Theory construction in high-energy particle physics

Adam Koberinski, The University of Western Ontario

Supervisor: Myrvold, Wayne, The University of Western Ontario
A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Philosophy
© Adam Koberinski 2019

Follow this and additional works at: https://ir.lib.uwo.ca/etd

Part of the Philosophy of Science Commons

Recommended Citation
https://ir.lib.uwo.ca/etd/6518

This Dissertation/Thesis is brought to you for free and open access by Scholarship@Western. It has been accepted for inclusion in Electronic Thesis and Dissertation Repository by an authorized administrator of Scholarship@Western. For more information, please contact wlswadmin@uwo.ca.
Abstract

Science is a process, through which theoretical frameworks are developed, new phenomena defined and discovered, and properties of entities tested. The goal of this dissertation is to illustrate how high-energy physics exemplified the process of theory construction from the 1950s to 1970s, and the promising ways in which it can continue to do so today. The lessons learned from the case studies examined here can inform future physics, and may provide methodological clues as to the best way forward today. I examine the discovery of parity nonconservation in weak interactions, the emergence of Yang-Mills theories as the foundation of the standard model, and contemporary precision testing of quantum electrodynamics. In each of these cases, I examine the details of the physicists’ practice to draw conclusions regarding the epistemology behind successful episodes of theory construction. I reconstruct the methodology of each episode in order to find generalizable lessons to apply to contemporary issues at the frontiers of the search for a theory of quantum gravity.

In order to understand the many moving parts in each case study, I introduce a new terminology to distinguish the “parts” of a scientific discipline, inspired by the literature on scientific modelling. These terms—theoretical framework, dynamical model, phenomenological model, experiment, and mathematical tools—are meant to aid in investigating other quantitative scientific disciplines beyond high-energy physics. Ultimately, high-energy physics is at its best when various avenues of theoretical ideas are being pursued, spurring the development of new mathematical techniques to use as tools, and new ideas are quickly and vigorously tested experimentally. Proliferation of new ideas in response to theoretical developments is characteristic of the era of construction of the standard model, and is still ongoing in precision testing of quantum electrodynamics today.

**Keywords:** Theory construction, high-energy physics, epistemology of physics
Summary for Lay Audience

Science is a process, through which theoretical frameworks are developed, new phenomena defined and discovered, and properties of entities tested. The goal of this dissertation is to illustrate how high-energy physics exemplified the process of theory construction from the 1950s to 1970s, and the promising ways in which it can continue to do so today. The lessons learned from the case studies examined here can inform future physics, and may provide methodological clues as to the best way forward today. I examine three major episodes in the development of the standard model of particle physics. In each of these cases, I examine the details of the physicists practice to draw conclusions regarding the epistemology behind successful episodes of theory construction. I reconstruct the methodology of each episode in order to find generalizable lessons to apply to contemporary issues at the frontiers of the search for a theory of quantum gravity.

In order to understand the many moving parts in each case study, I introduce a new terminology to distinguish the parts of a scientific discipline, inspired by the literature on scientific modelling. These terms are meant to aid in investigating other quantitative scientific disciplines beyond high-energy physics. Ultimately, high-energy physics is at its best when various avenues of theoretical ideas are being pursued, spurring the development of new mathematical techniques to use as tools, and new ideas are quickly and vigorously tested experimentally. Proliferation of new ideas in response to theoretical developments is characteristic of the era of construction of the standard model, and is still ongoing in precision testing of quantum electrodynamics today.
Acknowledgements

I am grateful to all of my many mentors throughout the last four years. I would like to thank my advisor, Dr. Wayne Myrvold, for his regular feedback on earlier drafts of all chapters and helpful connections made in the world of academic philosophy. I would like to thank Dr. Doreen Fraser for initially sparking my interest in the philosophy of physics, and for her continued mentorship throughout my time as a student. I would like to thank Dr. Chris Smeenk for his leadership and for the academic opportunities he has provided me. All three of these professors have had a major impact on my thinking and development as an academic.

