, its wavelength (\( \lambda \)):

\[ E = pc, \]

where \( p = h k, k = 2\pi/\lambda \) is the 'wave vector' and \( h = h/2\pi \) is the reduced Planck constant (more on this later). The magnitude of the momentum is

\[ p = h k = h \frac{2\pi}{\lambda} = \frac{h}{\lambda} \]

This relationship between particles with zero mass and particles with non-zero mass becomes an important element in the standard model of particle physics. The classical Yang–Mills theory, upon which the standard model is based, is a generalization of the Maxwell theory of electromagnetism, where the chromo-electromagnetic field carries charges. As a classical field theory, it has solutions, which travel at the speed of light, so that its quantum version should describe massless particles (gluons). However, the postulated phenomenon of color confinement in the theory of quantum chromodynamics permits only bound states of gluons, forming particles with mass. This is called the 'mass gap' [47]. How the mass gap emerges is still an unsolved problem in quantum physics.

De Broglie's fundamental equation relating the momentum of a particle to its wavelength was a precursor to the uncertainty principle. De Broglie introduced the idea of a 'wave packet'. A wave packet (see figure below) is a "burst" or "envelope" of wave action that travels through space as a unit.

![Wave packet](image)

It is described by a set of component waves of various energies that interfere constructively over a small region of space, but interfere destructively outside that region. The size and shape of the wave packet may remain constant or change while propagating. De Broglie was able to show that the wave packets displayed many of the properties of the electron – a particle [45].
De Broglie’s equation ‘\( p = h/\lambda \)’ implied that, if the wave packet was confined to a very small area (\( \lambda \ll 1 \)), \( p \gg 1 \), this made it hard to see how measuring the exact momentum and position simultaneously as the wave packet flashed by a suitable measuring device would be possible. As it turned out, the error in measuring the exact location of the wave packet was related to the wave length by \( \lambda \propto \Delta x \), where \( \Delta x \) is the error in the position measurement. Analogously, \( p \propto \Delta p \) represents the error in the momentum measurement. Plugging these two values into de Broglie’s equation:

\[
p = \frac{h}{\lambda} \rightarrow \lambda p = h \rightarrow \Delta x \Delta p \geq h,
\]

which says that the error in measuring momentum and position simultaneously is at least as great as \( h \). In the coordinate representation of a wave packet (such as the Cartesian coordinate system) the position of the wave packet is a function of \( \lambda \). Therefore, a spatially narrower wave packet allows a better measurement of position at the expense of the spread in the momentum of the wave. This trade-off between the ambiguity in position and spread in momentum is an example of the Heisenberg uncertainty principle [45].

### 2.2.3 Signs of a Brewing Controversy

From a philosophical standpoint, de Broglie thought that a real wave (having a direct physical meaning) was associated with the particles. In fact, the wave aspect of matter was formalized by the ‘wave function’ defined by the Schrödinger equation. Subsequently, however, Schrödinger equation was given a purely statistical interpretation by Max Born (11 December 1882 – 5 January 1970). This became part of the Copenhagen interpretation of quantum mechanics. Under this interpretation, the wave function carries no physical meaning, but only gives an appearance of wave behavior to matter, without making real physical waves appear. But until the end of his life, de Broglie believed in a real, physical interpretation of matter waves. Wrote de Broglie:

“\( \ldots \text{the particle must be the seat of an internal periodic movement and that it must move in a wave in order to remain in phase with it was ignored by the actual quantitative physicists [who are] wrong to consider a wave propagation without localization of the particle, which was quite contrary to my original ideas.} \)”

De Broglie ideas were later refined by David Bohm (20 December 1917 – 27 October 1992). The de Broglie–Bohm theory is a non-local hidden variable deterministic theory whose predictions agree with the indeterministic quantum theory. The de Broglie-Bohm theory is today the only interpretation giving real physical status to matter waves. But, since it does not go further in its predictions than the Copenhagen interpretation, it is largely ignored [48].

The sentiments expressed by de Broglie would later erupt into a full-fledged ideological battle that still exists somewhat today. On the one side are the relativists who maintain
that the quantum theory, as understood by the Copenhagen interpretation, does not make sense and should be replaced or be re-interpreted. On the other side are those who, more or less, support the Copenhagen interpretation. I'll have more to say on this topic later.

2.2.4 The Development of the Atomic Model

By this time, physicists began contemplating how these ultra-small particle waves of matter should be described. For the most part they concluded that the fundamental constitution of matter consisted of small indivisible entities (atoms). The concept of an atom dated from the time of Aristotle (384 BC – 322 BC), who thought the Earth was made from four basic elements (earth, wind, fire and water). Contemporaries of Aristotle (Democritus, Aristarchus and Archimedes), to one extent or another, rejected his principles in favor of other ideas (atomism and the idea that the planets revolved around the Sun) that were closer to modern conceptions. But “The Church” so favored Aristotle ideas, it deemed him the “Official Philosopher of the Roman Catholic Church”.

At the time, the Church played a dominant role in all aspects of human life. It dictated which ideas were worthy of being called ‘true’ ideas. It was heresy even to conduct an empirical test or experiment that might challenge the Church’s official position on a topic, Galileo being the most famous victim of the Church’s inquisitions. But eventually the Church’s influence waned, and during the period of enlightenment, scientific thinking became more persuasive.

In 1827, while examining grains of plant pollen suspended in water under a microscope, Robert Brown (21 December 1773 – 10 June 1858), a Scottish botanist, observed that the minute particles executed a continuous jittery motion. He then observed the same motion in particles of inorganic matter enabling him to rule out the hypothesis that the effect was life-related. Although Brown did not provide a theory to explain the motion, it eventually became known as ‘Brownian motion’ [49].

By the time Einstein took up the subject of Brownian motion in 1905, the positivist philosophical movement had begun to emerge. The positivists, at the time, supported the view that atoms did not exist because they were, for all practical purposes, unobservable. And it was their contention that the metaphysical ideas embodied in many of the notions in physics were the primary source of theoretical errors. They maintained that physics, as far as possible, should be purged of all metaphysics, which included the idea of the atom. But, Einstein was able to determine the size of atoms and how many atoms there were in a mole - the molecular weight (in grams) of a gas in accordance with Avogadro's law. Einstein’s paper is widely credited with proving that atoms exist – that they were indeed measurable objects that fit the requirements of positivist thinking at the time [50].

2.2.4.1 The “Plum Pudding” Model of the Atom

How was the atom to be described? Sir J. J. Thomson (18 December 1856 – 30 August 1940), a British physicist and 1906 Nobel Prize winner for the discovery of the electron, was one of the first to propose a model. His model was composed of electrons
(Thomson called them ‘corpuscles’) surrounded by a soup of positive charges that balanced out the electron’s negative charges, somewhat like negatively-charged "plums" surrounded by positively-charged “pudding”. The many arrangements of the electrons in the pudding constituted the numerous possible atoms [45].

Under the direction of Ernest Rutherford (30 August 1871 – 19 October 1937), in 1909, Hans Geiger (September 30, 1882 – September 24, 1945) and Ernest Marsden (19 February 1889 - 15 December 1970) conducted an experiment designed to test the validity of the Thomson model. A beam of alpha ‘+2e’ charged particles was shot thorough gold foil. Since the electrons were assumed uniformly distributed throughout the “pudding”, there was an expectation that the alpha particles would be deflected by, at most, a few degrees. But when the experiment was conducted, a small percentage of particles were deflected through angles much larger than 90 degrees. This suggested that very strong forces were present to enable such marked deflections [45].

2.2.4.2 The Rutherford Model of the Atom

The results led Rutherford to reject Thomson’s model in favor of a miniature solar system model, where a large positive electric charge was concentrated into a small (as compared with the size of the atom) central region, but outside of this "central region" the atom was mostly empty space. The large positive core was orbited by the smaller (in size), but equal and opposite electrically charged electrons rendering the atom roughly electrically neutral [45].

The deflection experiments proved useful in determining the nuclear structure of many materials. The deflection patterns varied depending on the type of material, giving insight into the atomic makeup of the atoms that made up the material. The experiments also showed that electric charge came in discrete units of ‘e’, providing further evidence that electric charge was a quantum phenomenon [45].

Rutherford was able, using purely Newtonian methods, to calculate the energy required for the electron to remain in orbit around the nucleus of a hydrogen atom. This also allowed him to calculate the radius of the electron orbit. His calculations agreed with estimates made in other ways [45].

But the Rutherford model didn’t square with the classical theory of electromagnetism, which predicted that accelerating electrons gave off radiation. As electrons emit radiation, they lose energy. And since electrons in circular orbits were, in effect, accelerating, they should lose energy eventually spiraling into the nucleus. Still, atoms were stable. The “spiraling into the nucleus” effect rarely happened. Reportedly, Rutherford was on the verge of scrapping the entire idea of a miniature solar system atomic model until Niels Bohr came to the rescue.

2.2.4.3 The Bohr Model of the Atom

Subsequently, Niels Bohr (7 October 1885 – 18 November 1962) joined Rutherford at Manchester University as an understudy. Bohr noticed that when passing an electric current through a gas, the atoms emitted radiation in a spectrum of only certain discrete
wavelengths. The discrete wavelengths came in a spectral series which uniquely identified each element. The first such series was found by J. J. Balmer (May 1 1825 – March 12 1898) in 1885 when he studied the visible part of the hydrogen spectrum. Later the spectral series of the invisible parts of the hydrogen spectrum were also developed. The equations that represented each series were remarkably similar. They all exhibited the mathematical form:

\[
\frac{1}{\lambda} = R \left( \frac{1}{m^2} - \frac{1}{n^2} \right),
\]

where \( n \) and \( m \) are integers, \( \lambda \) the wavelength and \( R \) is a constant called ‘Rydberg’s constant’. Bohr hypothesized that the electron could circle the nucleus indefinitely without radiating energy provided that the orbit contained an integral number of de Broglie wavelengths. Using the de Broglie theory, the electron wavelength could be used to calculate the orbital speed required to balance the electrostatic force of the nucleus. The various orbits represented the permitted discrete energy levels the electron could assume. Bohr was able to calculate the energy levels and relate them to the spectral series of the hydrogen atom in almost perfect agreement. This was a startling development and a monumental achievement. Bohr had saved the Rutherford model. Atoms would remain stable avoiding the destructive collisions between the spiraling electrons and the oppositely charged nucleus [45]. Bohr also showed that when electrons either absorb or lose energy, they jump from one energy level to another in discrete units \( E_i - E_f = h\nu \), where \( E_i \) is the initial energy state and \( E_f \) is the final state. In the process of losing energy, electrons gave off energy in the form of radiation [45].

### 2.2.5 The New Physics

The Rayleigh-Jeans law could be deduced from Maxwell’s electromagnetic theory and was known not to provide an accurate description of the emission and absorption of radiation that agreed with experiment. This fact, coupled with the discovery that electrons were confined to discrete orbits as described by Bohr’s model of the atom, de Broglie’s matter waves and Einstein’s analysis of the photoelectric effect, provided overwhelming evidence that, in some sense, the classical theories of the day harbored fundamental flaws.

The Bohr model accounted for certain experimental results in a convincing fashion. It correctly predicted the spectral series of the hydrogen atom. But it also had limitations. It was incapable of extending its predictions to more complex atoms – atoms with more electrons, explaining why certain transitions of energy levels were more likely to occur than others, accounting for small differences in certain spectral lines, and adequately accounting for the formation of molecules [45]. Clearly, a broader theory was required.

---

1 For a complete derivation see: Bohm, D., Quantum Theory, D. Bohm, 1979, pp 7-17.
2.2.5.1 Non-Relativistic Quantum Mechanics

The broader approach was developed by Erwin Schrödinger and Warner Heisenberg (5 December 1901 – 1 February 1976). Schrödinger formulated his version of quantum mechanics in 1925 for which he received the Nobel Prize in Physics in 1933. His equation would become to quantum mechanics what ‘\( F = ma \)’ was to Newtonian mechanics. In parallel, Heisenberg, along with Max Born and Pascual Jordan, developed the matrix formulation of quantum mechanics for which Heisenberg received the 1932 Nobel Prize. Schrödinger would later show that his and the Heisenberg formulations were equivalent.

2.2.5.1.1 The Wave Function

Quantum systems are described by a wave function. As a matter of tradition, the wave function is normally signified by the Greek letter ‘\( \psi \)’. The wave function represents solutions to Schrödinger equation. In other words, \( \psi \) represents the possible states that a quantum system can adopt at any moment in time. It contains all the information known about an atomic system. Of questions asked about a quantum system, answers can be obtained only if the information sought is contained within the system’s ‘\( \psi \)’.

Like classical mechanics, quantum mechanics consists of relationships between observable magnitudes such as ‘velocity’ and ‘position’. The primary difference between classical and quantum physics is that classical physics takes for granted that certain quantities such as a particle’s mass, velocity, acceleration etc. can be known simultaneously. But the uncertainty principle of the quantum mechanics places limitations on when and how much can be known about certain quantities. Unlike classical physics, where a series of measurements taken of a single observable gives the same value after each measurement, quantum measurements are, in general, unpredictable. It is only possible to give an average value of a series of measurements [51].

Under the current and most accepted explanation of quantum mechanics, \( \psi \) has no physical interpretation. On philosophical grounds, this seems quite paradoxical. How can such a fundamental concept within a theory of physics have no physical reality? But giving a physical interpretation to the ‘\( \psi \)’ is fraught with difficulties. Specific paths for quantum objects like electrons cannot be given. It is not possible to obtain repeatable exact measurements in every case. For these and other reasons, the inventors of quantum mechanics were pretty much driven toward the paradoxically challenged quantum wave function as a primary concept.

The uncontrollable nature of quantum processes forced its developers to work in terms of probabilities rather than real physical outcomes. Betting odds on an outcome can be given, but that’s it. At each moment a given particle is presumed to have a particular ‘\( \psi \)’. But unlike classical physics, where if the initial values are known at some point in space and time, future values are known with certainty, the quantum mechanical ‘\( \psi \)’ can only be associated with probabilities [52].
While given no physical interpretation, a valid $\psi$ possess certain characteristics; $\psi^2$ must be positive and proportional to the probability of finding, for example, a body at a given location at a given time. This removes any chance that negative probabilities will occur, since negative probabilities are meaningless. If all the values that $\psi^2$ can assume are added up, that sum must be finite. This says that an electron, for example, must inhabit a location at all times within the quantum system under consideration. All values of $\psi^2$ must be real (no complex numbers). ‘$\psi$’ must be a single valued function of location and time. It would not do, for instance, if a body could have both a 30% and a 50% chance of being at a given location at the same time. Finally, the partial derivatives of all quantum mechanical wave functions must exist. This is a mathematical statement about how smooth a wave function is required to be. How smooth is the ‘$\psi$’? Roughly speaking, its values cannot jump around randomly, but must comply with what mathematicians call ‘continuity’. Continuity is a classical concept and the observant reader will note that this requirement seems like an oddity. Are not quantum possesses by nature supposed to jump around unpredictably? Yes, but not in the manner that might be expected. Quantum processes are smooth, at least the part of the process that cannot be observed - the part that involves ‘$\psi$’. However, as soon as an interaction with a measuring device occurs, $\psi$ ceases to be a factor, and in its place, a real outcome appears. The process of changing the status of a particle by measuring it is called ‘collapsing the wave function’. It is in this sense that quantum processes jump [52]. Keep in mind, however, that although the ‘$\psi$’s’ are continuous (smooth) functions in the mathematical sense, they are not deterministic; rather, they are probability distributions.

2.2.5.1.2 The Role of the Operator in Quantum Mechanics

Nothing in the quantum theory induces a real physical outcome until and unless a measurement is taken i.e. an observation made. Until then, outcomes are described in terms of the possibilities associated with the wave function. How is the jump from the nebulous, unseen world of the wave function, to the concrete world of real physical outcomes represented in quantum mechanics? The mathematical answer to this question is: by an operator. The philosophical answer is called the ‘measurement problem’ and it is one of the outstanding issues in quantum mechanics and is mired in controversy.

The idea of an operator is simple. When presented with a mathematical object or a set of elements, an operator gives instructions on what to do to that object or to those elements. For instance, in the expression ‘$a + b$’, the ‘+’ sign gives the instruction: add the number ‘$a$’ to the number ‘$b$’. The ‘+’ sign is an example of an operator. In quantum mechanics operators act on the ‘$\psi$’ and are of a very special type, expressed by the following mathematical equation:

$$O\psi = \lambda\psi$$

Here ‘$O$’ represents the operator, $\psi$ is the wave function and $\lambda$ is a value called an ‘eigenvalue’. When the operator ($O$) acts on the wave function, it signifies that an
observation or measurement has taken place. At that point, the right hand side of the equation above shows a reproduction of the wave function multiplied by one of its eigenvalues. In essence, the operator has extracted information from the wave function. The eigenvalue represents the real outcome of a measurement.

With every physical observable (position, momentum, energy, angular momentum ...) there is an associated operator. Not all operators in quantum mechanics are created equal. It is perfectly justifiable to operate on a wave function with more than one operator - signified by 'ABψ', where 'A' and 'B' are operators. 'ABψ' signifies that B operates on ψ first followed by A. 'BAψ' signifies the opposite.

Is the order in which the operators act on the wave function significant? Interestingly, in quantum mechanics, the answer is sometimes 'yes' and sometimes 'no'. This is expressed mathematically by the commutator '[A, B] = AB - BA'. If [A, B] = 0, the order in which the operators are applied is immaterial. But if [A, B] ≠ 0, then order does matter and applying A then B would give a different physical outcome than applying B then A. What the commutator signifies is whether two observables, represented by the operators 'A' and 'B', can be measured simultaneously. If [A, B] = 0, the answer is 'yes'; otherwise, the answer is 'no'.

Why is this important? Operators express the unpredictable nature of quantum processes. Suppose a particle is in an unknown state, its momentum measured and found to be p. What if momentum is measured again? Quantum mechanics guarantees that the answer will still be p. A particle that is in an eigenstate of momentum has a definite momentum 'p' with probability '1' [52]. Suppose, however, instead of momentum, the second measurement measures the position of the particle. This time the answer becomes completely uncertain. The particle is equally likely to be in any position as in any other. A particle in a momentum eigenstate doesn't have a particular position. This is an example of the uncertainty principle, which requires that if the momentum of a particle is known exactly, then its position becomes completely indeterminate and vice-versa. Position and momentum operators in quantum mechanics do not commute i.e. [P, M] ≠ 0. It is not possible to obtain exact position and momentum measurements simultaneously [52].

Again, if momentum is measured with result 'p', the same measurement can be made 100 times successively with the same result each time. Now, if on the 101st measurement, position is measured, the result will be completely probabilistic, but a result is obtained nonetheless. The particle will be found somewhere. Now suppose on the 102nd measurement, momentum is measured again. Is it certain that the value will be 'p'? No! When the position was measured, it forced the wave function to jump to an eigenstate of position, meaning it is no longer in an eigenstate of momentum. The position measurement erases any memory of the momentum measurement. At that point, a momentum measurement becomes probabilistic [52].

What if, instead of measuring position, kinetic energy is measured and then momentum measured subsequently? In that case, the momentum measurement would still be p.
The reason is that momentum and kinetic energy have the same set of eigenstates, so a measurement of the kinetic energy doesn't erase the momentum information [52]. Two operators, momentum \( M \) and the kinetic energy operator \( K \), are examples of commuting operators (\([M, K] = 0\)). They can be in the same eigenstate simultaneously, and hence, they can be measured simultaneously.

### 2.2.5.1.3 The Wave Equation

Classical equations make deterministic predictions of the future status of physical systems. If the requisite number of initial conditions is known, classical physics predicts exactly the future state of the system. Classical systems can be divided into wave-like or particle-like phenomena. If the phenomenon under consideration is particle-like, particle-like equations describe the system. The same can be said for wave-like phenomena [24].

But, as de Broglie and others proposed, submicroscopic particles like an electron have a dual nature. There, particles display both particle and wave-like characteristics. From a classical standpoint, this creates a conceptual paradox. At the quantum level, prior to collapsing the wave function, there is no way of determining whether a particular phenomenon is wave-like or particle-like. This means that there is no way of selecting which types of equations (wave-like or particle-like) are applicable in describing the physical system of interest [24].

The solutions of Schrödinger’s equation for electron energy levels in the hydrogen atom agreed with experiment and with what the Bohr model predicted. Moreover, the equation could be extended to more complicated atoms and even to particles not bound in atoms. Despite its success, the interpretation of what these “matter waves” actually represent remains controversial. How was it possible that Schrödinger’s equation could describe both a wave, which is decentralized, spread out over an entire space, and a particle whose center is located at a single point? Schrödinger believed that the intensity of the wave at a point in space represented the “amount” of the electron that was present at that point. In other words, the electron density was spread out, rather than being localized. But this seemed to contradict what is actually experienced; particles localized.

In the final analysis, Max Born’s statistical interpretation of Schrödinger’s equation won out. He concluded that the wave associated with the electron was not a tangible ‘matter wave’, but a wave that determines the probability of finding a particle at a given location at a given time. In this sense, a particle could still be thought of as localized in space and time. In the region where this associated ‘matter wave’ had a large amplitude, the probability of finding the electron there would be high. If the amplitude was small in that region, the probability of finding the electron there would be low.

Schrödinger’s equation like Newton’s \( F = ma \) allowed for predictions, but in this case, the predictions were of a probabilistic nature. This was the only way, it seemed, to deal with the fact that an electron, for example, seemed to display both wave-like and
particle-like characteristics. The electron is described as spread out over an entire space until it interacts with a measuring device, at which point, it became localized.

But accepting the ‘Born interpretation’ meant rejecting the idea that an electron physically exists at a point in space and moves along a definite path. Instead, the electron could only be described, prior to taking a measurement, as a cloud of probabilities. Schrodinger equation is a way of describing the evolution of these probabilities in time.

The Schrödinger equation has a time dependent form used for describing progressive waves i.e. progressive probabilities. It also has a time independent form, which describes the probabilities associated with standing waves. Often called stationary waves, standing waves only allow specific wave amplitudes. For example, suppose the time independent Schrödinger equation is used to describe a simple pendulum. A pendulum can be thought of as a conserved energy system, assuming no energy is lost from the system due to friction or other external forces. On the down swing, the pendulum loses potential energy, but gains kinetic energy. On the upswing, it loses kinetic energy, but gains potential energy. This process repeats itself as the pendulum swings back and forth. Both classical and quantum pendulums conserve energy. But, described in classical terms, the kinetic energy of the pendulum can assume all possible values, including zero. But this is not the case for quantum mechanics. A quantum mechanical pendulum (usually called a harmonic oscillator) can only assume certain values, not including zero. The time independent Schrödinger equation predicts that only certain discrete energy levels materialize. An important distinction between quantum and classical physics is the discrete nature of quantum phenomena. The fact that quantum harmonic oscillators have no zero energy states is a direct consequence of the uncertainty principle [53].

Moreover, a classical pendulum is perfectly repeatable. On the upswing, it loses kinetic energy, but gains potential energy up to a maximum potential of, say, \(A\). It never exceeds \(A\). But quantum pendulums are not perfectly repeatable. Each swing of the arm may reach a level that is either more or less than \(A\). It is only possible to think of \(A\) in terms of an average.

The time-independent Schrödinger equation describes the hydrogen atom more completely than the Bohr model. The hydrogen atom is a three dimensional object, so its description requires a three dimensional time-independent Schrödinger equation. Fortunately, without much loss of generality, the three dimensional wave function can be mathematically separated into three one dimensional wave functions that can be treated individually [45]. This significantly eases the job of describing the hydrogen atom.

The first wave function predicts the probability that the electron is in one of the discrete energy levels that the electron can assume. It was shown that the Schrödinger equation predictions were in line with the Bohr model. This was a significant validation
of Schrödinger’s approach [45]. Energy levels are normally signified by the number ‘$n$’ ($n = 1, 2, 3, ...$).

The second wave function, usually signified by ‘$l$’, predicts the probability of obtaining certain values of orbital angular momentum of the electron. Angular momentum, like energy, comes in discrete quanta, which can assume only the values ‘$l = 0, 1, 2, ..., (n - 1)$’. Discrete angular momentum quanta were assumed within the Bohr model, but they are predicted by Schrödinger’s equation, further validating Schrödinger’s approach. From this, the idea of a ‘quantum state’ was spawned. Scientists began using convenient codes to describe the quantum states. For instance, the levels of angular momentum were coded with letters:

$$l = 0, 1, 2, 3, 4, 5, 6 \rightarrow 0 = s, 1 = p, 2 = d, 3 = f, 4 = g, 5 = h, 6 = i$$

For example, a ‘$2s$’ state meant that the electron was at energy level ‘two’ and had ‘zero’ orbital angular momentum. Whereas an atom with $n = 4$ and $l = 2$, was a ‘$4d$’ state [45].

The third wave function, usually signified by ‘$m_l$’, predicts the probability that the electron’s orbital angular momentum obtains a certain direction. The direction of orbital angular momentum is quantized with respect to an external magnetic field. The magnetic quantum number ‘$m_l$’ specifies the direction of $l$. The possible values are $m_l = l, ..., -l$. In other words, in the event an electron finds itself in an external magnetic field, the third wave function predicts the orientation of its orbital angular momentum. An atom in a state of principle quantum number ‘$n$’ breaks up into several sub-states when in an external magnetic field. The energies are slightly more or slightly less than the energies of the ‘$n$’ state in the absence of such a field. The splitting of the spectral lines in the presence of an external magnetic field is called the ‘Zeeman effect’, named for the Dutch physicist Pieter Zeeman (25 May 1865 – 9 October 1943) [45].

In the Bohr model of the hydrogen atom, the electron could always be found in the equatorial plane at $r = n^2r_0$, where $r_0$ represents the inner most orbit of the electron. But Schrödinger’s approach modifies the Bohr’s model in two important aspects. First, no definite values for $n, l, m_l$ could be given, but only relative probabilities for finding the electron in various states. Secondly, it became impossible to think of the electron as moving around the nucleus. Since the hydrogen atom is described by the time-independent Schrödinger equation, there is no way of describing the position of the electron as a function of time [45].

2.2.5.1.4 Electron Spin and Complex Atoms

Despite the elegance with which the three dimensional Schrödinger’s equation accounts for certain aspects of the hydrogen atom, it cannot approach an adequate description of atoms without the concept of ‘electron spin’ and the associated ‘exclusion principle’. Experiments had shown that the spectral lines associated with the hydrogen atom actually consisted of two separate but very minute lines. Referred to as the ‘fine
structure’ of the hydrogen atom, this spectral line separation is not predicted by the Schrödinger approach [45].

Secondly, the magnetic quantum number ‘\( m_l \)’ accurately explained the “normal” Zeeman effect, but many elements displayed an anomalous Zeeman effect, where more orientations for orbital angular momentum were found than Schrödinger’s theory predicted [45].

In an effort to account for both the fine structure and the anomalous Zeeman effect, S. A. Goudsmit (July 11, 1902 - December 4, 1978) and G. E. Unlenbeck (December 6, 1900 – October 31, 1988) proposed in 1925 that the electron possessed an intrinsic angular momentum independent of and in addition to any orbital angular momentum. Goudsmit and Unlenbeck had in mind the classical idea of a charged particle spinning around its axis. Since the electron was charged, the rotation produced a magnetic moment opposite in direction to its orbital angular momentum [55].

However, the classical idea in line with what Goudsmit and Unlenbeck had proposed proved to be conceptually infeasible. W. Pauli pointed out that if the electron really possessed the intrinsic spin that Goudsmit and Unlenbeck advocated, the speed of rotation would exceed the speed of light. This put the concept at odds with the theory of relativity, where nothing travels faster than light. But in 1928, P. A. M. Dirac showed on the basis of a relativistic quantum-theoretical treatment that the electron must have the intrinsic angular momentum and magnetic moment attributed to it by Goudsmit and Unlenbeck. Hence, electron spin became a purely quantum mechanical phenomenon with no classical analogue [55].

Electron spin comes in only two discrete values: \( \pm (1/2) \hbar \). The idea of ‘electron spin’ not only proved to be successful at explaining both the fine structure and the anomalous Zeeman effect, but many other atomic effects as well.

### 2.2.5.1.4.1 The Exclusion Principle

While Schrödinger’s equation could explain the electron configurations in the hydrogen atom, it failed to do so for the more complex atoms. Atoms with atomic structures differing by just one electron displayed vastly different chemical properties. While one of these atoms might be a gas, the other could be a metal. It was hard to explain why the chemical properties of atoms could change so abruptly with only a minute change in the number of electrons and protons [45].

In 1925, Wolfgang Pauli (25 April 1900 – 15 December 1958) found that each electron must have a different set of the atomic numbers ‘\( \{n, l, m_l, m_s\} \)’, where \( m_s \) is associated with electron spin. Pauli noticed that in all atoms but hydrogen certain state transitions were never found. Given that an electron had atomic numbers ‘\( n, l, m_l \)’, the forbidden transitions all had the characteristic that the electron spins could not add to 1 or −1. This meant that two electrons with the same quantum numbers ‘\( n, l, m_l \)’ could not have the same spin direction [45].
Moreover, not only electrons, but all elementary particles possess spin (some particles have spin-0). This was a monumental discovery. Ultimately, it allowed physicists to make a distinction between matter and the forces that act on matter. Matter consists of spin-1/2 particles that obey the Pauli exclusion principle. These particles are called ‘fermions’ named after the Italian-born physicist Enrico Fermi (29 September 1901 – 28 November 1954). The force carrying particles have integer spin usually 0, 1 or 2 and are called ‘bosons’ named after the Indian mathematician Satyendra Nath Bose (1 January 1894 – 4 February 1974). Bosons are not subject to the Pauli exclusion principle, and hence, can exist in the same quantum state [56].