I would like to thank Dr. Melissa Jacquart, Dr. Lucas Dunlap, Dr. Jamie Shaw, and Dr. Marie Gueguen for fostering a great atmosphere of intellectual development and fun among the philosophy of science students at Western. The time I spent with all of you has made this journey so much more enjoyable. I want to thank Austin Horn and John Lehmann for helping me stay active away from the office. My grandparents, Don and Jo-Anne Haycock, have supported me throughout my post-secondary education. I am grateful to Darian Fournie for putting up with the drawn-out process of finalizing this dissertation, and my sister Amber DeBruin for keeping my positive. Finally, I am most grateful to my mother, April DeBruin, without whom I would not be the person I am today.

This dissertation is dedicated to the memory of my father, Arthur Koberinski Jr.
Contents

Abstract ii
Summary for Lay Audience iii
Acknowledgements iv
List of Figures vii
List of Tables viii

1 Introduction: Theory Construction in Physics 1

2 Experimental priority in the discovery of weak parity nonconservation 4
  2.1 Introduction ......................................................... 4
  2.2 Distinctions and background ....................................... 6
    2.2.1 Terminology .................................................. 6
    2.2.2 The state of high-energy physics in “the fifties” ................. 9
  2.3 A (brief) history of parity nonconservation ........................ 11
    2.3.1 Strangeness .................................................... 13
    2.3.2 How many V-particles? ....................................... 14
    2.3.3 Lee and Yang ............................................... 16
    2.3.4 Toward electroweak unification ................................ 18
  2.4 Analysis ............................................................ 19
    2.4.1 Lessons from theory construction in high-energy physics ........ 19
    2.4.2 Surprising benefits of the absence of a dynamical model ........ 23
    2.4.3 Lessons for modern theoretical physics ....................... 24
  2.5 Conclusions ....................................................... 28

3 Mathematical developments in the rise of Yang-Mills gauge theories 29
  3.1 Introduction ........................................................ 29
    3.1.1 Terminology .................................................... 30
  3.2 Yang-Mills theories in the 1960s .................................. 32
    3.2.1 Gauge freedom ............................................... 32
    3.2.2 Renormalization ............................................. 33
    3.2.3 Rejection of local field theories ............................. 36
  3.3 The reemergence of Yang-Mills theories ............................ 38
    3.3.1 Proof of renormalizability ................................. 38
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.3.2</td>
<td>Development of the renormalization group</td>
<td>44</td>
</tr>
<tr>
<td>3.3.3</td>
<td>Lattice quantum field theory</td>
<td>48</td>
</tr>
<tr>
<td>3.4</td>
<td>Analysis</td>
<td>49</td>
</tr>
<tr>
<td>3.4.1</td>
<td>Implications for theory construction</td>
<td>49</td>
</tr>
<tr>
<td>3.4.2</td>
<td>The success of formal analogies with statistical mechanics</td>
<td>52</td>
</tr>
<tr>
<td>3.4.3</td>
<td>Renormalizability and effective field theory</td>
<td>54</td>
</tr>
<tr>
<td>4</td>
<td>QED, Q.E.D.</td>
<td>56</td>
</tr>
<tr>
<td>4.1</td>
<td>Introduction</td>
<td>56</td>
</tr>
<tr>
<td>4.2</td>
<td>How predictions are made in QED</td>
<td>57</td>
</tr>
<tr>
<td>4.3</td>
<td>State of the art precision measurement and derivation of $\alpha$</td>
<td>60</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Determining $\alpha$ through the electron’s magnetic moment</td>
<td>60</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Precision determination of $\alpha$ through the standard model</td>
<td>63</td>
</tr>
<tr>
<td>4.4</td>
<td>Independent measurements of $\alpha$: The quantum Hall effect</td>
<td>66</td>
</tr>
<tr>
<td>4.4.1</td>
<td>Determining $\alpha$ using the quantum Hall effect</td>
<td>67</td>
</tr>
<tr>
<td>4.5</td>
<td>Using precision tests to go beyond the standard model</td>
<td>68</td>
</tr>
<tr>
<td>4.6</td>
<td>Conclusions</td>
<td>73</td>
</tr>
<tr>
<td>5</td>
<td>Conclusions</td>
<td>74</td>
</tr>
<tr>
<td>Bibliography</td>
<td></td>
<td>76</td>
</tr>
<tr>
<td>Curriculum Vitae</td>
<td></td>
<td>87</td>
</tr>
</tbody>
</table>
List of Figures

2.1  A. The models as mediators view of the interplay between theory, models, and experiment; B. My proposed alternative. Models are split into two types: phenomenological and dynamical, and theory is split into a theoretical framework within which dynamical models are constructed. 

3.1  A single fermion loop contribution to the self-energy of the photon. From this diagram, one can easily understand the terminology of one-loop renormalization: terms like this must be properly renormalized to prove one-loop renormalizability of a dynamical model.

4.1  Scale drawing of a Penning trap, from Brown and Gabrielse [1986].

4.2  The motion of a particle in a Penning trap is a combination of three decoupled motions: magnetron motion at a frequency $\omega_m$, cyclotron motion at $\omega_c$, and an axial oscillation at $\omega_z$ (typically with $\omega_m << \omega_z << \omega_c$). Figure from Odom (2004).
List of Tables

4.1 Measurements of $\alpha^{-1}$. Low-energy QED values are taken from articles cited below. All other values are from Peskin and Schroeder [2018, p. 198]. The most precisely measured values come from low-energy QED tests and condensed matter. ............................................. 59
Chapter 1

Introduction: Theory Construction in Physics

Understanding the methods that bestow epistemic privilege to the products of science has long been a primary goal of the philosophy of science. Physics, widely thought to be the most fundamental and precise of the sciences, has therefore been of particular interest to philosophers. What features of our best physical theories allow for their impressive range and precision of quantitative description of nature? Early philosophy of science sought the answers by examining the logical structure of theories. Philosophy of science was therefore aiming towards a logical language for science, particularly physics. However, philosophers have long realized that scientific theories are not simply handed down from above, fully formed and ready to be applied to the world. Instead, science is a process, through which theoretical frameworks are constructed, new phenomena discovered or predicted, and properties of new entities are tested. All aspects of an experimental science—theory, experiment, and phenomenology—play an important role in this process, and their mutual interactions are the sign of a healthy science.

The goal of this dissertation is to illustrate how high-energy physics exemplified this process of theory construction from the 1950s to 1970s, and the promising ways in which it can continue to do so. The lessons learned from the case studies examined here can inform future physics, and may provide methodological clues as to the best way forward today. Fundamental physics is in a very different epistemic position from the early days of constructing the standard model; in the analysis provided here I acknowledge these differences, but argue that we can still learn much about the dynamics of successful theory construction. My approach to investigating this era is a mix of primary historical analysis and close philosophical scrutiny of the methodology and epistemology involved in each episode. I focus on two major events in the construction of the standard model—the discovery of parity nonconservation in weak interactions and the (re)emergence of Yang-Mills gauge theories as the foundation of the standard model—and contemporary precision testing of quantum electrodynamics. The former are used to illustrate the ways in which theory construction was successfully undertaken in the past, while the latter is an example of how similar methods can still be used in today’s epistemic environment.

Much of the work in the philosophy of high-energy physics focuses on the foundations of quantum field theory. The results of this focus have led to important shifts in understanding the framework on which high-energy physics is built. For example, we now know that the concept
of “particle” used in high-energy physics is very different from the concepts of classical or non-relativistic quantum particles [Malament, 1996, Fraser, 2008], and that the notion of a global vacuum is observer-dependent in relativistic spacetimes [Unruh, 1976, Clifton and Halvorson, 2001]. Critics of the foundational approach have argued that these results, while valuable, are too far removed from the actual practice of high-energy physics, and therefore do not provide insight into what makes high-energy physics, as practiced by the majority of physicists, such a successful science [Wallace, 2006, 2011]. On the other hand, there has been focused work on historically informed philosophy of high-energy physics. Prominent examples include Cao’s (2010) work on structural realism in the transition from current algebra techniques to quantum chromodynamics, Cushing’s (1990) work on the S-matrix program as a rival research program to quantum field theory, or the more recent work produced by the epistemology of the LHC research group [Mättig and Stöltzner, 2018, Stöltzner, 2017, Wüthrich, 2017].