2.2.5.1.4.2 Electron Configurations in Atoms

With the help of Schrödinger’s equation and the Pauli exclusion principle, atoms could be codified and classified into groupings with similar chemical properties. The coding took the form called a ‘subshell’:

\[ nl^e, \]

where \( n \) is the principle quantum number, \( l \) the orbital quantum number and \( e \), the number of electrons in the subshell [45]. For example,

\[ 1s^22s^23p^63s^1 \]

was the codification for a sodium atom. It has four subshells.

In 1871, Dmitri Mendeleev (February 8, 1834 - February 2, 1907) formulated an empirical law called the ‘periodic law of the elements’. Combining the periodic law with the more recent codifying of the elements eventually led to the creation of the modern day ‘periodic table of the elements’. For years, the periodic table of about 92 elements represented the fundamental building blocks of Nature. But with the development of the standard model of particle physics, those 92 “fundamental building blocks” have been reduced to about 17 [57].

2.2.5.2 Toward the Development of a Theory of Elementary Particles

By the mid 1960’s only the electron could be explained by a relatively satisfactory theory. In 1928, P. A. M. Dirac (8 August 1902 – 20 October 1984) discovered a wave equation that described elementary spin-1/2 particles, such as electrons, consistent with both the principles of quantum mechanics and the theory of special relativity. In addition, it accounted for the fine structure of the hydrogen spectrum in a rigorous way and implied the existence of a new form of matter: antimatter - later discovered experimentally. Dirac’s explanation of spin as a consequence of the union of quantum mechanics and special relativity, and the eventual discovery of the positron, represents one of the great triumphs of theoretical physics [58].
However, aspects of Dirac's theory were problematic. There was nothing to prevent an electron in a positive-energy state from decaying into a negative-energy state by emitting photons. But, experiments had shown that electrons did not act in this way [59]. To cope with the problem, Dirac envisioned the vacuum state as consisting of an infinitely large quantum state, where all negative-energy electron states are occupied. Since the Pauli exclusion principle forbids electrons from occupying the same state, any additional electron would be forced to occupy a positive-energy state. This 'vacuum state' was eventually deemed the 'Dirac Sea' [60]. When the negative-energy states were incompletely filled, each unoccupied state created "a hole" that acted like a positively charged particle. Dirac initially thought that the hole might be the proton, but Hermann Weyl pointed out that the hole should behave as if it had the same mass as an electron, whereas the proton was over 1800 times heavier. The hole was eventually identified as the positron (the antielectron), experimentally discovered in 1932 by Carl Anderson (3 September 1905 – 11 January 1991), an American physicist. He received the 1936 Nobel Prize in Physics for the discovery [45].

However, conceptually, Dirac's idea of a vacuum state implied infinitely many unobservable negative electrons filling all of space, which exactly canceled the infinity of positive energy electrons [45]. Most physicists, at the time, showed little patience for unobservable metaphysical declarations. Eventually Dirac's theory was replaced by 'quantum field theory' (\textit{QFT}), which advocated the more convincing approach of treating the positron as a real particle, rather than the absence of a particle. '\textit{QFT}' redefined the vacuum state as a state where no particles exist, instead of an infinite sea of particles. While more convincing, \textit{QFT} does not eliminate all the difficulties raised by the Dirac Sea; in particular, the problem of the vacuum state possessing infinite energy, where it becomes necessary to employ renormalization to cancel out the infinites that arise in the theory [61].

By the 1930's world developments impacted the march toward a unified theory of physics. Progress was slowed by a worldwide economic downturn. Universities found it increasingly difficult to fund research projects. Researchers often were forced to seek alternative ways of making ends meet. Collaboration plummeted, since sabbaticals were severely limited and journals found it difficult to publish on the same regular basis as they once had. Moreover, the 3\textsuperscript{rd} Reich came to power in Germany as World War II approached. Much of the previous theoretical work in physics had been accomplished by German physicists many of whom were of Jewish descent. To escape the horrors of the 3\textsuperscript{rd} Reich, significant numbers of Jewish physicists left Germany some migrating to the United States, including Einstein. Once the war broke out, the majority of physicists were recruited by their respective governments to help with the war effort. The influx of talented German physicists greatly benefited the United States. During the war significant time was devoted to studying the processes of nuclear decay, the phenomenon primarily associated with atomic explosions.

But more importantly, once the war ended, physics related activities enjoyed an influx of ready cash, particularly in the United States, as a reward for helping with the war effort. With new resources, physicists constructed research facilities. They conceptualized
and built sophisticated equipment in the form of cyclotrons and atom smashers. The result was a boon to experimental physics. Novel and creatively designed experiments led to an explosion in new particle discoveries. The new particles displayed a myriad of diverse characteristics. The avalanche of new discoveries left theorists dumbfounded. They had little choice but to gather as much information as possible, then hope, by discerning recognizable patterns in the data, that advancements would be realized, built upon earlier theoretical work.

2.2.5.2.1 The Atomic Nucleus

The Dirac theory along with the Pauli exclusion principle described atoms primarily in terms of how electrons behaved around the nucleus. Up to this point, much less was known about nuclei. By the mid 1960's, serious problems remained in understanding the properties of nuclei.

First, there was the problem of size. Atoms of a certain element were not always the same size. With only two particles, the proton and electron, known at the time, the hypothesis was that some of the protons within the nucleus contained an electron, which rendered the proton electrically neutral. Experiments had shown that in beta decay atoms spontaneously emitted electrons from the nucleus. But the amount of nuclear electron energy far exceeded the observed electron energies in beta decay. In addition, hypothesized spin configurations that included nuclear electrons did not square with what was observed. Eventually, the ‘nuclear electron’ hypothesis was abandoned [45].

Two German physicists, W. Bothe and H. Becker, bombarded beryllium with alpha particles from a sample of polonium and found that the radiation emitted did not carry a charge. They assumed the radiation emissions were gamma rays. But to produce the results that Bothe and Becker observed would have required gamma rays with energies of about 53 MeV, far exceeding any observed gamma radiation energies known at the time. Finally, in 1932, James Chadwick (20 October 1891 – 24 July 1974), an English Nobel laureate in physics, discovered the neutron. He assumed that the radiation Bothe and Becker had observed consisted of neutral particles whose mass is approximately equal to the proton. This hypothesis accounted nicely for the observed proton recoil energies – a maximum proton energy of 5.3 MeV implies a 5.3 MeV neutron energy instead of the 53 MeV required for gamma radiation to produce the same effect. Immediately upon its discovery, the neutron was recognized as the missing ingredient in describing atomic nuclei. It’s mass of slightly more than the proton, its electrical neutrality and its spin of 1/2 all fit in perfectly with the observed properties of the nuclei, when it is assumed that nuclei are made up solely of protons and neutrons [45].

The second mystery surrounding nuclei was the question of stability. The nucleus was conceived as being a tightly packed conglomeration of positively charged protons and neutrons. Since positive electrical charges repel each other, how did the nucleus stay together? Many theories were advanced, but the most intriguing was put forward by Osaka University physicist Hideki Yukawa (23 January 1907 – 8 September 1981) in
1935. Rather than a static compression of unchangeable particles, Yukawa conceived the nucleus as a teeming pot of activity, where protons were transformed into neutrons and vice versa. Yukawa theorized that protons and neutrons attract one another by exchanging mesons. In 1947, the British physicist Cecil Frank Powell (5 December 1903 – 9 August 1969) and his colleagues observed charged $\pi$-mesons confirming Yukawa’s prediction [45].

According to the meson theory of nuclear forces, all nucleons consist of identical cores surrounded by a “cloud” of one or more mesons. Mesons can have zero, ‘$+$’ or ‘$-$’ electric charge. The sole difference between a proton and a neutron was the composition of their respective meson cloud. Nuclear forces were created by the exchange of mesons between nucleons. The force between two neutrons or two protons was a result of exchanging a neutral meson ($\pi^0$). The force between a proton and a neutron was the result of exchanging charged mesons ($\pi^+, \pi^-$). A neutron that emits a $\pi^-$ is converted into a proton:

\[ n \rightarrow p + \pi^-, \]

while the absorption of a $\pi^-$ by the proton the neutron is interacting with converts it into a neutron [45]:

\[ p + \pi^- \rightarrow n \]

Yukawa was able to show that the exchange of mesons between nucleons led to mutually attractive forces. But, this led to an additional question: if nucleons constantly exchanged mesons, why were neutrons and protons never found with other than their usual masses? The emission of a meson by a nucleon, which does not change mass, is a clear violation of the conservation of energy. Yukawa explained that the non-conservation of energy would be permissible so long as the exchange of mesons occurred within the time limits dictated by the uncertainty principle [45]. With this information, Yukawa was able to predict the mass of the $\pi$-mesons. His calculations were confirmed experimentally.

2.2.5.2.2 Nuclear Decay

The process of nuclear decay comes in three broad categories signified by the names ‘alpha’, ‘beta’ and ‘gamma’. All nuclear decay involves processes so that atoms can transform themselves into a more stable state, where the decay process becomes less likely to occur over time.

When nuclei gain a mass containing 210 or more protons and neutrons, they become so large that the short-range forces holding the nucleus together are barely able to counterbalance the mutual electrostatic repulsion of the protons. At that point, Alpha particles (which are identical to a helium nucleus) are emitted from the nucleus. Alpha particles are the primary decay particles because it takes the least amount of energy to
liberate them than other combinations of particles. Alpha decay involves transforming an atom into one with two less neutrons and two less protons [45].

In 1928, George Gamow (4 March 1904 – August 19, 1968) modeled alpha decay by imagining that the nucleus acted like a well, where the alpha particle becomes trapped. Constantly jiggling, the alpha particle bounces against the well barrier. From a classical standpoint, the alpha particle would not have enough energy to escape the barrier, and hence, alpha decay should be impossible. But Gamow used quantum mechanical principles, where the amount of potential energy of an alpha particle is averaged. He reasoned that if some of the particle collisions against the barrier carried less and some more energy than the average energy, there would be a tiny probability that an alpha particle on one of its collisions against the barrier would carry enough energy to escape. Gamow derived a relationship between the half-life of alpha decay and the energy of the alpha particle emissions previously discovered empirically by Geiger and Nuttall. His explanation of alpha decay is one of the most striking conformations of the correctness of quantum mechanics over classical physics [45,64].

Beta decay, like alpha decay, is a means by which atoms gain greater stability. Why beta decay occurs is still somewhat mysterious, but appears related to the ratio of the number of neutrons to the number of protons. If that ratio grows much above 1, the probability that an atom undergoes beta decay increases.

When first observed, beta decay displayed a couple of troubling characteristics. During beta decay an electron is emitted from the nucleus. But the nucleus was not supposed to contain electrons. The problem was overcome by what amounts to a sleight of hand. It was assumed that the electron leaves the nucleus immediately after its creation. Far more troubling was the observation that beta decay violated the conservation laws of energy, momentum and angular momentum [45].

Beta decay involves the conversion of a neutron into a proton. Symbolically, this can be represented by

\[ n \rightarrow p + e^- \]

In 1930, W. Pauli suggested that if an uncharged particle of small or zero mass and spin-1/2 is emitted in beta decay along with the electron, then the issue that arose with the conservation laws disappeared. The hypothesized particle resolved the discrepancy between the energy of the electron and the recoiling daughter nucleus so that the conservation laws held. There were actually two hypothesized particles: the ‘neutrino’ and ‘antineutrino’, designated by the symbols ‘\( \nu \)’ and ‘\( \bar{\nu} \)’ respectfully. This also meant that there were two types of beta decay: positive and negative. The formulas that included the new particles could then be written

\[ n \rightarrow p + e^- + \bar{\nu} \]

\[ p \rightarrow n + e^+ + \nu \]
The first formula signifies negative beta decay, where a neutron decays into a proton, electron and an antineutrino, whereas, in positive beta decay, a proton decays into a neutron, positron and a neutrino. The neutrino was not discovered until 1959. Its discovery was difficult because it is electrically neutral, so it does not interact with charged particles. It also possesses either a very small or no mass. So feeble is its interaction with matter, it must pass through about 130 light-years of solid iron on average before an interaction with other matter takes place [45].

The standard model of particle physics treats neutrinos as particles with zero mass. But more recent experiments have indicated that neutrinos have a small but non-zero mass. Neutrinos have become an important topic because a neutrino with a non-zero mass suggests new physics beyond the standard model. It is believed that neutrinos were produced in great abundance in the early Universe, so there are a lot of them. Their tiny mass but huge numbers are thought to contribute to the total mass of the Universe and might have an impact on its expansion [65,66,67].

Like whole atoms, the nucleus can be in an excited state. Excited nuclei return to their lowest energy state by emitting photons whose energies correspond to the energy differences between the excited state and the lowest energy state. Such emitted photons are called ‘gamma rays’ and the process is referred to as ‘gamma decay’. Gamma rays are produced after the occurrence of other types of decay. The mechanism involves the nucleus emitting an α or β particle, so that the daughter nucleus is usually left in an excited state. It can then move to a lower energy state by emitting a gamma ray in much the same way that an electron can jump to a lower energy state by emitting a photon [45].

By the 1930’s the alpha, beta and gamma decay processes had all been observed experimentally and explained individually; but no comprehensive theory of radioactive decay was formulated until the development of the electro-weak theory in the 1970’s.

2.2.5.2.3 The Particle Zoo

By the mid 1960’s about 30 varieties of relatively stable particles had been discovered along with a myriad of unstable particles. Practically speaking, “relatively stable” meant that particle interactions could be detected in a bubble chamber, whereas unstable particle interactions were so fleeting as to be nearly impossible or impossible to observe. In an attempt to find an all-encompassing theory, the particles were organized and classified in some kind of meaningful fashion. One such organized display appears in Table 2.2.5.2.3-1.

---

<table>
<thead>
<tr>
<th>Class</th>
<th>Particle</th>
<th>Antiparticle</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Name</td>
<td>+e</td>
</tr>
<tr>
<td>Photon</td>
<td>Photon</td>
<td>Y</td>
</tr>
<tr>
<td>Lepton</td>
<td>e-</td>
<td>( \nu_e )</td>
</tr>
<tr>
<td></td>
<td>( \mu^- )-neutrino</td>
<td>( \nu_\mu )</td>
</tr>
<tr>
<td></td>
<td>Electron</td>
<td>e</td>
</tr>
<tr>
<td></td>
<td>( \mu^- )-meson</td>
<td>( \mu^- )</td>
</tr>
<tr>
<td>Meson</td>
<td>( \pi^- )-meson</td>
<td>( \pi^- )</td>
</tr>
<tr>
<td></td>
<td>K-meson</td>
<td>K*</td>
</tr>
<tr>
<td></td>
<td>( \eta^- )-meson</td>
<td>( \eta^- )</td>
</tr>
<tr>
<td>Baryon</td>
<td>Nucleon</td>
<td>p</td>
</tr>
<tr>
<td></td>
<td>n</td>
<td>( \bar{n} )</td>
</tr>
<tr>
<td></td>
<td>( \Lambda^- )-hyperon</td>
<td>( \Lambda^- )</td>
</tr>
<tr>
<td></td>
<td>( \Sigma^- )-hyperon</td>
<td>( \Sigma^- )</td>
</tr>
<tr>
<td></td>
<td>( \Xi^- )-hyperon</td>
<td>( \Xi^- )</td>
</tr>
<tr>
<td></td>
<td>( \Omega^- )-hyperon</td>
<td>( \Omega^- )</td>
</tr>
</tbody>
</table>

Elementary Particles Stable against Decay by the Strong Interaction

Table 2.2.5.2.3-1

While there are alternative ways of organizing the data, the manner in which information is organized in table 2.2.5.2.3-1 should give an appreciation for the difficulty of building a theory of elementary particles from what, for all intents and purposes, is a hodgepodge of disparate data. All the particles have an antiparticle except for the photon (\( \gamma \)) and the \( \pi^0 \)-meson. When a particle interacts with its antiparticle, the result is usually complete inhalation of both particles, but this is not always the case. If neutrons or protons interact with their antiparticles, \( \pi^- \)-mesons are usually produced. This provided evidence that \( \pi^- \)-mesons might be a mediator of the nuclear force.

Prior to 1956, the neutrino and antineutrino were thought to be the same particle. Neutrinos carried no electric charge, possessed no mass and differed only in spin direction. At the time, it was thought to be immaterial whether physical processes were right-handed or left-handed. This “law of Leibniz” was a universally accepted principle referred to as ‘spatial parity’. And this was exactly what the neutrino and antineutrino seemed to be – mirror images of one another, and hence, identical particles. The law of Leibniz seemed to hold just fine for nuclear and electromagnetic phenomena. But T. D. Lee and C. N. Yang noticed that many serious theoretical discrepancies would be removed if neutrinos and antineutrinos had different handedness, even though that
meant they could not be reflected in a mirror. Lee’s and Yang’s conjectures were later experimentally confirmed. Lack of right-left symmetry in neutrinos can only occur if the neutrino mass is exactly zero [45]. In the standard model, neutrinos are treated as having zero mass. However, current experiments have shown that the neutrino has a small, but non-zero mass. The issue of neutrino mass remains unresolved and suggests new physics beyond the standard model.

The $\pi$-mesons were believed to be the mediators of the nuclear force. But free $\pi$-mesons were discovered in the diffuse streams of cosmic radiation that bombards the Earth [45]. Where did free $\pi$-mesons come from? Current theory conjectures that they must have been produced in abundance during the big bang. In addition, charged $\pi$-mesons almost inevitably decay into lighter $\mu$-mesons and neutrinos. Remarkably, the neutrinos found in $\pi$-meson decay are not the same as those found in beta decay. The existence of two classes of neutrinos was established in 1962.

The discovery of the $\mu$-meson is still something of a puzzle. Its characteristics are well known: rest mass ‘207’, spin-1/2, half-life ‘2.26 × 10^{-6}’. The $\mu^+$-meson decays into a positron, a neutrino and an antineutrino. The $\mu^-$-meson decays into an electron, a neutrino and an antineutrino. The mysterious part of the $\mu$-meson is that it does not seem to have a function. It interacts only electrostatically with matter. Except for its mass and half-life, it is just a heavy electron. Strangely, about 99% of the time $\pi$-mesons decay into $\mu$-mesons and only 1% of the time directly into electrons. Since $\mu$-mesons decay into electrons and positrons, the question is: why is there this intermediate step?

Finally, there were the $K$-mesons. They are relatively fleeting particles, which decay into $\pi$-mesons, $\mu$-mesons, electrons/positrons and neutrinos in various mixtures. They are capable of both nuclear and electromagnetic interactions, but interact weakly with nuclei [45].

The final class of particles is the ‘baryons’. It consists of the neutron, proton and the hyperons. Hyperons are particles heavier than the proton and they come in four subclasses in order of increasing mass. All are fairly fleeting with extremely short half-lives. Their characteristics can be found in Table 2.2.5.2.3-1. All hyperons have definite interactions with nuclei.

2.2.5.2.4 The Systematics of the Particle Zoo

Despite the vast number of elementary particles and the diversity in their characteristics, from all that seemed like chaos, there emerged an underlying structure. The structure itself did not constitute a theory of elementary particles, but it did provide hope that one day a complete theoretical picture would emerge [45]. Ultimately the structure became part of the standard model of particle physics.

By this time, the belief that particle interactions were mediated through an exchange of particles was already well established. The photon was in a class by itself. Its place as
the force carrier of the electromagnetic interaction was well entrenched. But the other three force carrying particles that would ultimately become part of the standard model ($W$, $Z$, $g$ gauge bosons) had yet to be hypothesized or discovered. Interestingly, speculation arose that a massless, stable spin-2 particle called ‘the graviton’ was the mediator of gravitational force. But, the graviton failed to make it into the standard model and has yet to be discovered.

It was suspected that the $\pi$-mesons were the mediators of the nuclear force, but nuclear interactions were the least understood at the time and, unlike the photon, $\pi$-mesons were not massless. Hence, at this point, mesons were not classified as force carriers. It was well-known that if energy conservation was to be preserved, the uncertainty principle required that the range of a force be inversely proportional to the mass of the exchanging particles. The electromagnetic and gravitational forces were thought to have infinite range, and therefore, their interchanging particles had zero mass [45].

The next class of particles was the ‘leptons’ or ‘small particles’, all with spin of 1/2. Table 2.2.5.2.3-1 shows them classified pretty much the way they appear in the standard model less the tau-neutrino and tau-particle. Then there are the mesons with non-zero masses and zero spin. The final class, the baryons, included the nucleons (protons and neutrons) and the hyperons. Hyperons are heavy particles, heavier than the proton. All baryons have spin of 1/2. Eventually, the mesons and baryons would be replaced by the ‘quarks’ in the standard model, but at this point in time, the existence of quarks had yet to be established.

Introducing the new quantum numbers ‘$L$, $M$ and $B$’ did not cover all circumstances. It was not clear why some heavy particles decayed into lighter ones with the emission of a gamma ray, while other seemingly equally permissible decays did not occur. For instance, the ‘$\Sigma^0$’ baryon decayed into a ‘$\Lambda^0$’ baryon and a gamma ray, but the ‘$\Sigma^+$’ baryon never decayed into a proton and a gamma ray. The question was: why not? Moreover, these heavier, soon to be named ‘strange particles’, were never created singly, but always came in combinations of two or more at a time. To deal with these
situations, a new quantum number called ‘strangeness’ ($S$) was introduced. In all electromagnetic and strong interactions, $S$ is conserved. This explained why certain baryon decays occurred while others did not. Only in weak interactions did $S$ change, but even then, changed only by $+1$ or $-1$ in any given decay. It was also convenient to introduce a quantity called the ‘hypercharge’ $'Y$ ($Y = S + B$’). $'Y'$ was conserved in all strong interactions [45].

Looking at Table 2.2.5.2.3-1, there are a number of particle families whose members have the same mass and interaction properties, but a different charge. These families can be thought of as being a single entity, but in a different charge state. One way of categorizing these families is by assigning the family a number, call it $'I'$, where the number of charge states is given by the formula $'2I + 1'$. Hence, for the nucleon, $I = 1/2$, so that $2I + 1 = 2$. The two states are the proton and the neutron. The $\pi$-meson family has $I = 1$, so $2I + 1 = 3$. The three charge states are the $\pi^+$, $\pi^0$ and $\pi^-$. There appeared to be an analogy between these charged states and the quantum number for angular momentum, so the number $'I'$ was rather inappropriately named ‘isotropic spin’, although it had nothing to do with the spin state of a particle. The analogy to quantum angular momentum was carried further by designating a new quantum number $'I_3'$.

$'I_3'$ is restricted to the values $I, I - 1, ..., 0, ..., -(I - 1), -I$, so $I_3$ takes on half-integer values if $I$ is a half-integer and integer values if $I$ is an integer. For the nucleon, $I = 1/2$, and hence, $I_3 = 1/2$ or $-1/2$; the former is taken to represent the proton and the latter the neutron. In the case of $\pi$-mesons, $I = 1$, so $I_3 = 1$ corresponds to $\pi^+$, $I_3 = 0$ to $\pi^0$ and $I_3 = -1$ to $\pi^-$. The charge of a meson or baryon is related to its hypercharge $'Y'$ and $'I_3'$ by the formula

$$q = e \left( \frac{I_3 + Y}{2} \right) = e \left( I_3 + \frac{S + B}{2} \right)$$

Hence, in the case of the nucleons, the proton has $I_3 = 1/2$, $B = 1$ and $S = 0$, so $q = e$, the charge on an electron. For the neutron, $I_3 = -1/2$, $B = 1$ and $S = 0$, so $q = 0$. The other meson and baryon charge states can be calculated in a similar fashion. Charge and the baryon number $'B'$ are conserved in all interactions. Thus, $I_3$ is conserved whenever $S$ is, namely, in strong and electromagnetic, but not in weak interactions [45].

Such were the systematics of the particle zoo. But despite an abundance of regularities, a complete theory of elementary particles had yet to be accomplished. The development of the standard model, which represents a complete system of elementary particles, will be discussed in the next chapter.
Chapter 3

The Standard Model and Beyond

“It is often stated that of all the theories proposed in this century, the silliest is quantum theory. In fact, some say that the only thing that quantum theory has going for it is that it is unquestionably correct.”

— Michio Kaku, Hyperspace

3.0 Introduction

Why is symmetry important in physics? A significant appreciation for the relationship between symmetry and conservation laws was gained through Noether’s theorem, which states that, for each continuous symmetry, there is an associated conservation law. By the mid 1960’s, these symmetry/conservation relationships were well understood (see table 3-1).

<table>
<thead>
<tr>
<th>Symmetry Operation</th>
<th>Conserved Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>All interactions are independent of:</td>
<td></td>
</tr>
<tr>
<td>Translation in Space</td>
<td>Linear Momentum</td>
</tr>
<tr>
<td>Translation in Time</td>
<td>Energy</td>
</tr>
<tr>
<td>Rotation in Space</td>
<td>Angular Momentum</td>
</tr>
<tr>
<td>Electromagnetic Gauge Transformation</td>
<td>Electric Charge</td>
</tr>
<tr>
<td>Lorentz Transformation</td>
<td>Velocity of Center of Mass</td>
</tr>
<tr>
<td>Interchange of Identical Particles</td>
<td>SPIN Statistics</td>
</tr>
<tr>
<td>Inversion of Space, Time and Charge</td>
<td>CPT Parity</td>
</tr>
<tr>
<td>?</td>
<td>Baryon Number (B)</td>
</tr>
<tr>
<td>?</td>
<td>Lepton Number (L)</td>
</tr>
<tr>
<td>?</td>
<td>Meson Number (M)</td>
</tr>
<tr>
<td>The strong and electromagnetic interactions are independent of:</td>
<td></td>
</tr>
<tr>
<td>Inversion of Space</td>
<td>Parity</td>
</tr>
<tr>
<td>Reflection of Charge</td>
<td>Charge parity, $I_3$, Strangeness (S)</td>
</tr>
<tr>
<td>The strong interaction is independent of:</td>
<td></td>
</tr>
<tr>
<td>Charge</td>
<td>Isotropic spin ($I$)</td>
</tr>
</tbody>
</table>

Some Symmetry Operations and their Associated Conservations Laws

Table 3-1

---

Beneath the systematics of elementary particles lay various conservation laws. Without them, keeping track of the endless number of characteristics in each system would be virtually impossible. By concentrating on what remains unchanged, many of the variables that make up complex systems can be ignored [56]. It was the connection between symmetry and conservation laws that ultimately led to the complete development of a modern theory of elementary particles.

3.1 Group Theory: The Road to Understanding Elementary Particles

The information in table 3-1 gives an overview of symmetry from the physics standpoint. Another way of looking at symmetry is from a mathematical point of view. Mathematicians study symmetry under the name ‘group theory’. A group is an abstract algebraic system. It was not at all clear that the mathematical abstractions embodied in group theory would have any connection to the physical world. Even Einstein did not recognize that his special and general theories of relativity were studies in symmetry. But, in 1931, Herman Weyl (9 November 1885 – 8 December 1955) realized that group theory could be applied to quantum physics.

Algebraically, a group is described as a set of elements ‘\( G = \{a, b, c, \ldots \} \)’ together with an operation or binary composition rule signified by ‘\( \circ \)’. The composition rule is customarily omitted in most treatments of the subject. The elements, in conjunction with the composition rule, create an algebraic system having the following properties:

1. Closure: the composition ‘\( a \circ b \)’ of any two elements in \( G \), is itself an element of \( G \) i.e. \( a \circ b \in G \).

2. Associativity: for any of the elements ‘\( a, b, c \ldots \)’ in \( G \),
\[
(a \circ b) \circ c = a \circ (b \circ c).
\]

3. Identity: there exists a unique element in \( G \), called the ‘unit’ or ‘identity’ denoted ‘\( e \)’, such that, for every element in \( G \), \( a \circ e = e \circ a = a \). Therefore, \( b \circ e = e \circ b = b \) and \( c \circ e = e \circ c = c \) and so on.

4. Inverse: for every element in \( G \), there is a unique element called its ‘inverse’, also in \( G \). Hence, \( a \circ a^{-1} = e \) and \( b \circ b^{-1} = e \) and so on, where ‘\( a^{-1} \)’ signifies the inverse of \( a \) and ‘\( b^{-1} \)’ signifies the inverse of \( b \).

The closure property ensures that the binary composition rule ‘\( \circ \)’ does not generate elements outside \( G \). Associativity makes the composition rule unambiguous. The composition \( a \circ b \circ c \) is unambiguous because the two interpretations allowed by the existence of a binary composition rule, ‘\( (a \circ b) \circ c \)’ and ‘\( a \circ (b \circ c) \)’, are equal [68].

Although ‘\( \circ \)’ could correspond to a numerical operation like multiplication or addition, this is not required. The composition rule is an abstract rule, which combines a pair of group elements to obtain a third group element. Whereas ordinary numbers commute

60
i.e. \( a \circ b = b \circ a \), the composition rule need not be commutative, i.e. \( a \circ b \) may not equal \( b \circ a \).

Example: does the set ‘\( Z = \{ ..., -3, -2, -1, 0, 1, 2, 3, ... \} \)’, with composition rule ordinary addition, form a group?: 1) the sum of any two integers is an integer, ensuring closure; 2) the addition of integers is associative; 3) ‘0’ is the identity element, since \( n + 0 = n \), \( n \in Z \); 4) the inverse of \( n \) (\( n \) signifies any number in the set ‘\( Z \)’) is \( -n \), since \( n + (-n) = 0 \), \( -n \) is an integer and \( -n \) is a member of \( Z \). Thus, all four group conditions are satisfied. Hence, the integers form a group under addition [68].