In this dissertation I bridge the traditions of historically motivated philosophy of high-energy physics and foundational analysis of the theoretical framework. I examine the historical episodes (Chapters 2 and 3) or contemporary practices (Chapter 4) in detail, but the questions I am interested in answering should also be relevant to those more concerned with foundational analysis into the framework of high-energy physics. In Chapter 2 I discuss the problems with non-empirical confirmation (cf. Dawid [2013]), and suggest alternatives that may be tenable in today’s epistemic environment. I also comment on the prominent role of symmetries in the sparse theoretical framework of the 1950s, underscoring the continued relevance of symmetries for naturalness arguments in high-energy physics today. Chapter 3 examines the role of formal analogies in the development of symmetry breaking and lattice quantum field theory, and connects the historical role of the importance of renormalizability with more modern views of the standard model as a collection of effective field theories. Chapter 4 deals with precision testing of QED as a means to set narrow bounds on the possible variations of key parameters in the standard model. This will help to limit candidates in the search for future models seeking to go beyond the standard model. Looking at the history and practice of high-energy physics allows one to approach foundational issues in the philosophy of physics from a slightly different perspective. For example, the historical importance of renormalizability as a criterion for acceptable models of the strong and (electro)weak interactions contrasts heavily with the accepted contemporary view of the standard model as a set of effective field theories. Those who take the contemporary view at face value miss out on the details and developments of renormalizability proofs that were instrumental to the emergence of the standard model. Dimensional regularization—the regularization method used to prove the renormalizability of massless Yang-Mills models with spontaneously broken gauge symmetry—is a powerful tool that is valuable outside of the context of renormalizability proofs. By paying attention to their historical role, one becomes aware of their existence, and potential use to remedy foundational issues present today.

A secondary goal of this dissertation is a finer-grained and more nuanced understanding of the various “parts” of a scientific discipline. The division between theory and experiment is neither sharp nor focused; especially in the context of fundamental physics, theory and experiment are closely linked, while there are numerous levels of abstraction from experiment to fundamental equations. In this dissertation I divide high-energy physics into experiment, phenomenological models, dynamical models, and theoretical framework. Roughly speaking, a dynamical model is what is typically referred to as a theory, with quantum electrodynam-
ics being a paradigm example. These are termed dynamical models to distinguish them from the theoretical framework that many models will have in common. In the case of quantum electrodynamics, the underlying theoretical framework is quantum field theory. Phenomenological models have a more limited scope than dynamical models, and tend to be more directly compared to experiment. At all stages, especially in a mature science, the various parts of the discipline all mutually interact with one another. New mathematical methods also play an important role in the process of theory construction. Chapters 2 and 3 detail these distinctions, and use them to frame the discovery of parity nonconservation (Chapter 2) and the emergence of the models of the strong and electroweak interaction (Chapter 3). I argue that these distinctions better capture the relevant parts of high-energy physics at work during the construction of the standard model, and that they are likely to apply more broadly throughout physics.
Chapter 2

Experimental priority in the discovery of weak parity nonconservation

The final publication of this chapter is available at
https://doi.org/10.1016/j.shpsb.2019.05.001

2.1 Introduction

By far the largest scientific enterprise of the twentieth century was the industry of high energy physics (HEP). Post World War II, governments across the world increased funding for experimental particle physics, and encouraged the development of large, expensive collaborative efforts. The drastic increase in funding and focus led to huge experimental leaps. This was a period of proliferating new particle discoveries, many of which defied the minimal known theoretical constraints (cf. Brown et al. [1989]). Along with increased funding for experiment came a concerted effort to organize and predict new phenomena, which led to the rapid development of theories underlying the HEP phenomenology. Both the experimental dominance and the flurry of theoretical work played important roles in the development of the standard model of particle physics, and the rapid progress in the field was facilitated by a close relationship between theorists and experimentalists.

Much of the current literature in philosophy of HEP focuses on the underlying framework of quantum field theory, examining the logic and coherence of theories as they exist today, rather than the historical process of theory construction and the role quantum field theory plays in HEP more broadly. This era is an especially fruitful one to explore, as it provides an excellent modern example of successful theory construction in physics. In the span of about 25 years, the discipline of HEP evolved from a state of theoretical chaos in the wake of an explosion of new particle discoveries, to the establishment of a standard model for particle physics, which has since stood up to rigorous experimental scrutiny for decades. During this time, physicists

1There are exceptions to this trend. For example, histories of the era of theory construction can be found in the series edited by Brown, Dresden, Hoddeson, and Riordan (1989a, 1997), Cushing (1990), Schweber (1994), Cao (2010) all provide historical analyses that draw philosophical conclusions. Fraser (2018), Fraser and Koberinski (2016) are some more contemporary examples of the same. Nevertheless, this sort of work is the minority in the body of work on the philosophy of quantum field theory and HEP.
2.1. Introduction

I gained an understanding of the dynamics behind the three fundamental subatomic forces, and how these gave rise to the behaviour of the myriad particles detected in cosmic ray and collider experiments.