But, does the set ‘\( Z \)’, with the composition rule ordinary multiplication, form a group? The product of any two integers is an integer (closure is satisfied), multiplication is associative, the unit element is 1, since \( n \cdot 1 = n \), but the inverse of \( n \) is \( 1/n \), which, in general, is not an integer. The last group condition i.e. that all the elements of the set have a unique inverse is not satisfied, and hence, the integers under multiplication do not form a group [68].

It is certainly not obvious, at least at this point, that the group structure has anything to do with the symmetry of objects like people, atoms or even idealized objects like triangles or circles. But, somewhat surprisingly, it does. To see this, recall how symmetry is defined. An object is said to be ‘symmetric’ if there exists a change in the object, usually referred to as a ‘transformation’, such as a rotation or reflection, whereby the object looks the same after the transformation as it did before the transformation. For example, an equilateral triangle is indistinguishable after rotations of \( 2/3 \pi \) and \( 4/3 \pi \) radians around its geometric center or symmetry axis. A square is indistinguishable after rotations of \( 1/2 \pi, \pi, 3/2 \pi \) and \( 2\pi \) radians, and the circle is indistinguishable after all rotations around its symmetry axis, which happens to be its center. These transformations are referred to as ‘symmetry transformations’. This idea is elucidated by saying that the object remains ‘invariant’ under a symmetry transformation [68].

The equilateral triangle and the square are examples of discrete symmetries i.e. the symmetry rotation angles and reflections have only certain values. For instance, there are only certain degrees of rotation for which the square remains invariant. However, the symmetries of a circle are continuous. It can be rotated by any amount around its axis of symmetry and would remain invariant. In the physical sciences, symmetry is fundamental because there are transformations which leave the laws of physics invariant. Such transformations involve changing the variables within a physical law such that the equations describing the law retain their form when expressed in terms of the new variables [68].

To understand this better, consider the notion of a Galilean transformation, named after Galileo Galilei (5 February 1564 – 8 January 1642), which says that the laws of physics are invariant whether an experiment is conducted in a stationary laboratory or in a laboratory travelling at a constant speed. The physics governing a glass of wine are the same whether it is stationary or on a train moving at 200 miles per hour. To put the
Galilean transformation in terms of symmetry: the laws of physics remain symmetrical (invariant) if a transformation is made from a stationary state to one moving with a constant velocity.

During Newton’s time and long after, Galilean symmetry was considered a law of Nature. But later it was discovered that Maxwell’s theory of electromagnetism seemed to violate Galilean symmetry. Experiments with light showed that an experimenter actually could tell the difference between standing still and moving at a constant speed. One plausible explanation was that light traveled through a luminous aether. The aether, considered a substance, like water that waves travel through, could have accounted for the discrepancy between Newton’s and Maxwell’s theory. But when appropriately designed experiments failed to detect the aether, the contradictions between the two theories remained.

Galilean symmetry was saved by Einstein. He postulated that light travelled at a constant finite velocity as Maxwell’s theory claimed, not at an infinite velocity as Newton’s theory demanded. As a result, the Galilean law did not just involve a spatial transformation i.e. transforming from a stationary state to one moving with a constant velocity, but time had to be taken into account as well. Einstein turned the Galilean transformation from a transformation in spatial position only to a transformation in space-time. If time was taken into account, the laws of physics remained invariant under space-time transformations. This removed the contradiction between the theories of Newton and Maxwell. Instead of giving an observer’s location in terms of three spatial coordinates ‘(x, y, z)’, the time on each observers watch had to be specified i.e. ‘(x, y, z, t)’. Einstein’s equation for representing this idea was

\[-x^2 - y^2 - z^2 + (ct)^2 = -x'^2 - y'^2 - z'^2 + (ct')^2\]

The arrangement of minus and plus signs are reflective of the fact that nothing travels faster than light. The coordinates ‘(x’, y’, z’, t’)’ usually denote the stationary observer while ‘(x, y, z, t)’ represents an observer moving at a constant velocity. The symmetry appears in the sense that (x’, y’, z’, t’) can be replaced by (x, y, z, t) and (x, y, z, t) replaced by (x’, y’, z’, t’) and the equation would remain invariant.

The physical interpretation of this is that the two observers, one moving at a constant speed relative to a given clock and the other stationary to it, may not agree on the location and time of an event, but will agree on the space-time interval represented by the equation above. It is in this sense that the space-time interval is said to be ‘invariant’.

A permutation is an arrangement of n distinct objects. There is a way of counting the number of permutations. It is simply n! or n(n – 1)(n – 2) … 1. For example, if there are 3 distinct objects, then there are 3 · 2 · 1 = 6 permutations or 6 ways of arranging the 3 objects. As it turns out, when combined with a composition rule, these n distinct object arrangements form a group, which is usually denoted by ‘Sn’. The group composition rule, in this case, is the permutation of the objects.
Quite remarkably, the somewhat abstract permutation group ‘$S_3$’ is identical to the symmetry transformations of an equilateral triangle. To see this, the 6 elements of $S_3$ are listed below:

$$
e = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix} \quad a = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix} \quad b = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix}$$

$$
c = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix} \quad d = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix} \quad f = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}$$

In this notation, the top line represents the initial or reference order of the objects and the bottom line represents the effect of the permutation. The composition rule corresponds to performing successive permutations and is carried out by rearranging the objects according to the first permutation and then using this as the reference order to rearrange the objects according to the second permutation. So, in this case, the composition rule simply instructs that one permutation follows the other. As an example, consider the composition ‘$ad$’, where permutation ‘$d$’ is performed first, followed by permutation ‘$a$’. Element ‘$d$’ permutes the initial order ‘$(1 \ 2 \ 3) \to (3 \ 1 \ 2)$’. Element ‘$a$’ then permutes this by putting the first object, in this case 3, in the second position, the second object, in this case 1, into the first position, and leaves the third object in position three, i.e. $(3 \ 1 \ 2) \to (1 \ 3 \ 2)$. Hence,

$$ad = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix} = b$$

It is only the permutation of the distinct objects, not the label, which is important for specifying the arrangement. Analogously,

$$da = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix} = c$$

Since $ad \neq da$, unlike ordinary numbers, permutations do not commute. If they did, the group would be called an ‘Abelian’ group. Hence, $S_3$ is called a ‘non-Abelian’ group.

Recall that the set ‘$Z$’ (the set of integers under addition) formed a group. Since ordinary numbers commute under addition, the group ‘$(Z, +)$’ is an Abelian group.

To show that $S_3$ is, in fact, a group, examine the ‘group product table’ shown in table 3.1-1. The table itself is somewhat misnamed, since it has nothing to do with the arithmetical product of two numbers, but simply displays all the possible permutations of $S_3$. 
Table 3.1-1

For instance, the table shows how to find the effect of applying the permutation ‘$b$’ first and then ‘$a$’ i.e. $ab$ by finding $b$ in the first column and then looking down that row until reaching the ‘$a$’ column. The result is: $abe = ab = 312 = d$. If $a$ acts first, then $b$, the results is: $bae = ba = 231 = f$. It is easy to see that any composition of the elements results in an element of $S_3$, satisfying the closure property of a group. Although tedious, associativity can be demonstrated. The identity element is $e$, as can be seen by looking along the first row and down the first column. In this particular group, each element is its own inverse: $aa = e$, $bb = e$ and so on. This can be verified by looking left to right diagonally.

A geometric realization of $S_3$ can be established by considering the symmetry transformations of an equilateral triangle (fig. 3.1-2). The elements ‘$a$, $b$, $c$’ correspond to reflections through lines which intersect the vertices at 3, 1, and 2 respectively, and $d$ and $f$ correspond to clockwise rotations of $2/3 \pi$ and $4/3 \pi$ radians respectively. The effect of the transformations on the positions of the vertices of the triangle is identical with the corresponding elements of $S_3$. There is a one-to-one correspondence between the triangle transformations and the elements of $S_3$. For the equilateral triangle, the composition rule ‘$ad$’ corresponds to a rotation followed by a reflection.
Thus, beginning with the standard order for the identity, the successive application of these transformations is shown below:

First, $d$ acts on $e$, which is a rotation on the original triangle. Then, $a$ acts on $d$, which is a reflection resulting in $b$. Similarly, $da = c$. In fact, all the compositions in $S_3$ are identical to the symmetry transformations of the equilateral triangle. If two groups have the same algebraic structure, they are, for all intent and purposes, identical [68].

What is the upshot of this discussion on groups? There are two main points. First, the group concept, where an element acts on another element to produce third element of the group, is very similar to the way processes are described within quantum mechanics, particularly, quantum field theory. Further along, quantum field theory will employ creation and annihilation operators that act on particles to create other arrangements of particles. The mathematical way of describing this is very similar to group operations.

Seemingly, group theory is simply a purely abstract mathematical structure. A little less abstractly, group theory can describe the permutations of a set of objects. Even less abstractly, group theory can describe the symmetry of a particular object, in this case, the equilateral triangle. The next step in the progression is to see if group theory can be applied to objects in the real world, like atoms and molecules. Quantum field theory will answer this question in the affirmative.

This last point is important. Taking an abstract structure, a pure creation of the human mind, then applying it to the real world, allows, in some sense, an understanding of how the world actually works. It must be assumed that the abstract structure is, at least potentially, understandable to us, since the human mind created it. Imagine if this was not the case. Finding it impossible to put our ideas into one-to-one correspondence with the structure of the world would preclude us from understanding it. The laws of Nature, whatever they might be, would be beyond comprehension, which, as a matter of fact, they might be. This is, to a large extent, the crux of the problem in physics today, which suffers from an inability to combine two competing seemingly incompatible theories. The theory of relativity makes perfectly good mathematical sense. It is based on ordinary numbers and the differential calculus. Both sit on a fairly firm mathematical
foundation. The quantum theory, however, is seemingly based on an entirely different set of rules. In fact, quantum field theory is based on Yang-Mills theory. And Yang-Mills theory has yet to be given a firm mathematical foundation. This logical void has become so troubling that the Clay Mathematics Institute is offering a million dollar prize to anyone who can put Yang-Mills theory on a firm mathematical foundation.

### 3.1.1 The Symmetries of Nature

To gain a better understanding as to why modern physics revolves around the idea of a ‘symmetry group’, a few words should be devoted to the kinds of symmetries found in Nature. One of the simplest is translation in space. Systems are generally described in terms of spatial coordinates. An event takes place at a certain location. Mathematically, location is described by a set of coordinates, say \((x, y, z)\), where the location is measured from a point, usually designated ‘\((0,0,0)\)’ called the ‘origin’. The symmetry of translation in space simply states that the laws of physics are independent of where in space the origin is chosen. This symmetry leads directly to the conservation of linear momentum. In quantum field theory, the number of particles is not conserved in particle interactions. There may be more or less particles prior to an interaction than afterwards. But momentum is conserved. Particles, prior to an interaction, have the same total momentum as they do after the interaction, although the particles may inhabit different positions in space.

Interestingly, time translation leads to the conservation of energy. This symmetry says that it doesn’t matter what time an experiment is started, the amount of energy at the beginning of the experiment will equal the amount of energy at the end of the experiment. This is a key feature of Einstein’s general relativity, where energy can flow around as time passes, but is never created nor destroyed [56].

Invariance under rotations in space means that the laws of physics are independent of the orientation of the coordinate system in which they are expressed. This symmetry leads to the conservation angular momentum [45].

Electromagnetism is governed by gauge symmetry. Gauge symmetry refers to shifts in the electric and magnetic field potentials. The potentials describe the electric and magnetic fields around an electromagnetic source – like a magnet. There are various ways of describing the potentials, but the strength of the electric and magnetic field remains invariant under different choices for the potentials. This gauge symmetry leads directly to the conservation of electric charge. Changes in the electric or magnetic field potentials do not change the electric charge. Closely related to gauge symmetry is the isotropy of space-time discussed earlier. The laws of physics remain invariant under space-time transformations [69].

In quantum mechanics, particles are described by a symmetry that preserves the nature of the wave function. This important symmetry is used to distinguish the force carrying particles from the matter particles [56]. Wave functions that obey the Bose-Einstein statistics are bosons – the force carrying particles. All bosons have integer spin and do not conform to the Pauli exclusion principle. If, in a system of particles, two identical
bosons are exchanged, their associated wave function remains unchanged. On the other hand, wave functions that obey the 'Fermi-Dirac statistics' are called 'fermions'. They have half-integer spin and obey the Pauli exclusion principle. If two identical fermions within a system are exchanged, their associated wave function changes sign i.e. \( \psi(x, t) \rightarrow -\psi(x, t) \). The 'SPIN Statistics' theorem states: no process occurring in an isolated system can change the statistical behavior of that system. In other words, in any isolated system of elementary particles, a boson cannot be changed into a fermion and vice versa. Therefore, the nature of the two types of elementary particles (fermions and bosons) is conserved in any process in Nature \[45\].

The symmetries associated with the conservation of the quantum numbers \( B, L, M \) discussed earlier, remained unknown during much of the 1960s. This deficiency was corrected with the invention of the quark model and the introduction of the electro-weak theory in the 1970's. The quark model was presented in 1964 by physicists Murray Gell-Mann and George Zweig shortly after Gell-Mann's 1961 formulation of a particle classification scheme known as the 'Eightfold Way' — or, in more technical terms, 'SU(3) flavor symmetry'. Yuval Ne'eman independently developed a scheme similar to the Eightfold Way in the same year \[70,71,72,73,74,75,76\].

The remaining symmetries involve parities of one kind or another. Roughly speaking, parity can be thought of as mirror image symmetry. From a physics standpoint, parity refers to the behavior of the wave function under inversion of space. If \( \psi(x, t) = \psi(-x, t) \), the wave function has even parity. If \( \psi(x, t) = -\psi(-x, t) \), the wave function has odd parity. Whether even or odd, a system retains its parity throughout an interaction. As was learned earlier, \( \psi(x, t) \) has no physical interpretation in quantum mechanics. Only \( \psi^2(x, t) \) is related to a real outcome, and then, only to its probability. Hence, real outcomes are independent of parity.

So why is parity important? Parity can actually limit the number of physical outcomes that can occur in an isolated system. It is possible to describe systems that have a certain parity in an initial state and then predict interactions with equally plausible final states that do not occur in Nature. The conservation law is the conservation of parity, which states that whether an isolated system is described by a left-handed or right-handed coordinate system, the laws of physics will remain the same \[45\]. Interestingly, there are particles in Nature that violate spatial parity (\( P \)). Evidently, a neutrino cannot be described using a right-handed coordinate system nor can an anti-neutrino be described using a left-handed coordinate system. Neutrinos partake of the weak interaction, so spatial parity is not conserved in weak interactions, but only in the strong and electromagnet interactions.

There are two more important parities: time (\( T' \)) and charge (\( C \)). Time parity says that \( \psi(x, t) \) does not change if \( t \) is replaced by \(-t\). This symmetry is called 'time reversal': if a process can occur in Nature, its reverse occurs. It was believed for a long time that \( T' \) was a symmetry of Nature. But in 1964, the \( K_2^0 \) meson was found to decay into a '\( \pi^+ \)' and a '\( \pi^- \)', which violated the conservation of \( T' \).
Charge conjugation ‘C’ is the symmetry associated with replacing all the particles of a system with their anti-particles. This means that for any process that takes place with particles under the influence of the strong or electromagnetic interactions, there exists an identical process for their antiparticles. But C symmetry fails in weak interactions. However, if all three parities are taken together (CPT), then this does appear to be a symmetry of Nature: for every process, an antimatter mirror-image counterpart takes place in reverse. And this happens even though each parity fails individually during some interactions [45].

There are two supreme challenges associated with constructing a unified theory of physics based on symmetry. Firstly, not all interactions exhibit the same symmetries. For instance, the strong interaction is independent of isotopic spin, but this is not the case with the other interactions. Secondly, Nature is not entirely symmetrical. The symmetries are, in a sense, hidden. It is the hidden world of the wave function that is symmetrical. When an observation or a measurement is taken, the wave function ceases to be a factor; the symmetries disappear and leave only an asymmetrical result. The weak interaction involves the concept of ‘symmetry breaking’, which explains how all massive particles and, in particular, W and Z bosons gain mass.

3.2 The Standard Model of Particle Physics

The search for a theory of elementary particles culminated in the creation of the standard model of particle physics. It is not a completely predictive theory. Its predictive power requires experimental input as well as theoretical insight. The standard model is considered a hybrid: part experimentally derived inputs, part theory. But once the input parameters are known, the standard model makes unassailable predictions.

With that in mind, what constitutes the standard model?
Table 3.2-1 shows the current elementary particle classification scheme within the standard model. There are 17 known fundamental particles. One of them, the Higgs boson, is not shown. Its discovery was accomplished in the summer of 2012 at the European Organization for Nuclear Research (CERN). Physicists have identified four fundamental interactions or forces that govern the behavior of elementary particles: the strong force (primarily responsible for binding the nucleus together), the electromagnetic force (interactions between charges), the weak force (primarily responsible for nuclear decay) and the gravitational force (not included in the standard model). The two major elementary particle groups are the ‘fermions’ and the ‘bosons’. Fermions have half-integer spin, while bosons have integer spin; most bosons have spin of 1. This includes the photon. There are ‘$W^+$’, ‘$W^-$’ and ‘$Z^0$’ bosons, the force carriers of the weak interaction, which have spin of 1. Each elementary particle is associated with an antiparticle with the same mass, but opposite quantum number. Some particles, such as the photon, are identical to their antiparticle. Such particles must be neutral (have no electric charge), but not all neutral particles are identical to their antiparticle. Particle-antiparticle pairs can annihilate each other if they are in an appropriate quantum state, releasing an amount of energy equal to twice the rest energy of the particle. Most often, the antiparticle is denoted by the same symbol as the particle, but with a line over the symbol. For example, the antiparticle of the proton ‘$p$’, is denoted ‘$\bar{p}$’.

The fermions are further subdivided into the leptons and quarks. Leptons do not participate in strong interactions and are freely propagating particles. Quarks come in six varieties referred to as ‘flavors’: up (u), down (d), charm (c), strange (s), top (t), and bottom (sometimes referred to as ‘beauty’) (b). An isolated quark has never been found. Quarks appear in pairs or triplets with other quarks and antiquarks. The resulting particles are the hadrons (more than 200). Hadrons include the baryons and mesons.

A quick comparison of Table 2.2.5.2.3-1 with Table 3.2-1 reveals that the standard model particle classification scheme contains no hadrons. In 1964, physicists Murray Gell-Mann and George Zweig hypothesized that hadrons were composite particles composed of combinations of more elementary particles that Gell-Mann named ‘quarks’. Like other elementary particles, each quark had an antiquark. Their model involved the three quarks ‘u, d, s’ [70,77].

Initially, the prediction of quarks received a cool reception. Not only were quarks never isolated, but they carried a fractional electric charge. This generated suspicion because such a particle had not been found in Nature. There ensued a debate about whether the quark was a real particle or an abstraction employed to explain concepts that were not fully understood at the time. But shortly, Sheldon Glashow and James Bjorken predicted a fourth flavor of quark, which they named ‘charm’ and used it to give a better description of the weak interaction. This equalized the number of known quarks with the number of known leptons and delivered a mass formula that correctly reproduced the masses of the known mesons [75,76,74,73].
The theory that describes the strong interaction is called ‘quantum chromodynamics’ (QCD) and is governed by the $SU(3)$ symmetry group. Strong interactions remain invariant in an internal color space. There are three distinct color charges (red, blue, green). Of course, the charges have nothing to do with color. The term is used mainly to distinguish the strong from electromagnetic forces. In terms of group theory, the assertion that there are no color singlets is simply the statement that QCD has a ‘$SU(3)$’ rather than a ‘$U(3)$’ symmetry. The experimental evidence supports the $SU(3)$ group structure. If color singlets existed, then, mathematically speaking, the elements of the group would commute, which suggests an Abelian group. ‘$U(3)$’ is an Abelian group, ‘$SU(3)$’ is not [78].

In 1968, scattering experiments at the Stanford Linear Accelerator Center (SLAC) showed that the proton was made of smaller, point-like objects, providing evidence that the proton was not an elementary particle. The point-like objects, later identified, were ‘$u$’ and ‘$d$’ quarks. The ‘strange quark’ turned out not only to be a necessary component of Gell-Mann’s and Zweig's three-quark model, but it provided an explanation for the characteristics of the $K$ and $π$-mesons. The ‘strange quark’ was indirectly validated by SLAC. The number of quark flavors grew to the current six in 1973, when Kobayashi and Maskawa discovered that $CP$ violation could be explained if another quark pair existed [79,80,81,82,83].

In November 1974, ‘charm quarks’ were produced almost simultaneously by two teams—one at SLAC under Burton Richter, and the other at Brookhaven National Laboratory under Samuel Ting. The charm quarks were observed bound with charm antiquarks in mesons. This discovery finally convinced the physics community that quarks were real particles and not just convenient abstractions. In 1977, the bottom quark was observed by a team at Fermilab led by Leon Lederman. Finally, in 1995 the top quark was observed also at Fermilab. It had a mass much greater than expected—almost as great as a gold atom. Table 3.2-2 shows the ‘quark’ makeup of various baryons and mesons [70,84,85,86,87].

<table>
<thead>
<tr>
<th>Mesons</th>
<th>Name</th>
<th>Particle</th>
<th>Quark Makeup</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\pi$-meson</td>
<td>$\pi^0$</td>
<td>$u\bar{u} - d\bar{d}$</td>
<td></td>
</tr>
<tr>
<td>$\pi$-meson</td>
<td>$\pi^-$</td>
<td>$u\bar{d}$</td>
<td></td>
</tr>
<tr>
<td>$K$-meson</td>
<td>$K^+$</td>
<td>$u\bar{s}$</td>
<td></td>
</tr>
<tr>
<td>$\eta$-meson</td>
<td>$\eta^0$</td>
<td>$u\bar{u} + d\bar{d} - 2s\bar{s}$</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Baryons</th>
<th>Name</th>
<th>Particle</th>
<th>Quark Makeup</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nucleon</td>
<td>$p$</td>
<td>$uud$</td>
<td></td>
</tr>
<tr>
<td>Nucleon</td>
<td>$n$</td>
<td>$ddu$</td>
<td></td>
</tr>
<tr>
<td>Hyperon</td>
<td>$\Lambda^0$</td>
<td>$uds$</td>
<td></td>
</tr>
<tr>
<td>Hyperon</td>
<td>$\Sigma^0$</td>
<td>$uds$</td>
<td></td>
</tr>
<tr>
<td>Hyperon</td>
<td>$\Sigma^+$</td>
<td>$uus$</td>
<td></td>
</tr>
<tr>
<td>Hyperon</td>
<td>$\Sigma^-$</td>
<td>$dds$</td>
<td></td>
</tr>
<tr>
<td>Hyperon</td>
<td>$\Xi^0$</td>
<td>$uss$</td>
<td></td>
</tr>
<tr>
<td>Hyperon</td>
<td>$\Xi^-$</td>
<td>$dss$</td>
<td></td>
</tr>
<tr>
<td>Hyperon</td>
<td>$\Omega^-$</td>
<td>$sss$</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.2-2
3.2.1 The Characteristics of the Standard Model

The standard model is much more than an elementary particle classification scheme. It is a powerful predictive tool. Given the input parameters and fundamental particle spectra, the standard model makes specific predictions for the structure of composite systems such as nucleons, nuclei, atoms and their interactions such as collisions or decays. The theoretical part of the standard model is a relativistic quantum field theory (QFT) in which the dynamics of the field is generated from the assumption of local gauge symmetries.

Each of the interactions in the standard model is governed by a Lagrangian density ($\mathcal{L}$) function which describes the dynamics of the system and provides the equations of motion i.e. there is an $\mathcal{L}$ for each interaction ($\mathcal{L}_{QED}$, $\mathcal{L}_{ew}$, $\mathcal{L}_{QCD}$). Each Lagrangian density function must remain invariant under the gauge symmetry that constrains it. A gauge symmetry (also called gauge invariance) is the property of a field theory in which different configurations of the underlying fields, which are not themselves directly observable, result in identical observable quantities. A transformation from one such field configuration to another is called a 'gauge transformation'. For example, in electromagnetism the electric and magnetic fields, $'E'$ and $'B'$, are observable, while the potentials $'V'$ (voltage) and $'A'$ (the vector potential) are not. Under a gauge transformation in which a constant is added to $'V'$, no observable change occurs in $E$ or $B$ [87,5,88,89,90].

The gauge symmetries act as a constraint on the number of possible choices for a theory governed by a particular Lagrangian. For instance, the choices for a $'\mathcal{L}_{QED}'$ (the Lagrangian for the electromagnetic interaction) are constrained to equations that remain invariant under the $'U(1)'$ symmetry group. Gauge symmetries constrain the laws of physics, because all the changes induced by a gauge transformation have to cancel each other when written in terms of observable quantities [87,5,88,89,90].

Modern physical theories describe reality in terms of fields e.g. the electromagnetic field and the gravitational field. There are fields for the electron and all other elementary particles. In QFT, particles arise through the quantization of the fields. When a field is quantized, the continuous values of the fields are converted into a smaller number of discrete quantities.

In all of the gauge transformations of the standard model, the symmetries relate only to the matter and force fields themselves and are independent of space-time transformations. This fact represents a major impediment in the search for a way to unite general relativity with the standard model. General space-time transformations are gauge transformations, and as such, according to the quantum theory, are not observable. But, in general relativity, these space-time transformations represent gravitational effects. The upshot of this is that gravitational effects are quantum mechanically unobservable. So far, no one has found a solution to this dilemma.
‘QFT’ associated with the standard model is a completely renormalizable theory meaning that divergences (sums to infinity) appearing in certain mathematical calculations can be made finite by absorbing them into the parameters such as masses and coupling strengths [90].

The standard model possesses a few shortcomings. Some calculations are very difficult limiting the predictive power of the model. Outside the range of current experiments, its predictions become obviously wrong particularly at very high energies. But, there are no violations of the standard model’s predictions in the domain where it has currently been quantitatively tested. That may change since the model is expected to be tested at CERN in ways that it never has before [90].

There is no natural mechanism for electroweak symmetry breaking (explanation for why the $W$ and $Z$ boson masses are not zero). This is imposed “by hand” into the standard model and leads to the prediction of the Higgs boson. While the Higgs boson has been discovered, the theoretical part of the standard model does not give an explanation as to how the Higgs mechanism arises. It’s there because experiments say it is there. But, there is no theoretical explanation for it [90].

### 3.2.2 Experimental Inputs to the Standard Model

Before it can make predictions, the standard model requires 19 measured parameters in its current form and 25 measured parameters if neutrinos have mass [90]. The number of experimentally derived inputs is constrained by the structure of the theoretical part of the model. ‘QFT’ requires particle-antiparticle symmetry arising from $CPT$ invariance. Every particle has an antiparticle with the same mass and mechanical spin, but opposite in the other internal quantum numbers (e.g. electric charge). In counting the experimental inputs to the standard model, tally up only the particles, because their antiparticles are required by $QFT$ [90].

The symmetries governing the standard model predict the number and character of the force-carrying bosons ($\gamma, W^\pm, Z^0, 8$ bi-colored gluons). These particles are set by the theory and are not experimental inputs [90].

The ‘QFT’ part of the standard model contains incomplete information about the fermions (quarks and leptons). However, there are some constraints imposed by the theory. Quarks must come in (electroweak) doublets. A quark doublet is referred to as a ‘family’ or ‘generation’. This is required by the $SU(2)$ symmetry of the theory. The quark doublets must be paired with lepton doublets. This is required to retain renormalizability. If the symmetries of the theory are to be maintained, then the sum of all the charges of quarks and leptons must be zero. The symmetry constraint requires that the number of quark and lepton generations be equal [90].

---

4 Note: when the standard model was first completed, it was thought the neutrinos were massless. But recent experiments show this not to be the case.
Theory does not constrain the number of generations of quarks and leptons. It is possible to construct a consistent “standard model” with 1, 2, 3, 4 or more generations of quarks/leptons. This number is an input to the standard model determined by observing what occurs in Nature. For example, a “standard model” could be constructed with two quark/lepton generations. But, this “standard model” would not accommodate violation of $CP$ symmetry in electroweak interactions. The violation of $CP$ invariance requires a “standard model” with three generations of leptons and quarks. Nature may contain other fundamental particles, but current experiments have not found any [90].

What measurements from experiment are needed to make the standard model completely predictive? The list is provided in table 3.2.2-1.

<table>
<thead>
<tr>
<th>Type of Input</th>
<th>Symbol</th>
<th>Measured Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mass:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leptons</td>
<td>$m_e, m_\mu, m_\tau$</td>
<td>0.511 MeV, 0.106 GeV, 1.78 GeV</td>
</tr>
<tr>
<td>Quarks</td>
<td>$m_u, m_d, m_s, m_c, m_b, m_t$</td>
<td>2 MeV, 5 MeV, 104 MeV 1.27 GeV, 4.20 GeV, 175 GeV⁵</td>
</tr>
<tr>
<td>Bosons</td>
<td>$m_W, m_Z, m_H$</td>
<td>80.40 GeV, 91.19 GeV, ~126 GeV</td>
</tr>
<tr>
<td>Coupling Strength:</td>
<td>$\alpha_e, \alpha_s$</td>
<td>Depends on energy level</td>
</tr>
<tr>
<td>Electroweak Mixing Parameters:</td>
<td></td>
<td>Depends on circumstances</td>
</tr>
<tr>
<td>3 angles, 1 phase</td>
<td>θ₁, θ₂, θ₃, δ</td>
<td></td>
</tr>
<tr>
<td>QCD CP violating parameter:</td>
<td></td>
<td>~0</td>
</tr>
</tbody>
</table>

Experimental Inputs to the Standard Model  
Table 3.2.2-1

The distinguishable fundamental particles included in the standard model are 6 quarks each of which come in 3 colors for a total of 18. Adding to the quarks the 6 leptons gives a total of 24 leptons and quarks. Each lepton and quark has an antiparticle for a total of 48. There are 12 gauge bosons ($\gamma, W^{\pm}, Z^0$, 8 bi-colored gluons) and the Higgs for a total of 61. But, the number of input parameters required by the standard model is not so large because $QFT$ with $CPT$ invariance requires

$$\text{Mass}(\text{particle}) = \text{Mass}(\text{antiparticle})$$

and the QCD structure requires $\text{Mass}(q_r) = \text{Mass}(q_g) = \text{Mass}(q_b)$. In other words, the masses of each quark color must be equal. ‘$QFT$’ requires that the photon and gluon masses be zero.

---

⁵ Values are approximate.
The lepton masses can be measured directly as they propagate as free particles, but the quarks are bound in color singlet hadrons (except for the top quark) and their masses must be found indirectly. In the current standard model, neutrino masses are all zero [90].

There are two independent ‘coupling “constants’, one is associated with the electromagnetic interaction, the other with the strong interaction. In mathematical calculations coupling constants are related to the strength of the interaction. The term ‘constant’ is a misnomer, since neither is really constant. Their values depend on the energy scales at which they are measured. Technically, these parameters should come directly from the theoretical part of the model, but they don’t. Needless to say, this is quite troubling and is related to the hierarchy problem and the issue of fine-tuning, where adjusting parameters very precisely, sometimes by many orders of magnitude, is required in order that the outputs of the model agree with observations. This is less of a problem for the electromagnetic coupling constant, but it, too, needs a slight adjustment in higher energy experiments. More problematic is the strong interaction coupling constant; its larger value subjects it to more extensive fine-tuning.

The electromagnetic coupling constant is \( \ll 1 \), so in the mathematical series expansions in which it is used, after a few terms, the remaining terms don’t contribute much to the series and can be ignored. This is often referred to as a ‘perturbation’. But, the strong interaction coupling constant is much larger and perturbation techniques are not as effective or, in some cases, can’t be used at all making calculations more difficult.

The electroweak mixing parameters are associated with the weak interaction and are parameterized by three angles and one complex phase. The phase angle ‘\( \delta \)’ is the source of \( CP \) violation and is relevant to the cosmological separation of matter from antimatter. Mathematically, these parameters are used in the \( CKM \) matrix that describes the probability of a transition from one flavor of quark to another [90].

In \( QFT \), the Lagrangian of the \( QCD \) sector of the theory contains a term that would permit \( CP \) violation in the strong interaction, in addition to creating an electric dipole moment in the neutron. But, no \( CP \) violation has ever been observed in strong interactions and the measured dipole moment of the neutron is several orders of magnitude smaller than what theory predicts. Hence, that term in the \( QCD \) Lagrangian is augmented by a measured constant ‘\( \Lambda_{QCD} \)’ that is not fixed by the theory. This is another example of fine-tuning and is often referred to as the ‘strong \( CP \) violation problem’ [90].

Small variations of some of the input parameters, fine-tuned to our Universe, result in dramatically different predictions. For example, a small percentage variation in the coupling strengths results in the prediction that nuclei would be unstable [90].
3.2.3 The Theoretical Part of the Standard Model

The theoretical part of the standard model is based on quantum field theory (QFT). There is no general way in non-relativistic quantum mechanics (NRQM) of treating interactions between particles, such as when a particle and its antiparticle annihilate one another to yield neutral particles nor is there a way to describe the decay of a particle into other particles. 'QFT' provides a means whereby particles can be annihilated, created and transmigrated from one type of particle into other(s). Particle number is conserved in NRQM, meaning that the number and types of particles remain the same in all processes. In many interactions, the number of particles does not remain the same. 'QFT' does not require particle number conservation, which is what actually is observed in Nature [91].

'QFT' is a relativistic theory consistent with Einstein’s theory of special relativity. Extrapolation of NRQM to Dirac's relativistic quantum mechanics (RQM) results in states with negative energies. In the early days of the quantum theory, negative energy states were problematic because they were unobservable. 'QFT' treats negative energy states as antiparticles with positive energies overcoming the earlier objections to RQM [91].

Figure 3.2.3-1 illustrates the interaction between an electron and a positron known as 'Bhabha scattering' - an example of the type of problem QFT handles well. At event 'x2', the electron and positron annihilate one another to produce a photon (γ). At event 'x1', the photon is transmuted back into an electron and a positron. Antiparticles are represented by lines with arrows pointing opposite their direction of travel in time. The strange, reverse order of numbering in time, i.e., 2 → 1, is standard in QFT [91].

The annihilation of the electron and the positron at x2 is accompanied by the creation of a photon. The photon is a boson – a force carrying particle and unlike electrons and positrons, it is not a “real” particle, but is a transitory, short-lived and undetectable object called a 'virtual particle', which determines the nature of the interaction between real particles. Next, at event 'x1', the photon is annihilated and an electron and positron are created [91].

Figure 3.2.3-1 is an example of a 'Feynman diagram', named after the renowned physicist Richard P. Feynman (May 11, 1918 – February 15, 1988). Many and varied
and, at times, complicated, Feynman diagrams are the primary method of pictorially representing particle interactions and decays. Feynman diagrams can look simple, but each line in the diagram represents a mathematical equation, which can be quite complicated.

What QFT delivers mathematically is a ‘transition amplitude’, which describes a transition from an initial set of particles to a final set. The square of the transition amplitude equals the probability of finding (upon measurement) that the interaction occurred. This is similar to the square of the wave function in NRQM being proportional to the probability of finding a particle at a certain location [91].

‘QFT’ employs creation and destruction operators acting on states. States represent particle configurations and operators represent transmutations to those particle configurations. The initial state of a quantum system is represented by the symbol ‘|i⟩’ called a ‘ket’. The ‘i’ indicates the initial configuration of particles. The final state of the system is signified by ‘⟨f|’ called a ‘bra’ where ‘f’ indicates the final configuration of particles i.e. ‘transition amplitude’ = ⟨f|o₁o₂...oₙ|i⟩. The o_k’s, k = 1,2,...,n, inserted between the initial state ‘|i⟩’ and the final state ‘⟨f|’ are creation and destruction operators that act in reverse order. The ‘oₙ’ operator acts first, signifying a transmutation into an altered state. The next operator acts on the altered state and the process repeats itself until the final state is reached.

Notice that this process is similar to the permutation group studied earlier. There, three objects are permuted with the initial state represented as |123⟩. When a permutation took place, the initial state was transformed or rearranged (123 → 213) i.e. a|123⟩ = |213⟩. So, the first permutation ‘a’ transforms |123⟩ into |213⟩. The next permutation ‘b’ transforms |213⟩ into |321⟩, the final state. Symbolically, in terms of notation, similar to QFT, this process can be written ⟨321|ba|123⟩, where |i⟩ = |123⟩ and ⟨f⟩ = ⟨321⟩ and ‘a’ and ‘b’ represent the operators.

Figure fig. 3.2.3-1 illustrates, in a grossly oversimplified manner, how a transition amplitude is obtained. First, write down the entire transition amplitude in symbolic terms:

\[ 'Transition Amplitude' = \langle e^+e^-|O_{x_1}O_{x_2}|e^+e^- \rangle \]

The ket ‘|e^+e^-⟩’ represents the initial state of an incoming electron and positron. At x₂ the transmutation is represented by the operator ‘O_{x_2}’, acting on the initial state ‘|e^+e^-⟩’. This operator destroys both the electron and positron and creates a virtual photon. ‘O_{x_2}’ can be written in a more descriptive form: ‘(\tilde{\psi}_dP_c\psi_d)’. At the next event ‘x₁’, the operator ‘O_{x_1}’ destroys the photon and creates an electron and positron. ‘O_{x_1}’ can be written in the more descriptive form: ‘(\tilde{\psi}_cP_d\psi_c)’. After the event at x₁ the system reaches its final state represented by ‘|e^+e^-⟩’. So, the transition amplitude can be written symbolically:

\[ 'Transition Amplitude' = \langle e^+e^-| (\tilde{\psi}_cP_d\psi_c)_{x_1} (\tilde{\psi}_dP_c\psi_d)_{x_2} |e^+e^- \rangle, \]
replacing the operators ‘\(O_{x_2}\)’ and ‘\(O_{x_1}\)’ with their more descriptive forms. At the event ‘\(x_2\)’, the operator ‘\(O_{x_2}\)’ acts on the state ‘\(|e^+e^-\rangle\)’, where \(\psi_d\) destroys the electron and \(\bar{\psi}_d\) destroys the positron. At event ‘\(x_2\)’, the incoming particles (ket) are destroyed by the destruction operators ‘\(\psi_d\)’ and ‘\(\bar{\psi}_d\)’, so at an intermediate point:

\[
\text{‘Transition Amplitude’} = \langle e^+e^- | (\bar{\psi}_c P_d \psi_c)_{x_1} (P_c)_{x_2} K_2 | 0 \rangle
\]

Since the original electron and positron have been destroyed, the destruction operators leave the ‘vacuum ket’, denoted ‘\(|0\rangle\)’, which signifies that no real particles remain. This is accompanied by a purely numeric factor ‘\(K_2\)’ in front of ‘\(|0\rangle\)’. The value of this factor is determined by the formal mathematics of \(QFT\). At this point, understanding the numerical factors is not critical and can, without much loss of comprehension, be ignored [91].

Next, the creation operator ‘\(P_c\)’ creates a virtual photon (\(\gamma\)) at the event ‘\(x_2\)’. The virtual particle is known as ‘a propagator’, because it represents the propagation of a virtual particle from one event in the particle interaction to another. The photon then propagates from \(x_2\) to \(x_1\), where it is annihilated by the operator ‘\(P_d\)’. This process leaves ‘\(|0\rangle\)’ along with an additional numeric factor ‘\(K_\gamma\)’ coming out of the formal mathematics. The transition amplitude then looks like:

\[
\text{‘Transition Amplitude’} = \langle e^+e^- | (\bar{\psi}_c \psi_c)_{x_1} K_\gamma K_2 | 0 \rangle
\]

At \(x_1\), the remaining creation operators ‘(\(\bar{\psi}_c, \psi_c\)’) create an electron and positron out of the vacuum. This leaves the newly created ket ‘\(|e^+e^-\rangle\)’ plus a numerical factor \(K_1\) coming out of the formal mathematics of \(QFT\). The ket and the bra now represent the same state and the transition amplitude can be written:

\[
\text{‘Transition Amplitude’} = \langle e^+e^- | K_1 K_\gamma K_2 | e^+e^- \rangle
\]

‘\(K_1, K_\gamma, K_2\)’ are just numbers multiplied together, so as part of the formal mathematics of \(QFT\), they can be moved from the ket/bra structure and the transition amplitude can be written:

\[
\text{‘Transition Amplitude’} = K_1 K_\gamma K_2 \langle e^+e^- | e^+e^- \rangle
\]

An important point here is that in \(QFT\) the bracket of a multi-particle state ‘\(\langle e^+e^- | e^+e^- \rangle\)’ is defined so it always equals unity i.e. \(\langle e^+e^- | e^+e^- \rangle = 1\), if the ket and the bra are identical and zero otherwise. The transition amplitude can then be written:

\[
\text{‘Transition Amplitude’} = K_1 K_\gamma K_2
\]

The process illustrated in fig. 3.2.3-1 can be thought of as an evolution of the original state, represented by the ket, to the final state, represented by the bra. At each step along the way, the operators act on the ket to change it into the next part of the
progression. When the ket looks the same as the bra, the full transition has been completed, and the bracket then equals unity. What is left is the numerical quantity ‘$K_1K_yK_2$’ [91].

‘$(K_1K_yK_2)^2$’ equals the probability that the process described in fig. 3.2.3-1 actually happens. This process depends on particle momenta, spins and rest masses, as well as the inherent strength of the electromagnetic interaction, all of which play a role in determining the probability that the interaction takes place. These, among other subtleties, have been omitted in order to convey, as simply as possible, the essence of a transition amplitude [91].

In $NRQM$, the solutions to the relevant Schrödinger equation are states (kets). But, the solutions to the relevant wave equations in $QFT$ are not states (particles), but operators that create and destroy states. Different solutions exist that create or destroy every type of particle and antiparticle. In this difference lies the power of $QFT$ [91].

Each sector of the standard model is associated with a Lagrangian density function ($\mathcal{L}_{ew}, \mathcal{L}_s$). It describes the dynamics of the system, provides the equations of motion and determines which particle interactions and decays can and cannot take place. In $NRQM$, solutions to the dynamical equations are provided by an appropriately chosen wave function ‘$\psi$’. Here, wave functions are considered to be states (particles). By contrast, in $QFT$, the solutions to the dynamical equations are not states, but creation and destruction operators (things that create or destroy particles). Since each of the Lagrangian density functions associated with an interaction is constrained by a symmetry, ($SU(2)$ symmetry for weak interactions and $SU(3)$ symmetry for the strong interaction), the operators must meet the requirements of each theory i.e. they must be chosen to allow the Lagrangian density for the weak interaction to remain invariant under $SU(2)$ symmetry and the Lagrangian density for the strong interaction to remain invariant under $SU(3)$ symmetry.

This describes the theoretical part of the standard model in grossly oversimplified terms. While the details are complicated, the high level process is fairly simple. Creation and destruction operators determine what kinds of particle interactions are possible in Nature. As solutions, operators must be chosen to conform to certain symmetry constraints. And that’s pretty much it.

3.2.4 Summary of the Standard Model

To summarize: the standard model of particle physics describes the electromagnetic, weak, and strong interactions of subatomic particles. Developed throughout the mid to late $20^{th}$ century, the standard model was molded by the interplay of experimental discoveries and theoretical advances. It was a collaborative effort involving many physicists from many continents and its development took many decades. The current formulation, finalized in the mid-1970s was completely experimentally validated upon the apparent confirmation in 2012 of the Higgs boson. Because of its success in
explaining a wide variety of experimental results, the standard model is sometimes regarded as ‘the theory of almost everything’ [92].

Note the phrase ‘the theory of almost everything’. The key word here is ‘almost’ as the standard model harbors a number of deficiencies:

1. The self-consistency of the theoretical part of the model has not been mathematically demonstrated
2. Experiments indicate that neutrinos have a non-zero mass, which the current standard model does not allow
3. The standard model does not include gravity and there is no known way of describing general relativity in terms of a quantum field theory
4. The standard model requires $19 +$ unexplained numerical constants and including the neutrino masses will require an additional 7 or 8 constants, which will also likely remain unexplained
5. The Higgs mechanism gives rise to the hierarchy problem, if new physics is present at high energy scales, which leads to the fine-tuning of the standard model parameters through renormalization, which is related to deficiency ‘#1’
6. The standard model cannot explain the amount of cold dark matter and predicts dark energy contributions that are too large and cannot easily accommodate the predominance of matter over antimatter
7. The explanation for the evolution of the universe seems to require a mechanism like “cosmic inflation”, but no such mechanism exists in the standard model

3.3 Beyond the Standard Model

The deficiencies of the standard model inspired physicists to look beyond – to find a theory that could explain the mysteries that the standard model could not. There emerged several candidates albeit with deficiencies of their own.

3.3.1. Einstein’s Failure

While new quantum discoveries came continuously to light allowing the further development of the quantum theory, Einstein persisted in his pursuit of a unified field theory. While aware of the new theory, he openly despised its indeterministic characteristics. He believed that the true course to a unification of physics rested in extending the field equation of general relativity into the quantum realm. It remained for him to show how this could be accomplished. This, he was never able to do.

Einstein tried several approaches. After completing the theory of general relativity, his early work on unification was, by and large, a reaction to approaches advanced by
others. Those included: the geometrization of the electromagnetic field proposed in 1918 by H. Weyl; the five dimensional theory suggested by Theodor Kaluza in 1919 and the concept of the affine connection on the metric field as advanced by Arthur Eddington (28 December 1882 – 22 November 1944) in 1921. Einstein continued to explore modifications of the Kaluza approach until his death in 1955 [93].

In Weyl’s geometrization of the electromagnetic field, the solution included the notion that light emitted by a radiating atom would depend on the prehistory of that atom. Einstein remarked that, in such a case, no uniquely defined frequencies of the spectral lines of a chemical element could exist i.e. the frequencies would depend on the location of the emitter [154]. This was at odds with experimental facts.

Regarding Eddington’s affine connection on the metric field, Einstein found that the theory did not account for the differences in electron-proton masses and that no singularity-free electron solution seemed possible. And since an electron is a singularity-free particle, Eddington’s solution was a non-starter [93].

In 1931, Einstein and his assistant Mayer succeeded in deriving the gravitational and electromagnetic field equations from Kaluza’s idea of a five-dimensional approach. In their first paper, they could not account for the structure of matter and concluded that the existence of charged particles was incompatible with the field equations. In the end, Einstein and his coworkers found it impossible to describe particles by non-singular solutions and to account for the quantitative difference in the strength of the gravitational and electrostatic forces between material particles [93].

Einstein only concerned himself with explaining how the electron and proton emerged. More serious from the conceptual point of view was the discovery of mesons. The muon was discovered in 1937. The charged pion was postulated in 1935 and discovered in 1947, and the neutral pion was postulated in 1938 and discovered in 1950. By the mid-fifties, at the time of Einstein’s death, a dozen elementary particles were known. Ten years later, the list of experimentally confirmed subatomic particles numbered about one hundred [93]. With the discovery of elementary particles beyond the electron and proton, twentieth century physics had postulated two new kinds of fundamental interactions (the weak and strong atomic interactions). Physics was moving past the old master.

3.3.2 Supersymmetry

Supersymmetry is a new particle classification scheme which postulates that every elementary particle of integer spin comes with another “super-partner” particle of 1/2 spin. This imposes a symmetry between the fermions and the bosons; for every quark (spin-1/2) there is a squark (spin-0); for a photon (spin-1) there is a photino (spin-1/2) etc.

Potentially, supersymmetry could provide the added ingredient necessary for resolving several theoretical problems; hence, many theoretical formulations, including string theory, contain supersymmetry. If supersymmetry exists at high energy scales, it would
provide an explanation for the hierarchy problem. The Higgs boson mass is subject to quantum fluctuations that would make its mass so large as to undermine the internal consistency of the standard model. Currently the problem is overcome by fine-tuning the parameters of the model. In supersymmetric theories the quantum fluctuations are canceled more naturally by the contributions of the super-partners eliminating the need for fine-tuning.

Supersymmetry allows for the high-energy unification of the weak, strong and electromagnetic interactions and provides a candidate for dark matter. At high energies the standard model suggests that the three interactions are unified. But this can only be accomplished with extreme fine-tuning of the parameters, and even then, the forces don’t quite unify at higher energy scales. Supersymmetry closes the gap allowing near exact unification of the three forces without extreme fine-tuning.

The gauge symmetries of the standard model are internal symmetries, meaning they are independent of space-time transformations. The ‘Coleman–Mandula theorem’ prohibits space-time and internal symmetries from being combined in any nontrivial way. Supersymmetry appears to offer the only "loophole" to this. Another theorem called the ‘Haag-Lopuszanski-Sohnius theorem’ demonstrates that supersymmetry is the only way space-time and internal symmetries can be consistently combined. The upshot of this is that supersymmetry appears to be the only mechanism for unifying general relativity with the standard model [94].

There is no direct evidence that supersymmetry exists; no super-partners have been experimentally observed. While all supersymmetric extensions of the standard model cannot be definitively ruled out at the Large Hadron Collider (LHC), for supersymmetry to solve the problems for which it was invented, at least a few of the super-particles should not be too heavy. To constitute dark matter, for example, the super-particles need to weigh no more than a few tenths of 1 TeV, which is well within the capabilities of the LHC. Another reason for super-particles to be light lies in the nature of the Higgs boson. A Higgs mass of around 0.125 TeV, the mass at which it was discovered, implies the super-particles should have fairly light masses. Unfortunately, if super-particles exist, they must be even heavier than many physicists had hoped. The situation is this: a number of experiments have already been conducted at the LHC and a number of ‘easy’ models should have shown up experimentally, but didn’t. If the collider fails to find them, supersymmetry’s usefulness could fade away. That would be a blow, not just to supersymmetry, but also to the more ambitious unified theories of physics that presume it [95].

### 3.3.3 Technicolor

Technicolor theories postulate a new super-strong interaction that gives a more natural explanation of electro-weak symmetry breaking (the mechanism through which elementary particles acquire mass) that overcomes the hierarchy problem. In spite of the precise agreement of the electroweak theory with experiment at energies accessible so far, the ingredients for electro-weak symmetry breaking remain unknown. The Higgs mechanism predicts the existence of the Higgs boson. But, the Higgs boson is
"unnatural" in the sense that quantum mechanical fluctuations produce corrections to its mass that lift it to such high values that it cannot play the role for which it was introduced without delicate fine-tuning of standard model parameters [96].

Instead of introducing the Higgs boson, technicolor models generate masses for the $W$ and $Z$ bosons through the dynamics of new gauge interactions. The new strong interaction leads to a host of new composite, short-lived particles called 'leptoquarks' at energies accessible to the LHC. The approach is natural because there is no elementary Higgs boson and, hence, no fine-tuning of parameters. The discovery of a Higgs-like boson with mass approximately $125 \text{ GeV}/c^2$ is not predicted by technicolor models. But, the main issue with technicolor theories is that no new gauge bosons or leptoquarks have been experimentally observed. If the collider does not find the new particles, it is unlikely that technicolor theories could survive [96,97,98,99,100,101,102,103,104].

### 3.3.4 Grand Unified Theories (GUTS)

The standard model is based on three independent interactions, three symmetries and three coupling constants. By contrast, GUTS, at high energy, merge the electromagnetic, weak, and strong interactions into one interaction. ‘GUTS’ postulate that the standard model symmetry groups are subgroups of a larger symmetry group with one unifying coupling constant. Compared to the standard model, all realistic GUTS are quite complicated. They introduce additional fields and interactions or even additional dimensions of space to explain the observed fermion masses and mixing angles. Due to the complications, complexities and lack of any experimental evidence, there is no generally accepted GUT.

Most GUTS predict the existence of new particles in the form of quarks and leptons. But, the predicted particles cannot be observed because detecting their masses is beyond the reach of current and foreseen collider experiments. Instead, experiments designed to detect traces of grand unification are performed indirectly by observing phenomena such as proton decay, electric dipole moments of elementary particles or the properties of neutrinos. Some GUTS predict the existence of magnetic monopoles [105].

A few GUTS explicitly break the baryon number symmetry, allowing protons to decay via the Higgs particle, magnetic monopoles or new $X$ bosons. Proton decay is one of the few observable effects of the various proposed GUTS. To date, there have been no observations that would support any particular GUT. In addition, magnetic monopoles have not been observed in Nature.

### 3.3.5 Preons

Preon theories aim at finding theoretical answers to the large number of arbitrary constants within the standard model. Preon theories postulate that quarks and leptons are composite rather than fundamental particles. But, up to now, scattering experiments have observed no such structure [90]. The interest in preons was primarily
predicated on the assumption that the Higgs boson would not be discovered by the LHC. Since the Higgs has been tentatively discovered, it is doubtful that preon theories will garner much interest in the future.

3.3.6 String Theory

String theory speculates that matter does not consist of point-like entities as in the standard model but rather tiny loops of string. Strings are two dimensional objects that vibrate in space-time. Each harmonic of a string represents a different particle. From this beginning, the laws of physics emerge. General relativity, electromagnetism and Yang-Mills gauge theories all appear in a surprising fashion [12].

The first string theories were the very limited bosonic types. To be mathematically consistent, Bosonic string theory requires 25 extra spatial dimensions. It predicts the existence of a particle called a ‘tachyon’. The mass of a tachyon is represented by an imaginary number implying that it travels faster than light. Such a particle has never been observed and its existence would conflict with known physics. The theory requires strongly interacting massless particles that cannot be brought to rest. Finally, it leaves out known particles of Nature - namely fermions [34]. Bosonic string theories have never been considered serious candidates for unification.

But in 1970, Pierre Ramond was able to include fermions in a supersymmetric string theory. The new theory had no tachyons and instead of 25 extra spatial dimensions there were only nine. About the same time, Andrei Neveu and John Schwartz found a superstring theory, where the superstrings interacted with one another allowing it to be consistent with special relativity and the quantum theory. Two years later Joël Scherk, Tamiaki Yoneya and Schwartz found that some of the massless particles in the theory behaved like gravitons (the particles that are supposed to carry the gravitational force). This discovery amounted to a breakthrough and immediately suggested that string theory could unite all of the interactions [16].

In 1984, the first of two momentous events would take string theory from an obscure fuzzy idea to one of the more intriguing ideas ever to come along in physics. Schwartz in collaboration with Michael Green provided strong evidence that string theory was finite, freeing it from the mathematical anomalies that had plagued previous attempts at unification. Within weeks some of the most ardent critics of string theory began working on it. Theoretical physics had struck gold. They had found a quantum theory of gravity that also unified forces with matter [13].

But very quickly it was realized that string theory was not unique. Instead of one, there were five equally consistent 10 dimensional superstring theories. And to accept string theory meant assenting to the extra spatial dimensions and the requirements of supersymmetry without which the hope of creating a theory that solved the hierarchy problem and agreed with Nature vanished.
In any believable theory the extra dimensions would have to be curled up into small
difficult to detect loops. The many ways of curling up the extra dimensions amounted to
a new theory. When strings are allowed to move and vibrate in the complicated
geometry of six extra spatial dimensions, a vast array of different kinds of particles
arise. There are constants associated with the masses of the particles and with the
forces. As it turned out, string theory added many more, not less, freely varying
constants. Even so, string theorists were unable to produce a string theory with the
characteristics of the standard model [13].

Things looked bleak, until 1995, when drawing on the work of a number of string
theorists (Ashoke Sen, Chris Hull, Paul Townsend, Michael Duff and Schwarz), Edward
Witten suggested at a conference at USC that the five known string theories
(Type I, Type IIA, Type IIB, heterotic SO(32) and heterotic E8 × E8) were different
aspects of a single underlying theory Witten named ‘M-theory’. When the five string
theories were analyzed, Witten discovered that an object in one superstring theory was
a dual description of a different object in another superstring theory. Because the extra
dimensions can come in different shapes and sizes, string theory can come in many
variants i.e. different universes, where the different shapes and sizes of the extra
dimensions have different physics. However, Witten was able to show that many of
these geometries result in the same physics [107].

One theorist who took great interest in the Witten lecture was Joseph Polchinski - a
physicists at UC Santa Barbara. He showed that in order for the dualities in string
theory to work correctly, the theory must include not only strings but also objects of
higher dimension. He called these new objects ‘branes’ – short for membranes.

Strings are subject to boundary conditions. Just like fastening each end of the strings
on a guitar achieves the desired harmonics, the strings in string theory needed
something to which they can fasten themselves. Polchinski realized the dualities
between the five different 10 dimensional superstring theories required a class of
extended objects upon which open strings can end. He called these objects ‘D-branes’.

D-branes are dynamical objects, where the environment of strings stretching between
them becomes very rich. Polchinski was able to connect certain string modes to a non-
abelian gauge theory, which is the same type of theory as the standard model. He
showed that it might be possible to define observables in space in the sense of quantum
mechanics [111].

Polchinski’s discovery was both a blessing and a curse. Branes are additional features
added to the background geometry of the strings. Their addition greatly increases the
number of background geometries upon which the string can live. Branes made it
possible to describe some special types of black holes within string theory [16].

Imagine stealing an amount of hot gas from the universe and dropping it into a black
hole. Since the gas cannot escape from the hole’s gravitational pull, the entropy in the
universe would decrease in violation of the second law of thermodynamics. Saving the
second law requires that the black hole gain whatever entropy the in-falling gas originally had. Accordingly, Stephen Hawking surmised that a black hole should emit a characteristic spectrum of thermal radiation in order that the Universe's entropy does not decrease. It is possible to employ thermodynamic arguments to derive the "Bekenstein entropy" for black holes. Counterintuitively, Bekenstein entropy is proportional to the black hole's surface area rather than its volume.

Black hole entropy is closely associated with the holographic principle. Gerard 't Hooft in 1993 (with help from Leonard Susskind) showed that the amount of "information" a space contains may be related to the area of a region’s boundary, not its volume [109]. In quantum field theory, everything can be viewed as information. In short, the holographic principle amounts to the following two postulates:

1. A gravitational theory describing a region of space is equivalent to a theory defined only on the surface area that encloses the region

2. The boundary of a region of space contains at most one piece of information per square Planck length

String theorists have constructed models in which a black hole is described as a very long (and hence very massive) string. This model gives rough agreement with the expected entropy of a Schwarzschild black hole, although an exact proof of this has so far eluded its proponents. The chief difficulty is that it is relatively easy to count the number of degrees of freedom quantum strings possess if they do not interact with one another. But for the black hole problem, gravity is an interaction, and so if the "string coupling" is turned off, in other words, if gravity is turned off, no black hole could ever arise. Calculating black hole entropy requires working in a regime where string interactions exist.