Though it is certainly the case that some theoretical framework was commonly accepted at the time—physicists knew that quantum theory and relativistic effects would be needed to understand the high-energy behaviour of subatomic particles—a widely accepted detailed framework was markedly absent throughout most of the 1950s and 1960s. Quantum field theory initially provided a promising language in which to describe high-energy phenomena, though mathematical and conceptual issues in generalizing quantum field theory beyond the electromagnetic interaction led to distrust and its (temporary) abandonment.

In this paper I will focus on the case of the discovery of parity nonconservation in weak interactions from the period spanning 1947-1957, and the lessons this episode provides for successful theory construction in HEP. The history of the discovery is already well documented (cf. Franklin [1979], Brown et al. [1989a], Franklin [1989], Wróblewski [2008]), and so the purpose of this paper is not to uncover novel historical details. Instead, I aim to (a) summarize the history into a coherent story for philosophers of science, and (b) use the history as a case study for the epistemological evolution of the understanding of weak interactions in HEP. This epistemological evolution is at the heart of the process of constructing the standard model of particle physics.

Analysis of theory construction in the past does more than provide a philosophical depth of understanding to the history of physics; one potential contemporary use for such analysis is to learn generalizable epistemic principles that may serve as useful heuristics for theory construction today. One of the advantages of post-World War II HEP is that many of the current funding and community structures we see in physics were emerging or established here. Generalizations from the construction of the standard model might help with current problems in the frontiers of theoretical physics, especially with attempts to go beyond the standard model. There are of course important differences as well, but the higher degree of sociological similarity may mean that successful practices and heuristics from the 1950s will still be useful today.

One rapidly emerging strand of philosophy of science treats the process of doing science as dependent on models to mediate between theory and experiment. This view is comprehensively outlined in Morgan and Morrison [1999]. The literature on models in science is both wide-ranging and illuminating, highlighting the complexity of scientific testing. The distinction drawn is an important one. Between higher level theoretical principles and experimental production of phenomena, links need to be drawn through the construction of models. Recent work has focused on the nature of models in their own right, without an explicit discussion of how they fit into scientific theories [Jacquet 2016, Weisberg 2013, Godfrey-Smith 2009]. Though a useful framework for understanding local scientific practice—where a particular experiment is compared to a particular theory—the models as mediators view is a bit coarse-grained for the analysis of theory construction as a whole. Therefore, I will apply a further set of distinctions, and emphasize the mutual interactions between all “pieces” of scientific practice. A more fine-grained analysis of the process of theory construction in science reveals some of the ways in which a more robust framework can be built out of phenomenology, experiment, and heuristics. One outcome of this paper is a generalization and expansion of the models as mediators view, with finer distinctions drawn as a means to understand the process of theory
construction. Section 2.2.1 provides the terminology and distinctions I will use. One goal of the present work is to provide a set of distinctions fine enough to account for the details of successful (and unsuccessful) theory construction, but general enough to apply outside of the domain of post-World War II particle physics.

Following this, I give a brief summary of the state of the discipline of HEP in the fifties in Section 2.2.2, followed by a summary of the historical details from the discovery of new unstable particles up to the confirmation of parity nonconservation (Section 2.3). This details the major contributors to the discovery, as well as epistemic advances that altered the framework in which physicists conceived of weak interactions. Finally, I explicate the salient philosophical details of this case study in Section 2.4, starting with important local lessons, and moving on to morals to apply to theory construction in physics today. More general conclusions for the philosophy of science are provided in Section 2.5.

2.2 Distinctions and background

2.2.1 Terminology

My analysis below will split high-energy physics into distinct components. The goal of this division is to clarify how these parts—often considered independently in philosophical analyses of scientific practice, or on a more coarse-grained level—work together in this successful era of theory construction. There are many ways to divide the practice of physics, and I do not claim this division is unique. It does, however, illuminate the high degree of collaboration and interaction between what is typically called theory and experiment in this era, and is sufficiently general to serve as a starting point for analysis elsewhere in science.