String theory can solve the simpler case of non-interacting strings. This requires supersymmetry. In certain cases, the entropy calculation done for zero string coupling (no interaction) remains valid when the strings interact. The challenge for string theory is to devise a situation in which a black hole can exist without breaking supersymmetry. This has been done by building black holes out of D-branes. Calculating the entropies of these hypothetical holes gives results which agree with the expected Bekenstein entropy. Unfortunately, the cases studied so far all involve higher-dimensional spaces. They do not directly apply to black holes observed in our own universe [110,111]

String theorists must find a way of interpreting the extra spatial dimensions that makes sense in the context of the real world. In 1997, Argentinean physicist Juan Maldacena, inspired by the holographic principle, proposed the anti-de Sitter conformal field theory correspondence (AdS/CFT). His conjecture was: the Universe is a 4-dimensional boundary of a 5-dimensional space which contains the same information. In Maldacena’s AdS/CFT correspondence, he proposed a new duality between a gauge theory defined on a 4-dimensional boundary (three space dimensions and one time
dimension) and a 5-dimensional region (four space dimensions and one time dimension) [16].

The \textit{AdS/CFT} correspondence, called the ‘Maldacena duality’, is the conjectured equivalence between a string theory and gravity defined on one space, and a quantum field theory without gravity defined on the conformal boundary of this space [16].

Such were the successes of string theory over about a 30 year period. All the forces and particles of Nature could be explained as characteristic vibrations and actions of strings and branes. It included gravity and unified the bosons with the fermions. The laws of motion of the particles were united with the laws governing the forces. String theory has only two fundamental quantities: the string tension and the string coupling constant. It is governed by one simple principle: strings move so as to minimize the surface area drawn out by the strings moving in space-time. There was evidence that the five seemingly different 10 dimensional string theories were a manifestation of a single underlying theory (\textit{M}-theory). String theory gave insights into black hole entropy. And finally, there appeared to be a way of relating string theory to gravity and the gauge symmetries of quantum field theory, which gave a plausible connection with our own Universe [14].

But there are obstacles to overcome. \textit{M}-theory has yet to be defined. It remains a conjecture and attracts criticism for lacking predictive power and for being incomplete and untestable. Witten has suggested that a general formulation of \textit{M}-theory will require the development of new mathematical language. With no underlying foundation, it is becoming increasingly difficult to make meaningful progress on string theory.

John Schwartz and Michael Green provided strong evidence that string theory was finite. But the “finite theory” assertion has since come under scrutiny. Most string theorists point to a proof by Stanley Mandelstam as evidence that string theory is finite. But there does not seem to be general agreement that the Mandelstam’s proof is complete. Primarily, string theorists make their calculations using perturbation approximations. Without \textit{M}-theory, all the perturbation expansions in string theory would have to be proven finite or replaced with exact calculations [14].

While the introduction of branes brought a welcome amount of flexibility to string theory, it added several layers of complexity. The addition of branes increases the number of background geometries upon which the string can live [16]. This greatly increases the number of possible string theories.

The general theory of relativity is background independent. This means that the laws of gravity i.e. the space-time backgrounds are dynamic, not fixed. Einstein was concerned about this aspect of his theory. He could not bring himself to believe that the Universe was like a malleable piece of clay. He added to his equations a term: the ‘cosmological constant’. It cancelled the effect of the expansions or contractions essentially predicting the Universe as static. However, after Einstein had added the cosmological constant, Hubble, among others, argued that the Universe was expanding. That argument
eventually led to the big bang theory of the Universe. Once physical evidence appeared, which made Hubble’s hypothesis probable, Einstein felt that the inclusion of the cosmological constant in his equations had been a mistake.

But, all experimental and observational evidence today indicates that the equations of gravity require a non-zero cosmological constant. This is primarily due to the existence of dark matter and energy which is chiefly thought responsible for the expansion of the Universe. It also suggests that the value of the cosmological constant is positive rather than negative.

The Maldacena conjecture is described on anti-de Sitter spaces. All universes described on anti-de Sitter spaces have negative curvature, and hence, have negative cosmological constants. Our Universe, evidently, cannot be described on an anti-de Sitter space.

All string theories at present are background dependent. This leads to the ‘vacuum problem’. In classical physics a vacuum has zero energy. In quantum field theories, the notion of a zero energy space has been replaced with idea of a ‘vacuum state’. It is defined as the ground (lowest energy density) state of a collection of quantum fields. Quantum fields exhibit fluctuations everywhere in space, even in regions which are otherwise ‘empty’ i.e. devoid of matter and radiation. These zero-point fluctuations give rise to an enormous vacuum energy density which is believed to act as a contribution to the cosmological constant appearing in Einstein’s equations of general relativity [112]. In string theory the vacuum state is not given. It must be determined by specifying the background conditions necessary to make the vacuum state what it is. This means the 10 or 11 dimensional space-time configuration along with some other technical details must be specified. Unfortunately, there is an enormous number, maybe an infinite number, of consistent choices for a background. And each of these background spaces comes with a large set of parameters that specify the size and shape of the background space-time [113]. This gives rise to an enormous number of possible string theories.

The plethora of possible string theories raises the question of identifying which string theory describes our Universe, and further, invokes a requirement of explaining what all the other string theories represent. The circumstance has brought forth an abundance of unusual ideas. Not the least controversial is the notion of an ‘omniverse’. An omniverse is the conceptual ensemble of all possible universes, with all possible laws of physics. The definition of a "universe" includes one set of "physical laws and constants that govern it". The idea of an ‘omniverse’ includes multiple sets of physical laws and constants; each expresses a wholly or partially separate universe [5].

To some string theorists, the concept of an omniverse has become an acceptable alternative to the idea of a ‘universe’. One of the biggest proponents of this view is Leonard Susskind of Stanford University. He argues:
“... Physicists always wanted to believe that the answer was unique. Somehow there was something very special about the answer, but the myth of uniqueness is one that I think is a fool’s errand ... If there were some fundamental equation which, when you solved it, said that the world is exactly the way we see it, then it would be the same everywhere. On the other hand, you could have a theory which permitted many environments, and a theory which permitted many different environments would be one in which you would expect that it would vary from place to place. What we’ve discovered in the last several years is that string theory has an incredible diversity – a tremendous number of solutions – and allows different kinds of environments. A lot of the practitioners of this kind of mathematical theory have been in a state of denial about it. They didn’t want to recognize it. They want to believe the universe is elegant – and it’s not so elegant. It’s different over here, it’s that over here. It’s a Rube Goldberg machine over here. And this has created a sort of sense of denial about the facts of the theory. The theory is going to win, and physicists who are trying to deny what is going on are going to lose.”

In 2003, Susskind published “The Anthropic Landscape of String Theory” in which he very publicly gave up the idea that a unique string theory would be discovered [15,16]. In the paper, he introduced the concept of “the landscape” of string theories; a vast number of mathematically consistent possible universes.

While Susskind’s musings are both controversial and interesting, he raises two important issues:

1. Of the vast landscape of string theories, which one describes our Universe or our part of the Universe?

2. If some kind of Omniverse, then how is this fact to be verified?

The answer Susskind gives to the first of these questions is “the anthropic principle”. Many string theorists now champion the idea of using the anthropic principle as means of determining which string theory best describes the Universe/Omniverse.

What is the anthropic principle? It seems there are finely tuned connections between the values of the physical constants of a theory and the conditions necessary for life to exist. These connections are often referred to as ‘anthropic coincidences’. Here are some of the most significant [113]:

1. The electromagnetic force is 39 orders of magnitude stronger than the gravitational force and if they were more comparable in strength, stars would have collapsed long before life had a chance to evolve

2. The vacuum energy density of the Universe is at least 120 orders of magnitude lower than some theoretical estimates and if at any time it were as large as the calculations suggest, the Universe would have quickly blown apart
3. The electron mass is less than the difference in the masses of the neutron and proton, a free neutron can decay into a proton, electron, and anti-neutrino and if this were not the case, the neutron would be stable and most of the protons and electrons in the early universe would have combined to form neutrons, leaving little hydrogen to act as the main component and fuel of stars.

4. The neutron is heavier than the proton, but not so much heavier that neutrons cannot be bound in nuclei where conservation of energy prevents the neutrons from decaying and without neutrons there would not be the heavier elements needed for building complex systems such as life.

5. The carbon nucleus has an excited energy level at around $7.65 \, MeV$ and without this, insufficient carbon would be manufactured in stars to form a basis for life.

The anthropic principle comes in two versions:

1. The weak anthropic principle ($WAP$): our location in the Universe is necessarily privileged to the extent of being compatible with our existence as observers.

2. The strong anthropic principle ($SAP$): the Universe (and hence the fundamental parameters on which it depends) must admit the creation of observers within it at some stage.

Even string theorists who now embrace the anthropic principle such as Susskind and Polchinski once despised it as being totally unscientific. The unscientific criticism comes in large part because the $SAP$ is sometimes invoked to argue for a supernatural designer of the Universe. Ironically, the $WAP$ is often used as an argument against a supernatural designer, as Susskind does in his book “The Cosmic Landscape” [16].

The most divisive issue lingering around the anthropic principle is the idea of a finely tuned universe. A fine-tuned universe is one where the conditions that allow life to exist occur only when certain fundamental physical constants lie within a very narrow range of values. If any of several fundamental constants were only slightly different, the Universe would not be conducive to the establishment and development of matter, astronomical structures, elemental diversity or life [114].

The existence and extent of fine-tuning in the universe is a matter of dispute. Fine-tuning does not require multiple universes. What it does require is an explanation of why the critical constants of Nature have the values they do. This explanation would have to be a characteristic of $M$-theory if the reductionist dream of explaining all the fundamental interactions as part of a unified theory is to be realized.

Another possibility is that the parameters are not fine turned, but essentially random. The concept of multiple universes can then explain why our Universe seems fine-tuned for conscious life. If there were a large number (possibly infinite) of different physical laws (or fundamental constants) in as many universes, some of these universes would
have laws that were suitable for stars, planets and life to exist. The makeup of our universe is then a consequence of the laws of probability \[16\]. This is essentially the position that Susskind adopts.

But those who make the “no fine-tuning” argument must address the vacuum energy problem. This is essentially the cosmological constant problem, since any vacuum energy density is equivalent to this parameter in Einstein's theory of general relativity. Calculations give a value for the vacuum energy density that is some 120 orders of magnitude greater than the maximum value possible from observations. Since this density is constant, it would seem to have been fine-tuned from the early Universe, so that its value allowed for the existence of life \[114\].

Until recently, the value of cosmological constant was thought to be zero, in which case there is no need for fine-tuning. However, in 1998, two independent research groups studying distant supernovae discovered that the expansion of the Universe is accelerating. More recent observations have confirmed this result. The source of this cosmic acceleration is believed to be some still-unidentified dark energy, which constitutes 70 percent of the mass of the Universe. One possible explanation is gravitational repulsion by means of the cosmological constant by way of the vacuum energy field that is allowed by general relativity. If this is the case, then the cosmological constant seems to be fine-tuned \[115\].
Chapter 4

The Practical Challenges of Unification

What I am going to tell you about is what we teach our physics students in the third or fourth year of graduate school... It is my task to convince you not to turn away because you don’t understand it. You see my physics students don’t understand it... That is because I don’t understand it. Nobody does.

— Richard P. Feynman, QED: The Strange Theory of Light and Matter

4.0 Introduction

Chapter 1 discussed how physics is dominated by two great, but seemingly incompatible theories. ‘Quantum gravity’ is the name given to a future theory that, if uncovered, would unite the quantum theory with relativity. Presently, such a theory remains afflicted by both practical and conceptual challenges, which, up to now, have prevented creating a generally accepted solution.

Chapter’s 2 and 3 reviewed the major discoveries that ultimately led to the standard model of particle physics and the general theory of relativity. While arguably the most successful endeavor in the history of modern science, as a unifying theory, the standard model harbors shortcomings, most notably, lacking an explanation for the gravitational force at the quantum level. Its theoretical basis has not been proven consistent and it requires a number of experimentally derived inputs having no basis in theory. Outside the energy ranges of its domain, its predictions are clearly wrong. The input parameters used in the model require unnatural fine-tuning and artificial mechanisms are employed to make the predictions of the model agree with observations.

The theory of relativity is our current theory of gravitation. In view of its agreement with experimental results and ability to predict new phenomena, it is perhaps the most successful purely theoretical development in the history of science. The theory immediately explains most of the physics at large scales and has continued to explain new developments, including the existence of black holes and recent observations in cosmology [155]. But its domain of applicability is constrained to large scale physics. There does not seem to be a way of extending it to the smaller scales of atomic physics.

For these and other reasons, physicists have looked for an approach that surpasses the standard model and the theory of relativity – string theory being the most promising alternative. While ingenious, the alternatives come with their own set of problems. This chapter contains a discussion of the practical challenges faced in overcoming the dichotomy between relativity and the quantum theory. In the end, the barriers inhibiting the uniting of the theory of relativity with quantum mechanics will be found primarily logical in nature, not conceptual as most physics believe.
4.1 Quantum vs. Classical Theories

The theory of relativity is a classical theory. There are three features of quantum theories that distinguish them from classical theories and one feature that illustrates how the two are connected [24]:

1. The indivisibility of quantum processes
2. Incomplete determinism in quantum laws
3. The holistic nature of quantum processes
4. The correspondence principle

4.1.1 The Indivisibility of Quantum Processes

Classical theories assume that all physical processes happen gradually. There are no sudden jumps or spikes in energy transitions, for example. But experiments have shown that such an assumption is invalid [24]. Bohr’s theory of the atom, the photoelectric effect, the absorption and emission of radiation in black bodies illustrate that quantum processes come in discrete packages.

The indivisibility of quantum processes has influenced the mathematical approach to describing quantum phenomena. Classical theories are based on the idea of infinite divisibility, which assumes that a process, like the trajectory of a rocket, can be divided into smaller and smaller segments, in fact, into segments of any desired smallest. The differential and integral calculus describes the motion of objects as making smooth transitions from one stage to the next. But it is difficult to apply the calculus to processes that are discretely discontinuous i.e. processes that are not infinitely divisible.

4.1.2 Incomplete Determinism in Quantum Laws

Classical physics is casual in nature. If the current state of system is known in enough detail, all future states of the system can be predicted with certainty. The current state of a system determines, in a sense, all future states. All that is needed [24]:

1. The system position at any instant in time
2. The velocity at that time
3. The value of the force at all times

This characteristic of classical physics is often associated with ‘determinism’. But numerous experiments have shown that quantum processes are uncontrollable. It is not possible to predict when energy transfers occur, for example [24]. Only betting odds can be given. The quantum world is probabilistic, not deterministic.

The probabilistic nature of quantum phenomena is one of the more controversial aspects of the theory. From a classical standpoint, if a problem requires a statistical treatment, this generally implies that a large number of variables cannot be easily
accounted for. But, if those variables could be known with certainty and described, deterministic solutions would follow.

But ‘probability’, in the quantum sense, plays the role of a Natural law. There are no “difficult to account for” variables that, if known, would allow for a more detailed or better answer. The indeterminate processes at the quantum level are not more fundamentally discoverable [24]. And hence, those processes are not knowable beyond the limits of probability.

Many physicists of the day – particularly Einstein, took issue (some still do) with this particular feature of quantum mechanics. In the now famous EPR paper, Einstein and his collaborators argued that the quantum theory ignored variables, which, if known, could explain the nature of atomic phenomena more thoroughly, without some aspect of the system becoming uncontrollable. And this he felt would return determinism to the atomic realm. His objections led to the development of Bell’s theorem (more on this later). And Bell’s theorem, as it turns out, was accessible to experimental verification. Originally designed to validate the theorem, experiments showed that Bell’s theorem was violated, as predicted by the quantum theory. Although the experiments are not completely conclusive, the results imply that it is not lack of information preventing an explanation of the nature of quantum processes. And it gave credence to the notion that the quantum theory is consistent with the way it is currently interpreted.

Again, the probabilistic feature of the quantum theory impacted how quantum processes were described mathematically. For example, it is not possible to specify the exact location and momentum of a particle simultaneously. Such variables, like position and momentum, are called ‘conjugate variables’. This is a consequence of the uncertainty principle, which forbids simultaneously specifying exact values for conjugate variables. In quantum problems, conjugate variables are considered uncontrollable. The best that can be hope for is to calculate their statistical averages. This means that in experiments, where conjugate variables become important, each time the experiment is conducted, a slightly different value will be obtained. But if enough experiments are performed, values for conjugate variables will hover around a statistical average. Also, the classical equations containing conjugates variables could not be carried over into the atomic realm. Any problem requiring a complete knowledge of the status of a system at every moment in time will find no place in the quantum theory [24].

4.1.3 The Holistic Nature of Quantum Processes

The ‘particle’ and the ‘wave’ represent the two great concepts of classical physics. Virtually every problem in the subject can be explained in terms of one or the other of these ideas [9]. The two concepts, however, are quite different. Particles have a center of mass located at a point in its interior, which exists at a certain space-time location. Waves, on the other hand, are spread out with no specific location in space and time. Waves are characterized by interference, diffraction, polarity and periodicity - attributes that particles were thought not to possess. Classical physics could classify phenomena into two distinct categories: wave-like or particle-like [24].
But, de Broglie’s equation, which relates a particle’s momentum (a feature of a particle) to its wavelength (a feature of a wave), promotes the idea that treating wave-like and particle-like as completely distinct phenomena is a mistake. But then, how is the nature of matter to be explained? If matter is neither distinctly particle-like nor wave-like, then what is it?

4.1.3.1 Wave-Particle Duality

Eventually, physicists settled on the concept of ‘wave-particle duality’. What is this? The idea arose primarily through the study of light. What, it was asked, is light? Newton thought it consisted of tiny particles that interacted with the eye in such a way as to produce images. However, later experiments conducted by Thomas Young (13 June 1773 – 10 May 1829) showed, conclusively, that light acted like a wave [69]. And Maxwell’s theory of electromagnetism described light as an electromagnetic wave. For just short of a century, Maxwell’s wave description of light remained fairly satisfactory. But late in the nineteenth and early twentieth century, it was shown that light, under certain conditions (photoelectric effect), displayed characteristics normally associated with a particle.

Suppose two small slots are cut horizontally into an armored plate in relatively close proximity to one another. If a machine gun sprays in random fashion a barrage of bullets at the plate, how is the manner in which the bullets pass through the slots described? This is a relatively complicated experiment, as some bullets will ricochet off the edges of the slots in unpredictable directions and the number of bullets passing through will depend on the size and shape of the bullets and the slots. Closing off one of the slots and setting up a detector behind the plate reveals where on the detector each bullet strikes. An experiment can then be performed that will divulge the most likely place for a bullet strike. Not surprisingly, this location turns out to be directly across from the opening. At positions farther away from this center location, bullets will strike with much less frequency. From a probability standpoint, the pattern that emerges on the detector is called a ‘probability distribution’. If the experiment is repeated many times, similar results will occur. This information can then be used to predict the probability that a bullet (think particle) strikes the detector at a certain location [116].

What happens if both slots are open? The pattern does not change. The impact pattern shows a similar probability distribution, the most likely spot on the detector is the location directly across from the two openings. Bullet strikes happen less often at off center positions (see figure 4.1.3.1-1) [117].
Now suppose a similar experiment is conducted only this time with light and a suitable detector. When a light beam passes through two small slits of a properly designed apparatus, what happens? If light is wave-like, it should have entirely different characteristics than a particle. There would be no center of mass. Waves would be spread out over large areas. Characterized by two opposite poles, crests and troughs, waves vacillate up and down or back and forth as they move through space and time. When two waves meet the vacillating motion can cancel or be reinforced depending on whether a crest of one wave meets the trough of another to cancel or two crests or two troughs meet adding together to create a larger crest or trough.

If the slits are opened one at a time, the pattern resembles the one for bullets - two distinct peaks. But when both slits are open, the light waves pass through both slits at once and interfere with each other; in phase, they reinforce each other; out of phase they cancel each other (see fig 4.1.3.1-2) [117].
There appears on the detector alternating light and dark bands – an interference pattern. Light bands appear where the waves have reinforced while dark bands appear where the waves cancel. The bands show more intensity in the center part of the detector, but less intensity at locations away from the center. If one slit is closed and the other left open, the interference pattern vanishes and a pattern emerges similar to the bullet pattern. This is because only one wave is produced, so there is no reinforcement or cancelling effect created by the interfering waves.

By the early twentieth century, physicists had invented the electron gun. This was essentially a vacuum tube, the same kind used in the first black and white television sets. At the time, there was little doubt that an electron was a particle [116]. It had mass, momentum and energy, just like any other particle. Like a billiard ball, electrons could transfer momentum and energy to other particles.

Using an electron gun, an experiment could be conducted that amounted to the two slot experiment. When the experiment was conducted with one of the slits closed, the results were virtually identical to the bullet experiment. Electrons with enough momentum dislodged other electrons from the detector. But when both slits were open, not only did electrons dislodge other electrons from the detector, but an interference pattern appeared - areas where no electrons were dislodged. This was unexpected – a very different result from the machine gun experiment. In classical physics, particles did not act like waves and vice versa (see Fig 4.1.3.1-3) [117]. What was going on?
Surprisingly, an explanation, at least one that would satisfy all or most of the influential physicists of the day, did not materialize. To this day, the explanation that was finally adopted, after much discussion, still invokes a debate.

In an effort to gain a deeper understanding, the electron gun experiment was modified and repeated. It was possible to slow down the experiment so that only one electron at a time was emitted from the gun, ensuring that the electron would pass through only one of the slits. But the modification had no impact on the results, which were the same as before; when both slits were open an interference pattern emerged [116].

Again, the experiment could be modified by directing a light beam just behind the openings to observe which slit the electron passed through. Every time an electron passed through one of the slits, a flash of light would appear. What was observed are electrons behaving normally – 50% of the electrons would pass through one of the slits and 50% through the other. But something else unexpected happened. If both slits were open and the light beam was turned on, the interference pattern on the detector disappeared [116].

Evidently, the directed light beam caused a disturbance, which effects qualitatively and irreversibly the behavior of the electrons. It is possible to minimize this disturbance by increasing the wavelength of the light beam, which decreases its energy so that the impact on the electrons is not quite as severe. But just about the time the light lacks sufficient energy, where it becomes difficult seeing which slit the electron passes through, the interference pattern reemerges [116]. No one has been able to construct an experiment where the exact path of an electron could be known and, at the same time, have an interference pattern appear. In this sense, it became difficult to determine if the electron acted like a wave or a particle, until and unless, it interacted with a measuring device. It was the electron’s interaction with a measuring device that
determined the nature of the electron, not the electron itself. This is the essence of wave-particle duality.

4.1.3.2 Schrödinger’s Cat

Faced with the grim fact that matter was not made from the simple substances described by the classical theories, the only choice was to find as reasonable an interpretation as possible, given the realities of the confounding experimental results. But what should that interpretation be?

The disposition that a quantum mechanical system can assume at any point in time is called a ‘configuration of states’. This configuration of states is usually called a ‘configuration space’. Schrödinger equation is a deterministic equation that describes the time evolution of a given configuration space. It is a deterministic equation because, given the configuration space, the time evolution of that space is determined exactly by Schrödinger’s equation. Mathematically, Schrödinger’s equation follows the law of superposition, which says that if \( \psi_1 \) and \( \psi_2 \) are both solutions to Schrödinger’s equation, then so is \( \psi_1 + \psi_2 \) [118].

To gain a better understanding of the idea of a quantum mechanical state, consider the parable of ‘Schrödinger’s cat’. A cat is placed in an opaque box so that it is impossible to see into the box without opening it. The box is equipped with a tube split into two branches; one branch leads to a capsule filled with poisonous gas, the other leads away from the gas. An electron is released into the tube. If the electron follows the branch leading to the capsule of poisonous gas, the interaction will cause the capsule to burst releasing the gas, hence, killing the cat. If the electron follows the path away from the gas, the cat lives. The cat’s configuration space consists to two states: \( \psi_1 = "\text{cat alive}" \), \( \psi_2 = "\text{cat dead}" \). The question is: after the electron is released into the box, is the cat in state \( \psi_1 \) or \( \psi_2 \)? There is only one way to find out; open the box and look inside. The cat will be found either dead or alive.

But one of the characteristics of Schrödinger equation is the law of superposition: if \( \psi_1 \) and \( \psi_2 \) are solutions to the equation, so is \( \psi_1 + \psi_2 \). In terms of the cat problem, what does \( \psi_1 + \psi_2 \) represent? Evidently, it must indicate that the cat is in some alternating state of deadness and aliveness. But, quantum mechanical systems resist any attempt at observing them in superposition [5]. When the box is opened, the cat will be either dead or alive, not in some nebulous state between deadness and aliveness. This is often referred to as ‘the collapse of the wave function’. An observation will, in some sense, cause a quantum system to jump to one of the possible single states. In the cat problem, this involves jumping to either state \( \psi_1 \) or \( \psi_2 \).

The superposition law serves to represent the uncontrollable - the lack of complete determinism in the quantum theory. In the Schrödinger’s cat parable, without opening the box, it cannot be determined with certainty whether or not the cat is dead or alive.

---

6 Strictly speaking, the most general solution is \( c_1 \psi_1 + c_2 \psi_2 \), where \( c_1 \) and \( c_2 \) are constants, but these technical details are not important for the discussion here.
Only betting odds can be given – 50% chance of being alive, 50% chance of being dead. The state of the cat is unknown until a measurement is taken and the only way that can happen is to look inside the box. Likewise, the quantum theory withstands any attempt to determine, with certainty, if a physical phenomenon of interest will show wave-like or particle-like properties until it interacts with a measuring device, that is, until an observation is made.

This way of looking at the problem suggested that the quantum theory has a holistic nature. In classical physics it is always possible to separate Nature into distinctly definable parts [24]. There is the observer and observed and it is not particularly problematic to formulate physical problems by making a distinction between observer and observed and to assume that the observer has little measurable impact on an experiment. The quantum theory resists any such attempt. Systems are to be thought of in a holistic fashion. Breaking problems into distinctly definable parts is meaningless. It is simply not possible to ignore the impact of an observation on any given experiment.

4.1.4 The Correspondence Principle

At this point, three non-classical ideas have been introduced: 1) state changes in the quantum realm, such as energy transfers, come in discrete units \( E = h \nu \) as opposed to gradual state changes characteristic of classical physics; 2) only the probability of a transfer can be determined between states; deterministic descriptions of the state transfers are not possible; 3) in general, quantum phenomenon must be treated holistically; it is not possible to break a quantum problem into distinct parts [24]. The situation begs the obvious question: if quantum mechanics is successful in describing microscopic objects, atoms and elementary particles, and classical physics successfully explains macro-systems, like springs and capacitors, how are these two approaches reconciled?

The quantum mechanical answer to this question is called the ‘correspondence principle’. It was first formulated by Bohr. This principle states that the laws of quantum physics must be chosen so that, where many quanta are involved, classical results on average must hold. The “many quanta are involved” statement is usually called ‘in the classical limit’.

Bohr’s statement of the correspondence principle is very general. The key phrase here is ‘in the classical limit’. It was not at all clear when Bohr adopted the principle precisely what “in the classical limit” meant [119]. It is still the subject of research and debate that goes beyond the scope of this book.

From a unification standpoint, the correspondence principle is not completely satisfactory. The fact that, when many quanta are present, classical methods of the calculus can be employed for obtaining satisfactory solutions to physical questions only implies that the errors encountered by adopting classical methods are sufficiently small and can be ignored. That the quanta will be packed so closely together so as to approximate a smooth transition does nothing to address the deeper philosophical
divergence between how classical and quantum theories describe the nature of physical phenomena.

4.2 The Copenhagen Interpretation

The Copenhagen interpretation is an attempt to explain the results of atomic experiments in terms of mathematical formulations. Heisenberg was employed as a lecturer in Bohr's Institute for Theoretical Physics at the University of Copenhagen. In the course of their collaborations, they, along with Max Born and others, formulated the ‘Copenhagen interpretation’ of the quantum mechanics [45].

There is no official Copenhagen interpretation. It will not be found in any formal document. In fact, Bohr and Heisenberg disagreed on some key aspects of how the experimental results and equations that describe them should be interpreted [119]. But roughly speaking, the Copenhagen interpretation can be characterized by the following principles:

1. A system is completely described by a wave function ‘\( \psi \)’, which represents an observer’s knowledge of the system

2. Descriptions of Nature are probabilistic and the probability of an outcome is related to \( \psi^2 \) - the square of the wave function

3. It is not possible to know the values of all the properties of the system at the same time; those properties not known with precision can only be described probabilistically

4. Matter can only be described as wave-like or particle-like after it has interacted with a measuring device e.g. after an observation has occurred and trying to describe matter as wave-like or particle-like before a measurement is meaningless

5. Quantum mechanical descriptions of large systems should, on average, agree with classical descriptions (correspondence principle)

How does the Copenhagen Interpretation explain the two slit electron gun experiment previously discussed? First, the electron must pass through one of the openings. In fact, the probability it passes through one of the two openings is 50%. Conversely, there is a 50% chance it passes through the other opening. This is a reflection of Principle ‘1’ discussed above - the completeness principle. It is certain the electron passes through one of the two openings (50% + 50% = 100%). It cannot be determined \textit{a-priori} which of the two openings the electron passes through. This reflects the probabilistic nature of the system (Principle 2).

Principle ‘4’ is the most controversial. It involves the collapse of the wave function. If no directed light beam is projected upon the slits, then the probability waves recombine
at the detector to form the interference pattern. In other words, the electron behaves like a wave. But when shining a light beam on the slits to see which opening the electron passes through, the wave function ceases to be a factor and it will be known with certainty which opening the electron passes through. This means there is no chance it passed through the other opening. In this case, the electron acts like a particle and the interference pattern on the detector vanishes.

It was Max Born, who in the 1920’s, suggested that the waves associated with the wave function ‘\( \psi \)’ should not be interpreted as the waves of some material substance like water or sound waves. Instead, they should be interpreted as measures of probability. Waves of probability relate to the uncertainty principle in the sense that what individual waves will do cannot be predicted with certainty. Only betting odds can be given [45].

This interpretation represented a complete break from the determinism of classical physics. It meant that the outcome of any experiment is only statistically predictable. The exact nature of each individual electron can never be determined. Those experiments designed to measure wave behavior will see the electron as a wave. Those experiments designed to measure the particle properties of the electron will see the electron as a particle. No experiment can be designed that measures both the wave and particle properties of the electron at the same time. This principle, as part of the Copenhagen interpretation, is known as the ‘complimentary’ aspect of quantum mechanics [120]. And it is the fundamental notion behind the idea of ‘wave-particle duality’.

The adoption of the Copenhagen Interpretation for quantum phenomena created a sharp divide between the classical physics of determinism and the new quantum physics. The divide spilled over into debates between those who accepted the new interpretation and those who didn’t. Amongst those who didn’t included some of the eminent physicists of the day, including Einstein, Schrödinger and de Broglie. They were pitted against Bohr, Heisenberg and Born among others. Einstein, in particular, could not accept the indeterminacy of the quantum theory. In his famous debates with Bohr, he was said to proclaim “God does not play dice …”, to which Bohr reportedly replied, “Stop telling God what to do …”.

4.3 The Measurement Problem in Quantum Mechanics

Quantum mechanics is characterized by two diametrically opposed processes. There is the deterministic unitary process described by the Schrödinger equation, whereby a superposition of quantum states evolves in time in a certain and determined manner. Under the Copenhagen interpretation no physical reality is associated with this process. It simply gives the probabilities that potential outcomes actually materialize.

The second process is described by a quantum state reduction that brings forth a specific outcome. This takes place when a measurement is performed. Roger Penrose called the two processes the ‘\( U \)’ and the ‘\( R \)’ process respectfully. How an outcome ‘\( R \)’ can develop from the process ‘\( U \)’ is the crux of the measurement problem [5]. The
measurement problem is not unlike the torturous grappling's by 17th century philosophers as they tried vainly to explain how an 'non-material' idea could cause a physical reaction to take place – like when a student raises a hand in response to a teacher's question.

The measurement problem in quantum mechanics is the unresolved issue of how (or if) wave function collapse occurs. This corresponds to transitioning from $U$ status into $R$ status. The inability to observe this transition directly has given rise to different interpretations of quantum mechanics. The wave function in quantum mechanics evolves according to the Schrödinger equation into a linear superposition of different states ($U$ process), but actual measurements always find the physical system in a definite state ($R$ process). Any future evolution is based on the state the system was in when the measurement was made. This implies that the measurement "does something" to the process under examination [5]. Whatever that "something" might be is not explained by quantum mechanics.

Returning to the parable of Schrödinger's cat, under the $U$ process, the cat is evolving into a linear superposition of states characterized by "live cat" and "dead cat" (the cat seems to be in a "mixed" state) both dead and alive at the same time. However, a particular observation of the cat finds it either alive or dead. The question is: how are the probabilities associated with the $U$ process converted into an actual sharply well-defined reality associated with the $R$ process [5]?

According to Bohr, this jump represented a simple increase in the knowledge of the observer. When an observation is made, the observer becomes aware of the actual reality of the situation and sees a real live cat or a dead one. Many physicists find this explanation unsatisfying. According to Bohr's interpretation, the transition from the $U$ status to the $R$ status takes place wholly in the mind of the observer. The cat does not physically jump from its $U$ state to the physical reality of being either alive or dead. In fact, it is not clear that quantum mechanics has anything to do with the mind of an observer. The $U$ process, while not physical, gives probabilities of real physical outcomes and the $R$ process represents an actual real outcome. What did either of these processes have to do with a mind?

This suggested that Bohr's version of quantum mechanics was deeply flawed. While it does describe what happens when an observer takes a measurement, the observer and the act of measurement are treated classically. This is surely wrong. The observer and the measuring device are made of the same materials (protons, neutrons and electrons) that all other things are made of. They should be governed by a quantum mechanical wave function. And this is entirely missing from Bohr's explanation. Physicists and their apparatus must be governed by the same quantum mechanical rules that govern everything else in the Universe. These rules are expressed in terms of a wave function (or, more precisely, a state vector) that evolve in a perfectly deterministic way. So where do the probabilistic rules of the Copenhagen interpretation come from? These shortcomings in Bohr's interpretation led to efforts at finding a better explanation [121].
In opposition to Bohr, Wolfgang Pauli (25 April 1900 – 15 December 1958) and Heisenberg believed that it was the observer that produced the collapse. Pauli described quantum mechanics as lucid mysticism [122]. The involvement of consciousness in the collapse of the wave function can be summarized thusly:

*The rules of quantum mechanics are correct but there is only one system (one wave function) which may be treated with quantum mechanics, namely the entire material world. There exist external observers which cannot be treated within quantum mechanics, namely human minds, which perform measurements on the brain causing wave function collapse [123].*

This interpretation attributes the process of wave function collapse to consciousness itself. Specifically, a non-physical mind is postulated to be the only true measurement apparatus [123]. Since consciousness is not subject to physical law and, hence, is not quantum mechanical, the ‘consciousness causes collapse’ argument appears to overcome the objection that observers and measuring devices should be treated quantum mechanically.

This viewpoint leads to questions that are deeply philosophical. In many philosophies, the conscious mind is seen as a separate entity, existing in a realm not described by physical law. This perspective was first introduced in detail by Descartes in the 17th century. Those who support this view maintain that the description of the physical world provided by quantum mechanics suggests a kind of mind/body dualism. Of course, it would seem that the proponents of this way of thinking would run into the same difficulties in defending it as did Descartes. He failed to adequately describe the mechanism which allowed a non-physical mind to affect the actions of a physical body, which it surely does. And as far as I can make out, this interpretation of the quantum theory adds nothing that would resolve of this age old question.

Beyond that, as Hugh Everett points out [124], difficulties arise when the Universe is conceived as possessing more than one observer. Observers, depending on their various perspectives, do not necessarily perceive identical realities even when witnessing the same physical system. If it is indeed consciousness that causes us to perceive reality, each observer should possess their own individual consciousness and the reality associated with that conscious mind would be at least in part subjective. But then, why do some observers not see a green moon, others a red moon and still others a yellow moon, for example? Evidently the answer lies in the meshing of the quantum mechanics with Eastern Mysticism, including Hinduism, Taoism and Buddhism. All these beliefs include the idea of a transitory, interconnectedness in all of Nature - a shared consciousness, and hence, experience one reality. There is, according this belief, a single consciousness shared by all. And this, it would seem, overcomes the problem of having numerous realities embodied in many different conscious observers. However, it was soon recognized that this point of view was difficult to square with the other aspects of physics. While the ‘consciousness causes collapse’ doctrine holds an important role for the conscious mind, it is difficult to explain how the earlier expansion of the universe evolved those conscious minds. As John Stewart Bell put it (quiet
sarcastically): "Was the wave function waiting to jump for thousands of millions of years until a single-celled living creature appeared? Or did it have to wait a little longer for some highly qualified measurer - with a PhD?" [125].

4.3.1 The Many Worlds Interpretation

One way to handle the measurement problem is to deny that the wave function collapses at all. Hence, the problem of explaining the mechanism by which wave function collapse occurs simply vanishes. And this is exactly what Hugh Everett proposed [124]. He did this by merging the micro and macro worlds. The observer became an integral part of the system by introducing a universal wave function that included observers as well as the objects observed as parts of a single quantum system. In Everett's system the continuous evolution of a wave function is not interrupted by acts of measurement. Everett’s conjecture became known as the “many-worlds” interpretation. It asserts that the universal wave function is objectively real and denies wave function collapse. This implies that all the possibilities embodied in the wave function actually materialize in the form of alternative worlds or universes. The universal wave function contains branches for every possibility. Each branch has its own copy of the observer and each copy perceives one of the possibilities as the outcome. According to a fundamental mathematical property of the Schrödinger equation, once formed, the branches do not influence one another. Each branch embarks on a different future, independently of the others. This branching or splitting of the quantum possibilities into independent worlds each having their own classical reality has generally become known as ‘decoherence theory’. The many-worlds interpretation views reality as a multi-branched tree wherein every possible quantum outcome is realized [126].

There are two principal objections to the many-worlds interpretation. First, it is not clear when and how the splitting or the acts of decoherence happen. The present day understanding of decoherence is not a precise self-contained explanation. There are many different interpretations of the quantum theory based on slightly different versions of decoherence. And it is not clear which version is the correct one. Secondly, since each of the many-worlds is self-contained, they do not interact with one another. This raises questions about how the many-worlds conjecture can make testable predictions and be subject to falsification. It is not at all clear that the many-worlds interpretation buys much in terms of solving the measurement problem. On the one hand, it does remove the necessity of explaining how the wave function collapses, only to give back that advantage by postulating a multiverse of worlds, only one of which the observers in that world have any direct knowledge of.

4.3.2 Environmental Decoherence

The many-worlds interpretation shares similarities with later "post-Everett" interpretations of quantum mechanics that use decoherence to explain the measurement process. While the “many-worlds” treats the alternative worlds as real, since the wave function obeys a deterministic wave equation, the other decoherent
interpretations either regard the extra quantum worlds as metaphorical or are agnostic about their reality [124,127].

An interpretation that has gained popularity recently is called ‘environmental decoherence’. Environmental decoherence attempts to explain the transition from the quantum ‘U’ process to the classical ‘R’ process by analyzing the interaction of a system with a measuring device or with the environment. To clarify, imagine a quantum mechanical particle or system of particles as an isolated system floating in empty space. This simplification may be fine in some cases, but in the real world, there is no such thing as an isolated system. Typically a particle in flight will collide with air molecules or will emit thermal radiation that gets absorbed by the environment. Any interaction with the environment leads to an entanglement between the particle's state and the environment's state. As the entanglement diffuses throughout the environment, the total state can no longer be separated into the direct product of a particle ‘state’ and an environment ‘state’. What was once a superposition of particle states becomes a superposition of particle states and environmental states. At this point the particle ceases to act as if it were in a quantum superposition of states, instead acting as a statistical ensemble of states [128].

The end result of the decoherence process is that the particle will appear to have collapsed in a manner described by the Born probability law. Born's law then becomes a consequence of the interaction between particle and environment. In essence, unlike the many-worlds interpretation, environmental decoherence involves at least the appearance of wave function collapse caused by the particle interacting with its environment.

To understand better how environmental decoherence is supposed to work, consider this well-known and important property of quantum mechanics: a quantum mechanical superposition of states is fundamentally different from a classical ensemble of states. In classical physics, an object is never both “here” and over “there”. It is either “here” or over “there”. Objects are never in two places at once. So, in classical physics, the probability of an object being “here” and over “there” is zero. However, recall the two-slit interference experiment discussed previously where electrons pass individually (one at a time) through a double slit. Within the standard quantum-mechanical formalism, the electron must not be explained by either one of the wave functions describing the passage through a particular slit (ψ₁ or ψ₂), but only by the superposition of these wave functions ‘ψ₁ + ψ₂’. The correct density distribution of the pattern on the screen is not given by the sum of the squared wave functions individually i.e. P = |ψ₁|^2 + |ψ₂|^2, but by the square of the sum of the individual wave functions: P = |ψ₁ + ψ₂|^2. So, from a quantum mechanical standpoint, the electron is treated as though it were at two places at once. Of course, the “two places at once” superposition state of the electron is not observable. In the Copenhagen interpretation of the quantum formalism, when an observation is made, the electron “jumps” from a superposition of states to one of the singular states ‘ψ₁’ or ‘ψ₂’.
In opposition to the Copenhagen interpretation, environment decoherence argues that when a system in a superposition of states \( \psi_1 + \psi_2 \) becomes entangled with its environment, it quickly decoheres into an ensemble of classical states. In other words, it quickly devolves from \( P = |\psi_1 + \psi_2|^2 \) to \( P = |\psi_1|^2 + |\psi_2|^2 \). There is a well-defined mathematical process called a ‘density matrix’ that describes this transformation. The classical states can then be given in terms of classical probabilities, where the probability of observing the electron in superposition quickly becomes zero.

It is immediately clear why environmental decoherence has become popular with many physicists. It effectively removes the mysticism of the ‘observer causes collapse’ interpretation. And because the system and the environment are both treated quantum mechanically, it addresses the objections to the Bohr interpretation. Because of this, it is often argued that environmental decoherence effectively solves the measurement problem.

There are about as many opinions on this question as there are different approaches to environmental decoherence and those are many. It is better to describe environmental decoherence as a program rather than a single technique. There are so many variants to this approach, it is difficult to keep them straight. The key question seems to be: what constitutes the “environment” and what constitutes the “system”? The many variants of environmental decoherence treat these two key components differently.

Whether or not environmental decoherence solves the measurement problem appears to largely depend on point of view. Physicists with a “positivist” outlook, those who are not overly concerned with ontological interpretations of physical events, will likely be more inclined to answer the question in the affirmative. The issue lies within the nature of the density matrix, which is the mathematical tool for proceeding through the environmental decoherence process. Density matrices are not unique. The same density matrix can describe more than one system. While a system composed of “spin up” and “spin down” electrons in an EPR environment is certainly ontologically different from one composed of “spin right” and “spin left” electrons, their probability distributions, and hence, their density matrices are identical. As far as I can tell, the following quote expresses more or less the prevailing opinion on whether or not environmental decoherence solves the measurement problem:

\[
\text{Does decoherence solve the measurement problem? Clearly not. What decoherence tells us, is that certain objects appear classical when they are observed. But what is an observation? At some stage, we still have to apply the usual probability rules of quantum theory [131].}
\]

The previous paragraphs discussed some interpretations of the quantum mechanics. Some employ the idea of decoherence and some don’t. In addition to the interpretations discussed above, consistent histories, de Broglie–Bohm theory, relational quantum mechanics, transactional interpretation, stochastic mechanics, objective collapse theories, many minds, quantum logic, modal interpretations of quantum theory, time-symmetric theories are other current interpretations of quantum
mechanics. The number of interpretations indicates that physicists are not in general agreement as to what the mathematical formalism of the quantum theory actually implies.

Viewpoints on this issue fall roughly into three categories: 1) those who maintain that the interpretation problem is primarily a philosophical issue that remains outside the domain of physics; 2) those who maintain that physics is primarily concerned with whether or not a theory predicts the correct results of measurements and anything beyond that is outside the domain of physics; 3) those who maintain that, in its present form, quantum mechanics does not make sense and needs to be ontologically reinterpreted or revamped in a way that does make sense.

As it stands, quantum mechanics consists primarily of a mathematical formalism that explains a broad range of experimental results and, in fact, explains most of what is known about the Universe. And if your viewpoint falls within category ‘1’, then the quantum theory in its current form might seem entirely sufficient. Those who fall within category ‘2’ are inclined to agree with the category ‘1’ advocates accept as it pertains to the measurement problem. Since quantum mechanics is supposed to predict the correct outcome of a measurement, at the very least, advocates should be able to say what a measurement is. Although they sometimes have difficulty expressing it, category ‘3’ advocates believe, despite its successes, that the quantum theory is, on some level, deeply flawed. The mathematics is inconsistent and the complementarity aspect of the quantum theory violates the basic principles of propositional logic. What are observed in the real world are not probabilities, but real physical outcomes that have a basis in an external reality that the quantum theory cannot quite touch. Not only does the measurement problem have to be resolved, but the problem of quantum entanglement, as illustrated in the EPR paradox, which seemingly violates principles of local causality [132].

Within the wave/particle duality concept there is no logical picture, at least one obeying classical propositional logic, which can simultaneously describe all the properties of a quantum system ‘S’. This is often phrased by saying that there are “complementary” propositions ‘A’ and ‘B’ that can each describe S, but not at the same time.

Fundamental to propositional logic is the idea of an “atomic statement” or proposition. An atomic proposition is a statement that ‘can be the case’ or ‘not be the case’ and all other things remain the same [133]. ‘It is raining’ is often cited as an example of an atomic proposition. It can be either true or false. But the “complementary” propositions ‘A’ and ‘B’ within the quantum theory are not atomic propositions. In fact, if A is true, then B is necessarily false and vice versa. For example if A = “I get an exact position measurement” and B = “I get an exact momentum measurement”, then according to the uncertainty principle, if A is true, B is not only false, but indefinite. So, the truth of B, in essence, depends on the truth of A. This seems to violate the very foundation of propositional logic upon which most mathematical systems are based.
4.4 The EPR Paradox and Bell’s Theorem

To Einstein, the wave-particle duality in quantum phenomenon was due to a fundamental lack of knowledge. Thought Einstein, there were reasons, although possibly hidden from view that, once known, would solve the wave-particle paradox and restore determinism to physics. And he set out to show that the quantum theory was, at best, an incomplete theory. His efforts, along with collaborators Podolsky and Rosen, culminated in a thought experiment known as the ‘EPR experiment’. But once the EPR thought experiment was examined and tested, the results would swing the evidence dramatically in favor of the complementarity aspect of the quantum theory and against the idea of local realism – the position Einstein favored.

4.4.1 The EPR Paradox

The now famous paper published in 1935 by Einstein, Podolsky and Rosen (EPR) “Can Quantum-Mechanical Description of Physical Reality be Considered Complete” discussed what Einstein termed a “spooky action-at-a-distance”. At the time, a number of experiments had revealed that electrons displayed a strange correlation effect. They seemed capable of magically communicating with one another over relatively large distances - distances large enough that the observed correlations could not be explained by claiming that the electrons could send messages to one another. Einstein knew from his work on the special theory of relativity that not only was the speed of light finite, but nothing could travel faster. And if nothing could travel faster than light, neither could messages be sent nor information exchanged in less time than it took for light to travel from one place to the next. Thought Einstein, if the correlations between quantum objects could not be explained by postulating that, somehow, information was being exchanged between quantum objects, what was the explanation? The quantum theory didn’t seem to provide an adequate answer. The paper went on to argue that there must be something at work, whether measurable or not, that would explain the “spooky action-at-a-distance” correlations. And since the quantum theory provided no explanation of this 'something', it was at best incomplete and, at worst, outright wrong.

4.4.1.1 The Stern-Gerlach Experiments

In 1922, Otto Stern (17 February 1888 – 17 August 1969) and Walther Gerlach (1 August 1889 - 10 August 1979) conducted experiments related to the deflection of particles. The experiments showed that electrons and atoms have intrinsically quantum properties. In addition, they demonstrated how, at the quantum level, the act of taking a measurement affects the system being measured.

The experiments involved sending a beam of particles through an inhomogeneous magnetic field and observing their deflections. The results showed that particles possess an intrinsic angular momentum that is most closely analogous to the angular momentum of a classically spinning object. This ‘intrinsic angular momentum’ is referred to as ‘spin’.
If the magnetic field is inhomogeneous, the force on one end of the dipole will be slightly greater than the opposing force on the other end. This imbalance results in a net force, which deflects the particle’s trajectory. If the particles were classical spinning objects, the expected amount of spin would be random and continuous. Each particle would be deflected by a different magnitude, producing a smooth distribution on a detector screen. Instead, the particles passing through the Stern-Gerlach apparatus were found deflected either up or down by a specific amount. These results suggested that spin was a quantized phenomenon.

Electrons are spin-1/2 particles. They have only two possible spin values measured along any axis ($\pm \hbar/2$). Had these values been classical in nature as, for example, particles rotating the way a planet rotates, the individual particles would spin impossibly fast. The electron radius measured at 2.8 fm (the classical electron radius) would rotate at $2.3 \times 10^{11} \text{ m/s}$ - greatly in excess of the speed of light ($2.998 \times 10^8 \text{ m/s}$). And this would be impossible [134]. Instead, spin is considered a purely quantum mechanical phenomenon.

As electrons pass through the Stern–Gerlach device, they are measured, say, by a detector or by an observer who records spin measurements. In this case, two possible values can be observed, either ‘spin-up’ ($+\hbar/2$) or ‘spin-down’ ($-\hbar/2$). If electron spin is not filtered out in any way, 1/2 the electrons would be found spin-up, the other half spin-down. Imagine a detector with a filter that prevents, for example, all the spin-down electrons from arriving at the detector. If a measurement was made after the electrons passed through the filter, only spin-up electrons would be detected. If the filter is rotated by 180°, only spin-down electrons would be detected [135].

Suppose a second filter is placed behind the first. If the second detector is set to filter out spin-down electrons as does the first detector, the arrangement of the detectors has no effect on the outcome of the experiment. All electrons passing through the second filter will be spin-up. However, if the second filter is rotated by 180°, no electrons will be detected. The first filter will block all the spin-down electrons and the second filter will block all the spin-up electrons. Since electrons are either spin-up or spin-down, no electrons are detected [135]. This demonstrates the quantized nature of spin in electrons. Filters designed to block electrons based on spin orientation only need two discrete settings to block electrons from passing through the filters.

Now suppose a third detector is added. Assume the detectors are aligned sequentially and are set at 0°, 90°, 180° respectfully. In this case, the first detector is said to detect spin in the $+z$-direction. The second detector measures spin in the $x$-direction and the third detector measures spin the $-z$-direction. The electrons that emerge from the 0° (first) detector are all spin-up, since all the spin-down electrons have been filtered out. All the electrons that pass through the first detector will pass through the 90° detector, but 1/2 of the electrons will emerge from the second detector with spin in the $+x$-direction and 1/2 will emerge with spin in the $-x$-direction [135]. This is a uniquely quantum mechanical effect. It is an example of the uncertainty principle. Spin direction is a conjugate variable and these variables cannot be measured simultaneously. Only
probabilities for conjugate variables can be given. In this case, if an exact \( +z \) spin measurement is made in the \( z \)-direction, spin measured in the \( x \)-direction becomes completely uncertain. There will be a 50% chance of getting a \( +x \) spin measurement and a 50% chance of getting a measurement in the \( -x \)-direction in the second detector.

Suppose the 90° (second) detector, measures spin in the \( +x \)-direction (a filter is added to block electrons in the \( -x \)-direction), 1/2 of the \( +x \) electrons emerging from the 90° detector will also emerge from the third detector. This illustrates how taking a measurement disturbs a quantum system. The system loses its memory of what happened in the first detector. Remember, the first detector let only \( +z \) electrons through. But when a \( +x \)-direction measurement is recorded in the second detector, the system completely forgets about the first filtering. This forgetfulness can be understood by noting that if the second detector is removed, no electrons pass through the 180° detector. Taking a measurement impacts the state of a quantum system [135].

Stern-Gerlach type experiments can be designed to examine the correlation effect of spin in electrons. Spin is a conserved quantity which means, barring outside influences like external forces, the total spin of a system must remain the same at all times. Suppose, before electrons pass through a magnetic field, their total spin is zero. If 10 electrons emerge from the magnetic field as spin-up, there must be 10 that emerge as spin-down, so that the total spin remains zero. Einstein believed that there were unknown reasons (hidden variables) that would explain the conservation of spin in electrons and other radiating particles that the quantum theory could not account for.

To get a deeper understanding of the correlations between quantum spinning objects, imagine that two 0° detectors are separated by a distance great enough that the two observers stationed at each detector and taking measurements cannot communicate with one another in time to change the results of any of their measurements. Imagine a radioactive substance originally in a spin 0 state decaying into two electrons emitted nearly simultaneously, one headed toward one of the detectors, the second headed toward the other detector. If the decay process is continuous, there will be a continuous stream of electron pairs each heading toward separate detectors. In the type of decay considered here, 1/2 of the electrons incident upon each detector are detected. In addition, if an electron is detected by a given detector, its entangled partner is not detected by the other detector [135]. This is illustrative of the conservation of spin. There are only two types of spin: spin-up and spin-down. If one of the electron pairs is ‘spin-up’, the other must be ‘spin-down’ in a given direction: total spin zero. The 0° detector detects only ‘spin-up’ electrons. If an electron is detected by a given detector, it is spin-up. Its partner must be spin-down in that direction, and hence, not detectable by the other detector.

Suppose one of the detectors is rotated and made a 180° detector. In this case, 1/2 of the electrons incident upon each detector are detected. But this time, if an electron is detected by a given detector, its entangled partner is also detected by the second detector. If an electron was not detected by a given detector, its partner, likewise, would not be detected.
Suppose one of the detectors is rotated to a 90° detector measuring spin in the \(+x\)-direction. In this case, \(\frac{1}{2}\) of the electrons incident upon the 0° detector are detected while \(\frac{1}{2}\) of the electrons incident upon the 90° detector are detected. But if an electron is detected by the 0° detector, \(\frac{1}{2}\) of the time its partner is detected by the 90° detector and \(\frac{1}{2}\) of the time it is not detected. This is illustrative of the uncertainty principle. If a measurement is obtained for spin in the \(+z\)-direction, it is impossible to predict what the spin measurement (spin-up or spin-down) will be in the conjugate \(x\)-direction.

### 4.4.2 The EPR Argument

These sorts of correlations or “spooky action-at-a-distance” between quantum objects have been given the name ‘quantum entanglement’. Quantum entanglement is one of the most bizarre, misunderstood and controversial aspects of the quantum theory. There are basically two issues involving quantum entanglement: 1) how to make sense of quantum entanglement in terms of ideas that can be comprehended as a fundamental feature of our Universe? In other words, the concept of quantum entanglement seems to make no sense in terms of what is actually experienced in the world; 2) why, if most quantum states are entangled, are these entanglements never experienced in daily life [5]?

Einstein agreed. The 1935 *EPR* paper argued that there was a flaw in the Copenhagen interpretation i.e. that only the position or momentum of a particle, but not both, could be known at the same time with certainty (wave-particle duality). The *EPR* thought experiment involved two systems that initially interact with each other, but then become separated. The position or momentum of one of the systems is measured, and due to the known relationship between the measured value of the first particle and the second particle, the observer becomes aware of that value in the second particle. A measurement of the conjugate state is then made on the second particle, and, once again, due to the entangled relationship between the two particles, that value is then known in the first particle. This outcome seemed to violate the uncertainty principle, as both the position and the conjugate momentum of a single particle could be known with certainty [136]. For example, two particles, ‘\(A\)’ and ‘\(B\)’, interact briefly and then move off in opposite directions. It is possible to measure the exact position of particle ‘\(A\)’ and the exact momentum of particle ‘\(B\)’. Here, there is no violation of the uncertainty principle. But because quantum states are entangled, with the exact position of particle ‘\(A\)’ known, the exact position of particle ‘\(B\)’ can be known. Similarly, with the exact momentum of particle ‘\(B\)’ known, the exact momentum of particle ‘\(A\)’ can be worked out. The *EPR* argument claimed it had proved that particle ‘\(B\)’ can have simultaneously exact values of position and momentum [137]."

To see how this argument relates to electron spin, suppose an electron pair is sent toward two spin detectors. One detector is set to measure spin in the \(+z\)-direction and the other in the \(+x\)-direction. If an electron is detected as having spin \(+z\), then its partner has spin \(-z\) in the \(z\)-direction. Measurement is not required. Correspondingly, if an electron is detected as having spin in the \(+x\)-direction, then its partner has spin in
the $-x$-direction. But according to the uncertainty principle, an exact measurement of spin in the $z$-direction precludes an exact measurement in the conjugate $x$-direction. The EPR argument, so it seemed, had beaten the uncertainty principle, as it showed conclusively that conjugate variables could be known with certainty at the same time.

The EPR thought experiment was completely at odds with the quantum theory. According the Copenhagen interpretation, both the momentum and position of a particle was governed by a single wave function. When, say, a position measurement was taken on the particle, the act of measuring caused an instantaneous collapse of the wave function. At that point, the wave function ceased to be a factor, making any kind of exact momentum measurement impossible or, in other words, measuring momentum became completely uncertain. This was the EPR paradox. The quantum theory predicts that both values cannot be known simultaneously, and yet the EPR experiment shows that they can. Therefore, the quantum mechanical description of physical reality, the EPR paper concluded, was incomplete [137].

The EPR argument should remind the reader of the Stern-Gerlach correlation experiments discussed previously. Like spin in different directions, position and momentum are, according to the quantum theory, conjugate variables. It is impossible to measure the exact location and the exact momentum of a particle at the same time. In the correlation experiments discussed previously, if an electron is detected by a $0^\circ$ detector, it is a $+z$ (spin-up) electron. Once this information is known, it is certain that its entangled partner is a $-z$ (spin-down) electron. In fact, whatever axis spin is measured along, the spin of the two partner particles are always found to be equal and opposite. This can only be explained if the particles are linked in some way. Either they were created with a definite (opposite) spin about every axis—a "hidden variable" argument or they are linked so that one electron "feels" which axis the other is having its spin measured along, and becomes its opposite about that one axis — an "entanglement" argument. There are only two possibilities: either the electron partner knows that a measurement has taken place, or it has a definite spin already about a second axis. Einstein, Podolsky and Rosen asked: how can the second particle "know" to have a precisely defined momentum, but uncertain position? Since this implies that one particle is communicating with the other instantaneously across space faster than light speed, and because the quantum theory had no explanation for this "spooky-action-at-a-distance", it must be missing something. This was the crux of the EPR argument [138].

The EPR argument is based on the principle of locality, which states that physical processes occurring at one place should have no immediate effect on the elements of reality at another location. This is a consequence of special relativity, which states that information cannot be transmitted faster than the speed of light without violating causality. Causality is the relationship between an event (the cause) and a second event (the effect), where the second event is understood as a consequence of the first [139]. It is generally believed that any theory which violates causality would also be internally inconsistent, and thus, deeply unsatisfactory. It seems that the usual rules for combining quantum mechanical and classical descriptions violate the principle of locality.
without violating causality. In the correlation experiments, causality is preserved because there is no way to transmit messages between detectors to allow the manipulation of the measurement results. Therefore, neither the EPR experiment nor any quantum experiment demonstrates that faster-than-light signaling is possible. What, then, is the explanation for the apparent faster than light communication between entangled quantum objects? Evidently, the quantum theory predicts such entangled states in its mathematical formulism, but without providing an explanation of how these entangled states arise.

4.4.3 Bell’s Inequality

Almost 30 years passed before any definitive progress toward a resolution of the EPR paradox would appear. But in 1964, Northern Ireland physicist John S. Bell (28 June 1928 – 1 October 1990) published a paper that contained Bell’s Theorem. To understand the significance of Bell’s Theorem, it is first necessary to gain an appreciation for the inequality the theorem is based on.

Suppose an object, any object, can possess three distinctly different characteristics, which can be labeled ‘A, B’ and ‘C’ respectfully. Then symbolically, Bell’s inequality can be written:

\[ N(A, \sim B) + N(B, \sim C) \geq N(A, \sim C) \]

The above inequality is shorthand for stating that: the number of objects that possess characteristic ‘A’, but not characteristic ‘B’ (the “\( \sim \)” symbol stands for “not”) plus the number of objects that possess characteristic ‘B’, but not characteristic ‘C’ is greater than or equal to the number of objects that possess characteristic ‘A’ but not characteristic ‘C’.

Bell’s inequality is simply an exercise in propositional logic. Examine the Venn diagram in fig. 4.4.3-1. In classical logic, an argument is valid if its conclusion is entailed by its premises. In Bell’s inequality ‘\( N(A, \sim B), N(B, \sim C) \)’ are the premises and ‘\( N(A, \sim C) \)’ is the conclusion. In the Venn diagram below, there are three circles labeled ‘A, B, C’ divided into distinct areas marked \( a_1, a_2, a_3 \) … The individual circles can be thought of as representing each of the three possible characteristics (A, B or C) an object can possess. Then \( a_1 \) represents an object possessing characteristic ‘A’, but not characteristics ‘B’ or ‘C’. Area ‘\( a_2 \)’ signifies an object possessing characteristics ‘A’ and ‘B’, but not characteristic ‘C’. Similarly, area ‘\( a_4 \)’ signifies an object possessing all three characteristics.
In terms of Bell's inequality, as it relates to the Venn diagram,

\[ N(A, \sim B) = \text{"the number of objects that fall within the areas 'a1' and 'a5'"} \]
\[ N(B, \sim C) = \text{"the number of objects that fall within the areas 'a2' and 'a3'"} \]
\[ N(A, \sim C) = \text{"the number of objects that fall within the areas 'a1' and 'a2'"} \]

Remembering that Bell's inequality is written

\[ N(A, \sim B) + N(B, \sim C) \geq N(A, \sim C), \]

by substituting the Venn diagram values for their equivalent symbols in Bell's inequality gives

\[ N(a1 + a5) + N(a2 + a3) \geq N(a1 + a2). \]

This inequality is equivalent to Bell's inequality, but written in Venn diagram representation. It says symbolically that the number of objects that get counted in areas 'a1 + a5' plus the number of objects that get counted in areas 'a2 + a3' is greater than or equal to the number of objects that get counted in areas 'a1 + a2'. It should be obvious that the inequality is logically true, since the right-hand side of the inequality (the conclusion) is entailed in the left-hand side (the premises). And this, stated previously, is the definition of a valid logical argument.

David M. Harrison of the University of Toronto gives an excellent example of an application of Bell's inequality [740]. In the Harrison example, the objects consist of students at a university. The characteristics are:

A. Male
B. Taller than 5’8”
C. Blue eyes

Bell’s inequality states that the number of male students who are not taller than 5’8” plus the number of students (male or female) who are taller than 5’8”, but do not have blue eyes is greater than or equal to the number of male students who do not have blue eyes. According to Bell’s inequality this relationship between students is always true.
The validity of Bell’s inequality depends on two assumptions: 1) that the nature of the problem under consideration can be described using classical propositional logic. In other words, the structure underlying the problem is in one-to-one correspondence with classical propositional logic. If the structure underlying the problem was incompatible with classical propositional logic, then classical propositional logic would be an invalid way of reasoning about this particular problem; 2) the characteristics of the objects under consideration exist (are true or false) whether they are observed or not. For example, it can be shown that:

\[ N(A, \sim B, \sim C) + N(A, B, \sim C) = N(A, \sim C) \]

Note that the right-hand side of the above equation is equal to the right-hand side of Bell’s inequality. Even though the characteristic ‘B’ on the right-hand side of Bell’s theorem is not directly observed, the assumption is that either \( \sim B \) or \( B \) is true for every object \([140]\) or, in this case, every student.

The reader by now is probably wondering what all this has to do with physics. As it turns out, Bell’s inequality can be applied to the Stern-Gerlach type spin correlation experiments discussed earlier. Recall that in those experiments there were three detectors aligned at 0°, 90°, 180° respectfully. Two detectors are separated by a distance great enough that the two observers stationed at each detector and recording spin measurements could not communicate with one another in time to influence the results of a measurement. A radioactive substance originally in a spin-0 state decays into two electrons emitted nearly simultaneously, one electron headed toward one of the detectors, its partner headed toward the other detector. In any direction, there are only two types of spin measurements in electrons: ‘spin-up’ or ‘spin-down’. If one of the electrons in a pair is spin-up, the other must be spin-down: total spin zero. The detectors can be constructed so that each detector measures spin in a certain direction. If an electron is detected by a given detector, it is ‘spin-up’ in that direction. Its partner must be spin-down along the same direction.

It is customary in experiments involving Bell’s theorem to divide the detection angles in half, so instead of 0°, 90°, 180°, the angles are 0°, 45°, 90°. This change is necessary because one of the objectives of the experiment is to test the uncertainty principle. The uncertainty principle involves conjugate variables. According to the quantum theory, if an exact measurement is obtained on one of these variables, its conjugate becomes completely uncertain. But in the 0° – 180° detector configuration, recall that if there is detection at 0°, it is certain there will be detection at 180°. Therefore, the 0° – 180° detector configuration is not a conjugate configuration. Cutting the detection angles in half overcome this limitation.

The way in which Bell’s inequality is applied to the electron spin problem is to count the number of electron spin detections in various detector configurations and then examine the results. The characteristics, in this case, are the detection angles:
Hence, in spin correlation experiments, Bell’s inequality can be written

\[ N(0^\circ, \sim 45^\circ) + N(45^\circ, \sim 90^\circ) \geq N(0^\circ, \sim 90^\circ) \]

The degree numbers in parentheses represent the detector configurations. For example, \( N(0^\circ, \sim 45^\circ) \) indicates that the first detector is set to \( 0^\circ \) and the second detector set to \( 45^\circ \) and \( N(45^\circ, \sim 90^\circ) \) indicates that the first detector is set to \( 45^\circ \) and the second detector is set to \( 90^\circ \). So in terms of the electron spin, the values inserted into Bell’s inequality are the number of detections registered in the \( 0^\circ \) detector, but not in the \( 45^\circ \) detector when the detectors are in the \( 0^\circ - 45^\circ \) configuration. Added to that amount is the number of detections registered in the \( 45^\circ \) detector, but not in the \( 90^\circ \) detector when the detectors are in the \( 45^\circ - 90^\circ \) configuration. According to Bell’s inequality, that sum should be greater than or equal to the number of detections registered in the \( 0^\circ \) detector, but not in the \( 90^\circ \) detector when the detectors are in the \( 0^\circ - 90^\circ \) configuration.

The experiments are usually performed by first setting the detectors to the \( 0^\circ - 45^\circ \) configuration, then sending a billion electron pairs through the detectors and recording the results. The detectors are then set to the next configuration and another billion electrons are sent through the detectors and those results recorded and so on. The upshot of this is that, in the spin correlation experiments, Bell’s inequality is violated. In other words, Bell’s inequality does not hold for quantum mechanical systems. This was an extremely important result [140].

As might be imagined, the profundity of this result has come under much scrutiny. One argument is that, since the two detectors are set at one configuration of angles and then data is collected for, say, a billion electrons, there is time for the detectors to “know” each other’s orientation, although not by any known mechanism. But tests have been conducted where the angles of the filters in the detectors are set randomly after the electrons have left the source. The results are the same: Bell’s inequality is violated and the predicted quantum correlations confirmed [140]. Other critiques point to the fact that the correlated pairs emitted go in all directions, so only a very small fraction of them are actually measured by the detectors. But experiments using Beryllium atoms have measured almost all of the pairs. The results again confirmed the quantum correlations [140]. There are other objections, but suffice it to say that most physicists believe the quantum correlations have been sufficiently confirmed.

### 4.4.4 Bell’s Theorem

So far the discussion has been limited to Bell’s inequality. Where does Bell’s Theorem come in? Bell’s Theorem is proof by contradiction, which involves making certain assumptions about the thing to be proved and then showing that the assumptions lead to a contradiction, and hence, at least one of the assumptions must be false. Recall
that Bell's inequality is based on two assumptions: 1) classical propositional logic is a valid way of reasoning about a specific problem; 2) an object has a definite characteristic, whether or not it is actually observed or measured.

Bell's theorem applied to the electron spin experiments actually makes three assumptions [140]:

1. Classical propositional logic is a valid way of reasoning
2. Electrons have spin in a given direction whether measured or not
3. No information can travel faster than light

The third assumption has been added as a result of the EPR paradox and is often referred to as the ‘locality’ assumption. In fact, if assumptions two and three are true, it would mean that the uncertainty principle is beaten, since 1) nothing travels faster than light and 2) the spin orientation of electron pairs can be known with certainty. In other words, the exact values of conjugate variables could be known simultaneously, which is in direct violation with the uncertainty principle of the quantum mechanics.

Sometimes the three assumptions are stated [140]:

1. Logic is valid
2. There is a reality separate from its observation
3. Locality

The key insight into the electron spin experiments is that Bell's inequality is violated. This means that at least one of the above assumptions is false. This is a valid conclusion reached through proof by negation. Harrison, in his treatment of the subject, goes on to discuss the consequences if each of the assumptions individually turned out to be false [140]. He appears to treat each assumption as if it is independent of the others.

At this point, I should say that my perspective on Bell’s Theorem differs from Harrison’s. Bell’s inequality is an exercise in classical propositional logic. And classical propositional logic is based on the idea of an ‘atomic proposition’, which is fundamental to the subject. An atomic proposition can be described as a statement that “can be the case or not be the case and everything else remain the same” [133]. To be atomic, a proposition is assigned a truth value (true or false) without affecting the truth value of any other atomic proposition. In this sense, the idea of an ‘atomic proposition’ is local by definition. And it seems reasonable that the statements that make up the assumptions associated with Bell's inequality should be atomic propositions, since they are the only meaningful statements within that system of logic. For example, \( P \rightarrow Q \) (\( P \) implies \( Q \)) is the definition of a compound proposition called an ‘implication’, where \( P \) and \( Q \) are atomic propositions. If \( P \) is true and \( Q \) is true, then \( P \rightarrow Q \) is true. But, if \( P \) is true and \( Q \) is false, then \( P \rightarrow Q \) is false. The point here is that before a truth value can be assigned to the implication, it is necessary to assign a truth value to each individual
atomic proposition that comprises the implication. The same can be said of compound statements like the conjunction \( P \land Q \) or the disjunction \( P \lor Q \).

The characteristics in Harrison’s student example are certainly atomic. For instance, if it is false that a student is taller than 5’8”, there is no implication that he or she must have blue eyes. The same can be said for any combination of the statements associated with the “student example” characteristics. The truth value of a characteristic does not depend on the truth value of one of the other characteristics. But is this true of the assumptions associated with Bell’s inequality? It is very hard to see that it is.

Recall the assumptions associated with Bell’s inequality:

1. Logic is valid
2. There is a reality separate from its observation
3. Locality

Assuming the assumptions associated with Bell’s theorem are atomic propositions, suppose that the third assumption, the ‘locality’ assumption, is false. Could the first assumption ‘logic is valid’ be true simultaneously? In other words, could classical propositional logic be used in some sense to describe a non-local reality? It is very hard to understand how it could. To see this, suppose \( P \) represents the statement ‘the electron is ‘spin-up’ and \( Q \) represent the statement ‘the electron is not ‘spin-up’ (which means it is spin-down). Considered individually both \( P \) and \( Q \) are atomic propositions. Either could be true or false. But, because electron spin comes in entangled states, both statements \( P \) and \( Q \) could be true of a single electron at the same time. This is simply not allowed in classical propositional logic. In a sense, the statement ‘\( Q \)’ is the same as the statement ‘\( \sim P \)’. And the statement ‘\( P \land \sim P \)’ (\( P \) and \( \sim P \)) within classical logic is always false. In propositional logic no single object can have a characteristic and not have that same characteristic at the same time. In plain language, such statements are contradictions. In essence, the statement ‘\( P \) and \( Q \) are atomic propositions’ is the same thing as saying that \( P \) and \( Q \) are local – their truth values should not depend on the truth values of any other proposition. But if the ‘locality’ assumption is false, then \( P \)’s truth value determines \( Q \)’s truth value by necessity. This is in direct conflict with the very idea of an atomic proposition. It is very hard to see how the ‘locality’ assumption could be false while the ‘logic is valid’ assumption remains true, at least, as it applies to the electron spin experiments and by extension the quantum theory. The only possible conclusion is that if the ‘locality’ assumption is false, so is the ‘logic is valid’ assumption.

Care must be taken here, because many physicists regard the locality assumption as true, since it implies that no signal can travel faster than light – a contention accepted by virtually all physicists. For instance, if a coin is cut in half so that one half contains a head and the other half a tail and if one of halves of the coin is selected at random, then the statement: “this half contains a head” is certainly true or false. The only way of knowing if the half selected contains a head is by looking at it. Many physicists regard this as a local phenomenon, since, if one half of the coin contains a head or a tail, it
does not in any way change the fact that the other half of the coin could be a head or a tail. And since there is no actual direct communication between the two halves of the coin, the phenomenon is deemed local.

But the notion of locality given above is completely at odds with propositional logic. The statement “this half of the coin contains a head” is indeed true or false, and so, gives the appearance of being an atomic proposition, and therefore, a meaningful statement within propositional logic. But, in fact, it is not a valid proposition. To see this, suppose

\[ P \text{ = “this half of the coin contains a head”} \]
\[ Q \text{ = “the other half of the coin contains a head”} \]

Now consider the implication \( P \rightarrow Q \). If \( P \) is true, commonsense dictates that the implication \( P \rightarrow Q \) is true, since if \( P \) is true, then \( Q \) must be false and if \( P \) is false then \( Q \) must be true, and hence, the truth value of \( Q \) is determined by the truth value of \( P \) i.e. the truth value of \( P \) implies the truth value of \( Q \). But within propositional logic, if the antecedent of an implication is true and the consequence false, then the implication is false. But, if the antecedent is false and the consequence true then the implication is true. Hence, there is a direct contradiction between the notion of locality given above and propositional logic. Of course, physicists are free to define locality in any manner they choose. But this does not imply that their notion of locality is consistent with propositional logic. It is not. To be atomic, and hence, a meaningful statement within propositional logic, the truth value of \( P \) should not in any way determine the truth value of \( Q \). And since if \( P \) is true implies that \( Q \) is false, the statements ‘\( P \)’ and ‘\( Q \)’ can be in no way considered atomic and must be considered statements outside of propositional logic.

Classical propositional logic is largely based on the law of the excluded middle, which states that propositions are either true or false. There is no declarative statement within classical propositional logic that is sort of true or sort of false. If the assumption ‘there is a reality separate from its observation’ is false, can the assumption ‘logic is valid’ be true? Again, it is hard to see how it could. Consider the statements:

\[ P \text{ = ‘An electron has an exact position’} \]
\[ Q \text{ = ‘An electron has an exact momentum’} \]

If the assumption ‘there is a reality separate from its observation’ is false, as it applies to the quantum theory, then, before an observation is made, it would be impossible to assign either \( P \) or \( Q \) a truth value. Before the observation, the electron is supposedly in a state, where it does not have an exact position or an exact momentum. Only the probabilities associated with the two propositions can be given. Left in a logical no man’s land, only a system of logic not based on the law of the excluded middle might apply. Again, the only reasonable conclusion is that if the assumption ‘there is a reality separate from its observation’ is false, then so is the assumption ‘logic is valid’.
On the other hand, once an observation is made, if $P$ is true, then $Q$ is false and vice versa. At that point, truth values can be assigned, but then the same problem arises as in the previous example: assigning a truth value to one of the propositions determines the truth value of the other. Neither $P$ nor $Q$, in this case, is an atomic proposition in the sense of classical propositional logic. If either of the assumptions, two or three, associated with the electron spin experiments were false, it is difficult to see how the 'logic is valid' assumption could be true. The truth value of the 'logic is valid' assumption becomes dependent on the truth value of the other assumptions and this violates the very essence of an atomic proposition fundamental to classical propositional logic. Therefore, is the following statement true?

*All three assumptions associated with Bell’s Theorem are one and the same.*

Actually, not quite. It is certainly possible that the 'locality' assumption is true, while the assumption that 'logic is valid' is false. And it is certainly possible that ‘there is a reality separate from its observation’ assumption is true, while the assumption that ‘logic is valid’ is false. But, what can be said is if either the ‘there is a reality separate from its observation’ assumption or the ‘locality’ assumption is false, then the assumption that 'logic is valid' is false.

Bell's theorem has often been cited as being a resolution to the *EPR* paradox. The argument goes that since the concept of local realism, favored by Einstein, yields predictions that disagree with those of the quantum theory and because numerous experiments agree with those predictions, the concept of 'local realism' is refuted [141]. It is not hard to see why this is an often repeated interpretation of Bell’s Theorem. The whole objective of applying Bell’s Theorem to the quantum theory is to resolve issues regarding the nature of the physical world. If, by chance, Bell’s Theorem also has profound implications regarding the language (the logic) physics is written in, that's a bonus, but surely less noteworthy than the "nature of the world" implications.

Considering the arguments just made, it would seem the only possible conclusion is that the quantum theory cannot be described in term of classical logic or any system derivable from it. Of course, this begs the question: what system of logic, if any, describes quantum processes? Book IV will discuss in detail how a change in logical approach will impact the nature of the problem of unifying all of physics.
Chapter 5

The Philosophical Challenges of Unification

The time has come to realise that an interpretation of the universe—even a positivist one—remains unsatisfying unless it covers the interior as well as the exterior of things; mind as well as matter. The true physics is that which will, one day, achieve the inclusion of man in his wholeness in a coherent picture of the world.

— Pierre Teilhard de Chardin,
The Phenomenon of Man

5.0 Introduction

The limitations on laboratory experimentation have pushed the practitioners of theoretical physics from theorizing about things accessible to experimental conformation to prognostications that reach beyond, to the edges of speculative philosophy. There is the problem of dark matter and energy and the issue of force carrying bosons. Neither of which is directly observable. Quantum theories depend on the notion of a wave function, which admits to no basis in reality.

Physics has never been able to completely extract itself from metaphysics, although not for lack of trying. But, which metaphysical notions should be included and which excluded from physics? Which metaphysical ideas are important in formulating a solid physical theory and which are unimportant and why? These issues, along with the discussion of the philosophical challenges of finding a unifying theory for physics, are the topics of this chapter.

5.1 Two Philosophical Questions

Two unresolved philosophical questions play a role in preventing physics from being explained by a single theory. The first is often called the 'mind-body problem'. Is the mind independent of the existence of the body or are the mind (if such a thing exists) and the body undivided? Physicists often state the problem in the following terms: what is consciousness and what role should it play in physics?

The second unresolved question: does an external world exist independently of our perception of it? This question reflects the antithesis in the realist/idealist worldview. Realists answer the question in the affirmative, idealists in the negative.

To philosophers and I should think to most physicists too, it is important to know whether observations and the measurements extracted from those observations reflect the true nature of the object under investigation. Interestingly, some physicists take an agnostic view toward this question – claiming that theories should only predict the correct outcome of a measurement. Anything beyond that is not physics.
Positivism asserts that the only authentic knowledge of the world is based on sense experience and positive verification. The Logical Positivists of the early twentieth century concluded that pure abstractions could tell us nothing about the external world. According to this view, in the world, there are only facts and those facts are arranged into certain relationships called ‘states of affairs’. The facts and states of affairs make the world what it is. Those facts and states of affairs are facts and states of affairs of a particular kind. To the positivist, a man could point to an animal and say ‘cat’. This was to assign a sound (symbol) to a particular object in space and time. But the symbol represented a particular object and a particular state of affairs. If the man thought he had formed the abstract notion of a ‘cat’, for example, he was mistaken.

Of course, people form abstract notions all the time or at least think they do. If an ordinary citizen was asked: “what is a cat”, the reaction would likely be one of shock that such the question was even broached. But a positivist would point out that even though a man could point to dozens of objects and say “this is a cat”, he could not, with certainty, recognize the next cat if he saw it. His abstract idea of a ‘cat’ was just that, purely an idea. There would be no physical object that exactly matched the man’s idea of a ‘cat’. Cats are particular objects characterized by a particular set of sense data. They cannot be further analyzed beyond sense experience. The man was simply associating a particular array of sense data with the word ‘cat’ that represented what he saw with his eyes etc. The positivist, while admitting the existence of an external world, would deny there is a one-to-one correspondence between our abstract notions or ideas of an external object and the object itself.

The positivist view has undergone several modifications in an attempt to clarify what “positive verification” actually means. I think it is safe to say the positivist tenant of direct “positive verification” is today not strictly adhered to. Its influence, however, carries considerable weight within the physics community.

What does modern physics say about the two questions just posed? It has a lot to say about the “is there an external world?” question. The answer modern physics gives turns out to be one of the most controversial of current theoretical thought. The theory of relativity answers the question in the affirmative while the quantum theory gives a somewhat qualified negative answer. I think most quantum theorists would accept the notion that an external world exists, but would add that the nature of that external world is modified by observation. And, some would argue, modified by consciousness. In this view, there is not really an independent external world separate from our observation of it or, even if there is, nothing really definitive can be said about it. Any attempt at penetrating the world of the wave function to a deeper level of reality beyond what quantum mechanics allows has generally failed.

Unlike relativity that assigns a reality to external objects, whether they are observed or not, quantum mechanics replaces local realism with a kind of positivism. The quantum theory regards the unobserved world as basically unreachable accept to give probabilities to outcomes. It can say nothing about the realism of the world until an observation is made. How these two theories can be reconciled into a sensible theory
that resolves the dichotomies in each of their philosophical positions is one of the most perplexing mysteries in theoretical physics.

As to the first question, at least in its present form, current theoretical thought in physics appears to have little to say. This is not to imply that physicists have ignored the question. Amongst physicists there does not appear to be one approach or angle of attacking the question that has gained any real momentum. The idea of a conscious observer plays a significant role in the measurement problem for those physicists who maintain that, in some sense, consciousness collapses the wave function. But to say that a conclusive position has been reached on this issue, one that is acceptable to the majority of physicists is, I think, inaccurate. The philosophers of the past, however, had plenty to say about the first question and the second too. And it will be instructive to examine what they had to say, if, for no other reason, but to gain a better understanding of the two issues from the varied contemplations of some of the finest minds of the past centuries.

5.1.1 René Descartes (1596 – 1650)

Descartes regarded knowledge gained through sense experience as unreliable. Only through thinking thoughts completely divorced from the influence of sense experience could man arrive at a true understanding of anything. To Descartes, the mind was a powerful independently acting agent through which true knowledge could be gained. He believed that the body worked like a machine, had the material properties of extension and motion reflected in the laws of physics. The mind, on the other hand, was non-material, lacked extension and motion, and was not subject to the laws of physics.

Descartes believed that the mind was distinct from the body. Here, Descartes means a distinction between two or more substances. A real distinction is perceived when one substance can be clearly and distinctly understood without knowing anything about the other and vice versa. In arguing for a real distinction between mind and body, Descartes claimed that 1) the mind is a substance, 2) it can be clearly and distinctly understood without any other substance, including bodies, and 3) that ultimately minds or souls exist without bodies [142].

Descartes elaborates on the differences between material and non-material substance by claiming that the nature of body or extension was divisible into parts, while the nature of the mind is “something quite simple and complete”, in a word, indivisible. The mind and body cannot have the same nature, for if this were true, then the same thing would be both divisible and indivisible, which, to Descartes, was impossible. Hence, mind and body must have two completely different natures in order for each to be understood by itself without knowing anything about the other [142].

The origins of the mind-body problem can be traced to the meditations of Descartes and his conclusion that mind and body are really distinct substances. But Descartes ran into a difficulty. If the mind is an entirely non-material thing without any extension in it
whatsoever and the body is an entirely material thing without any thinking in it at all, then each substance can have only its particular kind of modes. The mind can only have modes of understanding, will and, in some sense, sensation, while the body can only have modes of size, shape, motion, and quantity. But if that is the case, why can the “will” move the body, for example, when a question in a student’s mind causes the raising of the arm, and certain motions in the body cause the mind to have sensations? In other words, how can two substances with completely different natures causally interact? Descartes could not provide a convincing explanation [143].

The dualism (read separation) between the mind and the body was distinctive of Descartes’ philosophy. He thought that the only path to a true and complete understanding of nature was through deductive reasoning. And this was only to be achieved through a purely mental discipline free from the deceptions of sense experience. Later, using the deductive method he endorsed, Descartes gave an ontological proof of the existence of God. Still later, the proof was shown to be embodied in the premise of the argument. Descartes argument was circular. But once Descartes proved the existence of God, he could have faith that God would not deceive him and through the power and understanding God gave him, Descartes believed he could gain an understanding of the true nature of the external world [144].

Descartes also had a response to the skepticism concerning the existence of an external world. He argued that sensory perceptions are not willed by him. They are external to his own senses, and, according to Descartes, this is evidence of the existence of something outside of his mind, and thus, an external world. Since the mind was non-material, it was able to make a distinction between itself and substances different from itself and this was sufficient evidence that he could not be deceived that the external world contained material things and was totally different from his mind. He also maintained that the cause of the ideas in his mind were the bodies that existed in the external world [144].

Descartes was a rationalist and a realist. He believed in the independence of mind and body and in the independent existence of an external world. But he failed to show that the only true path to a sensible theory of nature was gained through deductive reasoning. Later, theoretical thinkers more inclined to modern scientific philosophy (positivism) insisted that metaphysical statements of the kind that Descartes advocated should be accompanied by verification through experiment and observation.

5.1.2 Benedict DeSpinoza (1632 – 1677)

Spinoza thought that Descartes’ had erred in believing that the mind and the body were made of entirely different substances. He believed that the mind and the body acted as a single unit, but were in some sense independent of one another. Spinoza is considered a monist, but his monism is unique. Usually monistic theories purport that all existence is made of the same kind of ‘stuff’, material or non-material whatever the case may be. But Spinoza’s system is different and he goes to great lengths to explain how the mind and body could be considered a single unit, yet maintain a fundamentally
independent relationship. It will be beneficial to discuss in some detail the nuances of his system.

Reading Spinoza’s monumental work ‘The Ethics’, not far past the first few words gains insight into how he planned to treat the mind-body problem. Part One, Definition II of “The Ethics” reads [145]:

“That thing is called finite in its own kind (in suo genere) which can be limited by another thing of the same nature. For example, a body is called finite because we always conceive another which is greater. So a thought is limited by another thought; but a body is not limited by a thought, nor a thought by a body.”

Spinoza reckoned thoughts amongst things non-material and bodies amongst material things. He believed in an independent external world. In this, he agreed with Descartes: “The Ethics”, Book II, axiom II:

“Modes of thought, such as love, desire, or the emotions of the mind, by whatever name they may be called, do not exist unless in the same individual exists the idea of the thing loved, desired etc…..”

Here Spinoza appears to say that an idea must be associated with the object of that idea. Presumably, these objects, whether they were real or imagined, are things perceived as external to the mind. They are things extended in space and time, in other words, material objects. It is interesting that there does not seem to be any room in Spinoza’s system for statements like “I desire to be happy”. Happiness in Spinoza’s system is not a state of mind, but arises from the mind’s perception of the disposition of the human body. If the body’s power of acting, and by power of acting Spinoza means power of existing, is perceived by the mind as enhanced, the mind will rejoice. If, on the other hand, the mind perceives that the body’s power of acting is hindered in any way, the mind will suffer. Spinoza makes this mind-body connection very clear in Proposition XIII:

“The object of the idea constituting the human mind is a body, or a certain mode of extension actually existing, and nothing else.”

Spinoza believed in the existence of a single substance. This substance he called ‘God’. But Spinoza’s God was not a God who stood on the outside looking in as in some religious systems – Christianity comes to mind. His “substance” was reflected in the essence of all existing things. Whether it was substance thinking or substance extended, it was the same substance. His stance on this question was diametrically opposed to Descartes’, where extended (material) things were subject to the laws of physics, but thoughts (the non-material) were not. While Spinoza agreed with Descartes that thoughts were non-material and extended things were material, he believed that both extended things and non-material thoughts were governed by the same set of laws. This he made clear in Proposition VII, of Book II of “The Ethics”:
“The order and connection of ideas is the same as the order and connection of things.”

Apparently, this means that if I could write down a set of mathematical equations, the laws of physics so to speak, that explained the entire workings of the material world, those same equations would also explain the laws of the non-materiality of the mind, even though the material (things) and the non-material (mind), are to be understood as completely separate.

This idea is further expressed in proposition XX, part II:

“There exists in god the idea or knowledge of the human mind, which follows in Him and is related to Him in the same way as the idea or knowledge of the human body.”

This position is diametrically opposed to the positivist view in which abstract concepts such as “I can think of a cat” are generally considered meaningless. A positivist might concede that God, if such a being actually exists, might be capable of thinking abstract thoughts of the kind Spinoza describes, but humans certainly could not. But as will be seen shortly, Spinoza believed man was also capable of forming abstract ideas, if only in a somewhat indirect fashion.

So, how did Spinoza foresee the connection between the mind and the body and what did he mean by the mind and body functioning as a single unit? He addresses the first of these questions in proposition XXIII, part II:

“The mind does not know itself except in so far as it perceives the ideas of the modifications of the human body.”

It is interesting to note that, in Spinoza’s system, the human mind cannot know anything except through the disposition of the human body. So how it is possible to form an adequate concept of the nature the external world? His views on this are reflected in proposition XXV, part II:

“The idea of each modification of the human body does not involve an adequate knowledge of an external body.”

And again in Proposition XXVI:

“The human mind perceives no external body as actually existing unless through the ideas of the modifications of its own body.”

Here Spinoza addresses one of the most interesting and key aspects of the realist worldview – that there exists an external world independent of our perception of it. Any realist must accept as a possibility that the perceptions of the mind presumably caused by the presence of an external object do not necessarily reflect the nature of that object.
Spinoza concedes that all bodies have some things in common, and to the extent of that commonality, their natures agree. But he also concludes that the natures of different bodies disagree to some extent, so the knowledge of the nature of an external object cannot be gained through a complete knowledge of the modifications of the human body which, in his system, is the only thing that the human mind can perceive.

It remains to explain what Spinoza’s position was regarding man’s ability to gain a complete understanding of the nature of the external world and to explain how he thought this knowledge was acquired. Here Spinoza gives a somewhat unconvincing and difficult to follow argument. In proposition XXXVIII, part II, he declares:

“Those things which are common to everything, and which are equally in the part and in the whole, can only be adequately conceived.”

Here Spinoza gives hope that finding those things that, in some sense, can be conceived “equality in the part and in the whole” (whatever this means), a true idea of substance can be gained. It should be clear by now that, in Spinoza’s system, a true idea is an idea of something that agrees in total with the actual nature of the thing being considered. So, the something ‘𝑋’ conceived through thought and the something ‘𝑋’ external to the mind must have exactly the same structure. This reflects Spinoza’s conception of a substance. Apparently, the knowledge gained is not the nature of the external objects that exist in the world, but the general nature of substance i.e. God. Proposition XXXIX, Book II:

“There will exist in the human mind an adequate idea of that which is common and proper to the human body, and to any external bodies by which the human body is generally affected – of that which equally in the part of each of these external bodies and in the whole is common and proper.”

Even though Spinoza’s true (adequate) idea is an idea of no individual thing, the human mind is capable of forming an abstract (general) idea of the substance reflected in the nature of what is common and proper to the human body which, in turn, is reflected as common to all external bodies, because, as with everything in Spinoza’s system, all things are derived from a single substance. This knowledge Spinoza called ‘scientific intuition’. It was characterized by the ability to grasp the entire nature of substance all at once – in a flash of purely luminous insight.

By whatever name they called it, the rational philosophers, to one extent or another, believed that man was endowed with a higher ability that surpassed all the other inhabitants of the Earth. This ability is normally called ‘reason’. Brand Blanchard, the Yale rationalist, describes reason this way [146]:

“When we say that man is a rational animal, then, we seem to imply that he can command ideas independently of sense, that he can abstract, that he can infer explicitly, and that he can sit in judgment on himself. The highest of animals can
do none of these things. The stupidest man, if not a pathological case, can in some measure do them all.”

5.1.3 David Hume (1711 – 1776)

No sooner had the ink dried on the pages of the rationalist’s tablets, their optimism for man’s future, their confident insights and sure judgments were rudely and skillfully challenged by the skepticism of their empirically leaning counterparts. While the empirical philosophers did not always agree with one another on various topics, the one thing they could agree on was that the rationalists had erred. Man possessed no such ability of abstract thinking of the kind the rationalists claimed, no full proof general approach could be found that would lead man to the truth.

David Hume was particularly adamant that the only power the mind possessed was the power of associations. It was not reason but custom and the repetition of associating one impression with another that determined our beliefs. Those beliefs then could only be based on probabilities. Our beliefs were only likely. The exact truth or falsity of any idea was uncertain. Hume classifies the perceptions of the mind into two distinct kinds: impressions (sensations, passions and emotions) and ideas. But, he regards ideas as simply faint impressions. All impressions are sense impressions. How he conceives ideas and impression as two distinct kinds, when the difference is one of degree, is not clear. But, Hume’s philosophy became the forerunner of scientific philosophy, where observation and experience play fundamental roles and metaphysics (meta: meaning beyond), any idea which could not be linked to a particular impression, is regarded as nonsense. If Spinoza thought he could come to know a single substance through some obscure mental abstracting ability, he was mistaken. There was clearly nothing but particular substances reflected in the impressions of sense [147].

Hume’s philosophy greatly influenced the twentieth century logical positivists, who maintained that statements that could not be verified by sense or were not logical tautologies (truth statements) or contradictions (logically false statements) were either emotive statements (statements that do not express propositions (true or false), but emotional attitudes), or the statements were simply meaningless.

5.1.4 John Locke (1632 - 1704)

But the first skeptical blow was struck by John Locke. By the time Locke wrote his “An Essay Concerning Human Understanding”, Newton had long ago finished his monumental work. Locke was profoundly influenced by Newton. He took it as established beyond all doubt that the world was mechanical, acting like a giant machine having a fundamentally corpuscular (atomic) nature – made from a conglomeration of insensible particles acting on one another by impulse, just as Newton had explained [148].

Like the rationalists, he accepted the notion of the independent existence of an external world. Perceptions, he thought, were a result of atoms acting mechanically on the
sense organs which sent impulses to the brain. But at the moment the impulses reached the brain, the material interactions produced a non-material idea. And it was those non-material ideas of which human awareness consisted [148].

This is where Locke parted ways with the rationalists. Although he endorsed the idea of an independently existing external world, he maintained that man was irreversibly cut off from it. Direct awareness was given to only our own ideas, not the objects existing in the external world.

But were the ideas held in our mind faithful representations of the external objects that are taken to be their cause? Locke maintained that some of the qualities associated with external objects were primary, some secondary. The primary qualities (extension, moving, having mass/weight) really are “in” external objects, but owing to the nature of our sensibility, secondary qualities (color, odor etc.) are merely appearances, figments of the imagination that are not really “in” extended objects [148].

Upon this point, his critics pounced. What criteria had Locke used to distinguish primary from secondary qualities? By Locke’s own admission what could be perceived directly were the ideas in our heads, not the external objects that supposedly caused those ideas. In Locke’s system, non-material ideas were the only things that man could be aware of. What grounds did he have for supposing that there was something included in those ideas that made some sensible qualities real while others were imaginary? Locke really never gave a satisfactory retort to this objection [148].

Locke’s position on the question of substance was much like that of Descartes. He firmly believed in material as well as non-material substance. There must be, thought Locke, something to which the qualities of objects belonged. For it cannot intelligently be supposed that the qualities of objects exist in their own right. There must be something for them to be qualities of. But this is where the similarities between Locke’s view and Descartes’ ended. In Locke’s mind there was no way of saying what substance was. For, to anything that pertained to substance was to unavoidably assign a quality to it. And that got no closer to the truth of what substance entailed [148]. This position reflected the thoughts and words that Locke will forever be remembered: substance is “something, I know not what”.

His position on substance pushed Locke into further difficulties. He wished to claim that there were two kinds of substance: material (extended world) and non-material (ideas). But if substance was “something, I know not what”, on what grounds would Locke be entitled to claim that there were two? In addition, from a theory of knowledge standpoint, Locke’s forecast portends a dreary outcome. Of what, then, could man know through contemplating his own ideas? If all that man can perceive is his own ideas and those ideas, as Locke admitted, were not faithful representations of an external world, neither were they faithful reflections of an underlying substance, then what were they? How could anything be known at all? On this question, Locke could provide no reply [148].

129
5.1.5 George Berkeley (1685 – 1753)

Berkeley found Locke’s system not only distasteful, but ridiculous and an affront to common sense. Berkeley was devoutly religious and in Locke’s system he saw an intolerable pretext for religious skepticism. Locke’s system was founded on the Newtonian idea of a physical universe operating mechanically like a vast machine. But where in this system did God come in? Locke had assumed that the universe was put in to existence and got going by a supreme deity. But Berkeley rightly realized that this assumption could be easily challenged. What if the universe was eternal? If the universe was not created, there would be no room for a creator. In Berkeley’s mind, such a scenario was unthinkable [148].

Berkeley was keenly aware of the struggles Locke encountered trying to explain the differences between the primary and the secondary qualities of matter and his torturous grappling with the idea of two distinct substances, the existence of which, Locke had little right to claim [148].

Berkeley found it patently absurd that the “visible beauty of creation” is no more than a “false imaginary glare” that the flowers in our gardens really have no color and no scent. He realized early on that Locke really had no basis on which to claim that the qualities of matter differed in any way. Locke’s insistence on the existence of an external world of which nothing could be known directly left him no grounds for asserting the existence of the primary and secondary qualities of matter. Upon what would this assertion be based? Moreover, Locke’s tenant that the only perceivable things are our own ideas, yet supposing those ideas reflect qualities and are connected to an external “something, I know not what”, in Berkeley’s view, was outrageous, bordering on blatant insanity [148].

So what was Berkeley’s solution? It was simply to deny the existence of matter. Berkeley believed he could wipe out all of Locke’s tortuous struggles in one blow. It removed at once Locke’s bumbling dualistic treatment of substance. Material substance simply did not exist, so the difficulties associated with Locke’s groundless insistence that two different types of substance existed vanished. There existed only a non-material spiritual substance and no other. All Locke’s excruciatingly painful grappling with the absurd hypothesis of the existence of an inaccessible and unknowable external world were wiped away. It simply did not exist. The houses, the trees and the birds would be associated directly with our ideas and did really exist. Even better, the need for contrasting our ideas with a baffling array of inaccessible and inexplicable causes disappeared [148].

Still, it could be objected that Berkeley’s explanation left no account for the existence of our ideas. If it wasn’t external objects that caused our ideas, where did they come from? Berkeley explanation: a truly incorporeal, active substance or spirit whose force of will implanted ideas in our minds in an orderly, systematic and rational manner [148].
It could further be objected that Berkeley’s system left no room for the existence of evil. If God was the cause of all our ideas and God was good, why then was there so much evil in the world – a condition with which Berkeley himself was incessantly preoccupied? Berkeley’s reply was that God simply provided the orderly connection and arrangement of our world of ideas as experienced. Our relationship with that orderly world of ideas was left up to us [148].

Finally, doesn’t this system require that ideas exist only when those ideas are perceived? Yes, Berkeley admitted, but he saw no difficulty in this. This insight more than any other is undoubtedly the one for which he will be most remembered. In this, Berkeley became prophetic. The scientists of the twentieth century found this insight remarkably useful in explaining the emerging field of quantum mechanics [148].

So what was Berkeley’s position on abstract ideas? Early on, he riled vehemently against them. Abstract ideas were, he thought, the primary reason that the philosophers of the enlightenment had gone wrong, failed to clear up old errors and introduced new ones, filled the public consciousness with all sorts of unnecessary notions of “something, I know not what” that not only was an affront to common sense, but laid thicker the learned dust of the middle ages. His distrust of abstract notions was one of the primary reasons he took on the task of challenging the ruinous and obscure ramblings of the philosophers of the day. Berkeley thought the answers to all these questions were quite simple and luminously clear [148].

But later, Berkeley realized he could not square his position on abstract ideas with his own philosophy. He wished to argue that the only things directly perceived were our own ideas and that those ideas were non-materi ally spiritual in nature. But what was the mechanism that allowed the conclusion that ideas were non-material? Certainly this perception was not gained through the ideas themselves, the content of which was concrete sense experience. This left no room for the concept of a deity. Berkeley argued, rather weakly, that in addition to perceptions, there are notions of spiritual things. He argued this without further elucidation on what he meant by a notion [148]. In fact, a notion seemed lot like an abstract idea. And if it wasn’t a perception or an abstract idea, then what was it? This, assuredly, is the weakest of Berkeley’s arguments and he never really adequately addressed the issue.

Finally, what account can Berkeley give of the physical sciences? The scientists of the day gave a rather convincing account of how the world acted like a finely tuned machine, a world of atoms acting impulsively on the sense organs providing an awareness of an intelligently inspired external world – an external world that in Berkeley’s view did not exist. Since a physical substance did not exist, there was nothing in Berkeley’s view for the so called “laws of nature” to be true of. Was the truth of an independently existing external world really a figment of the scientific imagination? Berkeley answered: yes, the scientific notion of an independently existing external world was a fiction, but it was a useful fiction. If the scientists found it necessary to invent cleverly devised abstractions to explain in general terms the nature of how our ideas
worked, so be it, Berkeley had no objection. But these abstractions were only useful, not true [148].

Berkeley’s insights predate those of some physicists today who are inclined to argue just as he did that the physics theories of today are not factual truth, but are merely mathematical and predictive conveniences. Not surprisingly, this view encounters stiff resistance. If it is not truth the physical scientist is after, then what is it? Berkeley deeply despised the high praise the physical scientists of the day were receiving and wished to imply that they had no exclusive claim to be an authority on the “nature of things”. He plainly failed to gain acceptance of this argument during his lifetime. And it is revealing that in our own day it has won more general support than ever before [148].

5.1.6 Immanuel Kant (1724–1804)

Kant wrote his Critique of Pure Reason toward the end of the Enlightenment, which by then was in a state of crisis. The Enlightenment had seen the spectacular achievement of Newton, enjoyed a clear break from the oppressive church dominated darkness of the Middle Ages and uncovered a new power embodied in human reason. This new confidence in reason brought with it the hope that humans could break free from traditional political or religious authorities and dart into a new era of self-government guided by reason. For why should political or religious authorities tell us how to live or what to believe, if each of us has the capacity to figure these things out for ourselves? Kant was truly committed to preserving this rationalist dream. To Kant, reason embodied the true path to freedom [149].

One of the inspirations of the Enlightenment was the new physics. But this new physics was mechanistic – a sea of impulsively guided atoms and bodies governed by causal laws. Where did freedom, a soul, or anything but matter in motion come in? Freedom was a necessary ingredient for choosing what is right over what is wrong, because otherwise who can be held accountable. And where was there room for the traditional religious belief in a soul that can survive death or be resurrected in an afterlife? Modern science, the source of so much optimism, threatened to undermine the traditional moral and religious beliefs that free rational thought was expected to support. This was the main intellectual crisis of the Enlightenment [149].

The Critique of Pure Reason was Kant's response to this crisis. The authority of reason was in question and Kant's main goal was to show through a critique of reason by reason that all essential human interests were mutually consistent. He attempted to show that reason really does deserve the sovereignty attributed to it by the Enlightenment [149].

Kant benefited from an understanding of the struggles his predecessor’s had encountered in attempting to answer the fundamental questions of the time and the associated disputes that naturally arose from those attempts. Against the empiricists, particularly Berkeley, he argued that if, indeed, all our perceptions were our own ideas, a contention upon which both Locke and Berkeley agreed, how could an understanding
of self be accomplished? Here Kant attempts to reestablish the existence of an independent external world – a claim Berkeley flatly rejected. Kant argued that if the only objects of our awareness are our own ideas, it would then be impossible to distinguish the ideas pertaining to my own existence from the ideas pertaining to objects, which evidently, have nothing to do with me. Kant firmly maintained that the only way I could distinguish “myself” from things “not myself” is for there to be an external world that would allow me to compare “myself” from things “not myself”. Kant’s methodological innovation was to employ what he called a ‘transcendental argument’. Kant argues that “there are objects that exist in space and time outside of me, which is a necessary condition for establishing one’s own existence. It would not be possible to be aware of myself as existing, he says, without presupposing the existence of something permanent outside of me to distinguish ‘myself’ from. Since I am aware of myself as existing, there is something permanent outside of me [149].

He charged the empiricists with a type of hypocrisy. If Berkeley could ontologically state that matter did not exist, then that claim, according to Kant, necessarily required the exercising of judgment and judgment did not come from sense experience alone. Judgment, thought Kant, was mostly a rational endeavor. Man could judge because he was rational. The empiricists, Kant argued, had committed the ultimate logical sin. They had used reason, in the sense that the rationalists had employed the word, to argue that man had no such reasoning ability [149].

Against the rationalists Kant argues that a-priori ideas like “God is a perfect being” are not possible. Contrary to what the rationalists thought, Kant rejects completely the notion that propositions like this one are etched on the fabric of the mind. Statements like “God is a perfect being” are a-priori statements about things as they are in themselves. In this case the statement purports to tell us something about God as he is in himself. Such statements, Kant maintains, are not possible because the rational judgments made by the mind can only apply to the objects existing in the external world and those objects are revealed in sense experience. The mind is devoid of content until interaction with the world actuates these formal constraints. The mind possesses a priori templates for judgments, not a priori judgments completely apart from sense experience [149].

Kant held that knowing the objects in the external world as they are in themselves was not possible. Because man was rational, he necessarily interjected upon those external objects a certain order and connection that were not necessarily inherent in the objects themselves. In essence, reason allowed us to create an orderly structure out of sense experiences [149].

Kant claims that the mind of the knower makes an active contribution to how external objects are experienced. This aspect of Kant system was fundamental to his transcendental idealism. His arguments are aimed at limiting knowledge. The rationalists believed that metaphysical knowledge about God, souls, substance etc. was real. But Kant argues that knowledge cannot go beyond the realm of sense experience [149]. Kant’s arguments are not unlike those found in quantum mechanics i.e. that the
external world is really unknowable until an observation is made. But the quantum mechanism's normally accept the idea that the observation is real and not “cooked” by the rationality of the human mind as Kant suggested.

Kant argues that objects cannot be experienced without being represented spatially. It is impossible to grasp an object as an object unless there is an ability to delineate the region of space it occupies. Without a spatial representation, sensations are undifferentiated and not ascribable to a particular object. Time is also a necessary intuition. The idea of time cannot be gathered from experience because succession and simultaneity of objects, both things that would indicate the passage of time, are impossible to represent if the capacity to represent objects in time was not an inherent human quality. In other words it is impossible to have any experience of objects that are not in time and space. Space and time themselves are templates and cannot be perceived directly. Any consciousness like ours must apprehend objects as occupying a region of space and persisting for some duration of time. Kant argues that the understanding must provide the concepts, the rules for identifying what is common or universal. By applying concepts, the understanding takes the particulars that are given in sensation and identifies what is common and general about them. The understanding allows thinking in abstract terms to form general notions of houses, cats and trees [149]. Here Kant reestablishes the ability of man to form abstract or general ideas, something empiricists particularly Berkeley and Hume flatly rejected.

Kant enumerated his special set of a priori contributions to experience in a table of categories. His list of a priori concepts made all empirical judgments possible. These concepts cannot be experienced directly; they are the manifestations which particular judgments of objects take [149].

What did Kant say specifically about the two fundamental questions that are the primary topic of this discussion: does an external world exist independently of our perception of it? And is there an active mind independent of the body?

It is clear that to both of these questions, Kant answered in the affirmative. An external world existed because I could distinguish myself from it. The mind existed because it possessed its own attributes manifested through a priori concepts (embodied in the categories) that were in some sense independent of pure sense experience.

But to the aspirations of his rationalist predecessors, who believed that reason could provide a faithful knowledge and representation of existence itself, Kant levied a severe limitation. In Kant’s view, the answers to purely metaphysical questions such as “does God exist?” were beyond reach. Reason could only be applied to the domain of sense experience. Moreover, a faithful knowledge of an external world as it existed in itself was too out of reach. It was man and his rational mind that imposed a particular order and connection on the things in the external world. Man could make sense of the external world because his rational mind impinged a certain order and connection on things, not because the things in the external world actually possessed the order and connection they appeared to have. In this, Kant sympathized with the “good Berkeley”.

134
The real world really was not the huge impersonal machine described by Newtonian physics. It was man’s a priori template of space and time that projected a mechanistic order and connection into the external world.

Kant’s system expressed four possible types of judgments. There were “a priori” judgements which were made independently of sense experience and “a post-priori” judgements that were not. Those two categories were further broken down into analytical and synthetic judgements. Analytical judgements were those which did not supply any new information about its subject. “That tree has a trunk” is an example of an “a post-priori” analytical judgement. Presumably the knowledge of “that tree” could only come by experiencing it. But, having a trunk is a necessary condition to being a tree. When no additional information is provided about the subject, then the judgement is analytic. But, “that tree is 50 feet tall” is an example of an “a post-priori” synthetic judgement. It is not necessarily in the nature of a tree to be 50 feet tall. Since being 50 feet tall is new information not inherent in the subject, the judgement is synthetic [146].

The foundation of Kant’s system rested in large part on there being “a prior” synthetic judgments. Kant believed that this type of synthetic judgement was manifested in mathematics and geometry. ‘5 + 7 = 12’ was an example of an “a priori” synthetic judgement, thought Kant. In order to produce 12 (evidently in this case 12 is the subject), knowledge of addition is required. And knowing how to add was not part of being the number ‘12’. It was something new not contained in the definition of 12. By that argument, the judgement was synthetic.

Kant felt that the only way man could rationally project objects in space and time, and hence, give order to the objects in the external world was the capacity to make “a priori” synthetic judgments. Those judgments revealed themselves in Euclidean geometry applied to Newtonian physics. In this, Kant had the upmost confidence. Euclidean geometry was etched into the fabric of the human mind as certainly as 2 + 2 = 4. It was Euclidean geometry that allowed for making sense of the external world [146].

It was upon these last two ideas that Kant’s critics pounced, particularly the logical positivists. Kant believed that man’s reason gave order and connection to the physical world. Reason became the rose colored glasses through which man formulated ideas of how the world worked. The Universe was orderly not because that order was inherent in the external world, but because man himself had “cooked” the result. Man’s own reasoning abilities forced upon Nature an orderly structure whose foundations were found in Euclidean geometry. It wasn’t Nature that was the source of the orderly connections man saw, it was man’s own a priori synthetic judgments that was the source of Natural order and those a-priori judgments were embodied in Euclidean geometry. Kant was as much a believer in Newtonian Mechanics as was Locke. Newton had employed Euclidean geometry to describe the workings of the world and it never seemed to have failed him. To Kant, Euclidean geometry was the necessary connection between pure mathematics and physics. It was what allowed man to apply mathematics to the description of the physical world [146].
But the positivists of the early twentieth century were well versed in mathematics, and by now, non-Euclidean geometries had been invented and thoroughly researched. And Einstein had employed non-Euclidean geometry in his theory of gravity. This showed conclusively that Euclidean geometry was not the necessary prescription for describing the world that Kant thought it was. And hence, it showed irrefutably that he had badly erred. Not only had Kant erred about the nature of Euclidean geometry, concluded the positivists, but he was also wrong about the existence of “a-priori” synthetic judgements. There were no such judgements, the positivists decided. Statements like “5 + 7 = 12” that Kant thought synthetic were, according to the positivists, analytic in nature. In fact, the positivists came to the conclusion that all mathematical and logical statements were ‘a-priori’ analytic [146].

This was an important departure. It cut off man’s connection to the external world that Kant had worked so hard to reestablish against the suspicions of the empiricists. The Euclidean geometry and ‘a-priori’ synthetic judgements embodied in logic and mathematics that were supposed to make the world a rational and intelligible place vanished. Instead, logic and mathematics became self-contained tautologies – self-consistent systems of pure thought. Lost was any connection to the physical world. To the positivists, logic and mathematics became valid ways of reasoning, but those ways said nothing about the external world [146].

Like Kant but unlike Berkeley, the positivists believed in an external world independent of our perception of it, but rejected the idea of any connection between mind and body. Instead, logical and mathematical statements were purely rational with no connection to the external world. Statements like “this apple is red” were considered propositions which could be either true or false and the only way of establishing the truth or falsity of such a statement was to verify it through sense experience. While logic and mathematics could give a self-consistent way of using language, it could say nothing about the external world. Of course, the verification argument depended somewhat on whether or not statements like “this apple is red” were actually contextually meaningful – a debate that will not be examined here.

5.1.7 Monistic Idealism

In recent years, there have appeared on the scene several books that promote the idea that the spiritualism of Eastern Mysticism is manifested in the concepts of quantum mechanics. These works have the apparent goal of bridging the conceptual chasms that divide science from religion. Roughly speaking, these views can be grouped under the idea of ‘monistic idealism’. Monistic idealism holds that consciousness, not matter, is the ground of all being. It is monist because there is only one type of thing in the Universe and idealist because that one thing is consciousness. In a sense, the journey has come full circle, back to the original question: what is consciousness?

To a monistic idealist consciousness is both material and non-material – mind and matter. There is no mind without matter and no matter without mind. Monistic idealism conceptualizes the meshing of opposites. There is no male without female, no up
without down, no existence without non-existence. Consciousness then is the unification of the internal with the external. Evidently, the relationship between monistic idealism and the quantum mechanics comes in the form of its connection to the quantum mechanical idea of ‘wave-particle duality’ revealed through the concept that “consciousness causes the collapse of the wave function” interpretation of the quantum mechanics. As far as I can make out, according to monistic idealism, everything in the Universe is connected by a vast oneness [150,151] – a consciousness of everything until that consciousness becomes aware of something. At that point, a certain and specific outcome or event occurs. I become aware of that pencil in front of me. Apparently, the goal of the eastern mystic is to become unaware of particular things, and hence, to become aware of only consciousness itself. This appears to be the supreme spiritual epiphany.

As it pertains to the two philosophical questions, it should be clear that, according to monistic idealism, an external world depends on the existence of an internal world. There can be no separation of the two. The mind and the body are inseparable. Awareness of a body without a mind is impossible. Hence, what our consciousness is aware of is having a body – a certain event or outcome in this vast landscape of interconnected consciousness.

In general, physicists have looked upon monistic idealism with a great deal of skepticism, primarily because it portends to be an explanation of the measurement problem in quantum mechanics. And while there are many proposed solutions to the measurement problem, as yet no one solution has gained general acceptance. Monistic idealism is not very scientific in this sense: it claims that what exists is nothing but consciousness and yet does not explain the mechanism of how that consciousness arises. If the mechanism cannot be explained, on what grounds can it be said that consciousness is understood? To put it another way, how did conscious beings arise out of the big bang?

5.2 Conclusion

This concludes a rather lengthy discussion of the status, the history and the logical and philosophical implications of modern physics. What was learned? There are two great theories that describe the nature of our Universe – the theory of relativity and quantum mechanics. One of the theories describes the physics of the macro-universe; the other describes the physics of the micro-universe. Together, the two theories explain just about everything, but not quite.

The micro part of our Universe is described by the Standard Model of particle physics. In fact, it successfully describes three of the four fundamental interactions of Nature. But, it also suffers from a number of shortcomings. It requires 19 or so adjustable constants whose values represent characteristics of our Universe about which nothing is known. The Standard Model does not include gravity nor does it have an explanation for dark energy/matter. It lacks an accommodation for the observed predominance of matter over anti-matter. The current theory requires that neutrinos have zero mass
which does not jibe with recent experiments. In order that its predictions agree with observations, the model must be supplemented with apparently unnatural mechanisms, renormalization and the Higgs mechanism, for example, which have no explanation within the theoretical framework of the model. At given energy levels, the input parameters to the model must be unnaturally fine-tuned. Finally, the self-consistency of the theoretical part of the Model has yet to be mathematically demonstrated.

On the other hand, while the theory of relativity successfully describes the macro part of our Universe and does not suffer from obvious logical or mathematical foundational issues, it cannot explain the nature of our Universe at the micro level. It simply isn’t up to the job. To apply general relativity to the physics of the small requires that Einstein’s non-linear equations be quantized in some manner. But, any attempt to do so results in a non-renormalizable theory – a theory whose infinities cannot be discarded or subsumed into the parameters of the theory. This fact has driven most physicists to the safer harbingers of string theory.

Of most problems with which physicists deal, the dichotomy between the theory of relativity and the quantum theory does not create serious issues. Relativists can study the cosmos without worrying too much about quantum effects. And quantum physicists can study the micro-world without worrying about the effects of gravity. But there are physical phenomena, black holes, for example, that simply cannot be explained by employing one or the other of the two theories. Such explanations need both. It is these situations that the incompatibilities between the two theories become problematical.

All attempts at unifying the two theories, of finding a quantum theory of gravity, have, so far, failed. Chapter 4 uncovered what appears to be the root of the problem. The logical foundations of the two theories are vastly different. This raises the question: where to go from here?

There are those who would argue that the fundamental differences in the two theories will not be overcome – that the correct path is simply to accept this fact and relinquish the idea that a unified theory will never come to fruition. In this view, physics would consist of a plurality of theories depending on the particular phenomenon under consideration. Another approach is to hope that string theory can remove the logical inconsistencies without deviating too much from already established physics. But, which string theory will emerge? Will \textit{M}-theory give a “\textit{theory of everything}”, or will there be a landscape of string theories, where only one of many theories in a vast array of possible universes describes the physics of our Universe? Or, finally, is there a different approach – a road not yet travelled?

If you’ve read this far, you’ve probably surmised that this study will pursue the third alternative – a path not yet travelled. The challenge has been set. A unification of physics will require that the logical inconsistencies in current theories be somehow understood and overcome. The philosophical issues raised in Chapter 5 have to be addressed. This will be accomplished in Book IV. But, before all this can happen, it will
be necessary to become familiar with current logical and mathematical concepts and then to apply those concepts in order to understand, as far as possible, the foundations of current physics. The first of these tasks will be addressed in Book II, the second in Book III.