By theoretical framework, I mean the network of principles and general mathematical constraints that serve as the common language of a research program. Currently, the theoretical framework underlying high-energy physics is quantum field theory. In the fifties, however, there was distrust in quantum field theory as a general framework, and so the agreed upon theoretical framework was much more minimal. It consisted of a relativistic causal structure and conservation laws carried over from non-relativistic quantum theory. Newtonian classical mechanics is a further example of a theoretical framework in physics, containing concepts such as force, inertia, mass, and so on. Within a theoretical framework, one constructs dynamical models to describe particular interactions. As a paradigmatic example, quantum electrodynamics constitutes a dynamical model in HEP, as it describes the electromagnetic interaction between electrons, positrons, and photons. I use dynamical model in the way most would use the term "theory," to disambiguate the particular models of interactions from the theoretical framework guiding and constraining their construction. I include the word "dynamical" to highlight the fact that in physics, these models are often encoded in some set of dynamical evolution equations. Typically, quantum field theory and quantum electrodynamics would both be described

This is independent of the way in which dynamical models are interpreted. Dynamical models do not require a realist or mechanistic underlying interpretation. The dynamical models in the standard model—quantum chromodynamics and the electroweak model—are still the subject of heavy interpretive controversy, and many physicists involved in its construction take a clear instrumentalist view of the standard model. Nevertheless, the standard model is a clear case of a collection of dynamical models.
as “theories,” though the latter is but an instance—or a model—of the former. Given this distinction, it may be unclear what I mean by “theory construction.” For the purposes of this analysis, theory construction is the process by which a theoretical framework is established, and a consensus collection of important dynamical models emerges within that framework. For HEP, this is equivalent to the construction of the standard model and the working out of quantum field theory as its basis.

Figure 2.1 highlights the contrast between this view of science and the models as mediators view discussed in the introduction. The outlines in the figure highlights the added distinctions: models have been split into phenomenological and dynamical, while theory has been split into dynamical models and theoretical framework. There are a few reasons for this. First, the term “model” is ambiguous. In the first sense, we can understand the term as used in model theory. Then a model is simply an interpretation of a theory. Take, as an example, the theory of general relativity. Mathematically, any model of the theory is given by a specification of a tuple $\langle M, g_{\mu\nu}, T_{\mu\nu} \rangle$ including the manifold $M$, a pseudo-Riemannian metric tensor $g_{\mu\nu}$, and a stress energy tensor encoding the matter content, $T_{\mu\nu}$. In terms of model theory, the class of these models satisfying the Einstein equations constitutes the theory of general relativity, and any particular specification is a model of the theory. Hence, an FLRW cosmological solution to the Einstein equations is a model of general relativity, though it forms the basis for the theory of cosmology. This is not usually the sense of the word “model” meant in the modeling literature in philosophy of science. This second meaning usually refers to partial constructions—with input from the relevant theory, other auxiliary theories, and perhaps phenomenology—meant to more directly model some proper subsystem that falls under a particular theory. The terminology schematized in the figure is partially meant to disambiguate between these two senses of “model”: model-theoretic models would fall under the class of dynamical models, and the theory specified by a particular class of models would be the theoretical framework. Models aimed at representing specific phenomena in the world fit under the heading of phenomenological models. The divisions in Figure 2.1B could likely be split even finer, though my primary focus is on the mutual interaction between all classes, as indicated by the double-arrows. The case of the discovery of parity violation will help to highlight the mutual interactions between, and evolution of, experiment, phenomenological models, and theoretical framework in HEP.

Experiments in HEP produce data, and from these data phenomena are constructed. Phenomena are built from experimental data and expectations given by dynamical models or the theoretical framework. Mathematical methods and tools are used at every step of the process, in order to generate predictions, construct phenomena, and compare the two. A final theoreti-

---

3I use the example of general relativity here because it fits particularly well with model theory. Quantum field theory, on the other hand, is nowhere near as clearly or rigorously defined, and specifying models of quantum field theory in this sense is extremely difficult. The project of constructing mathematically rigorous quantum field theories describing realistic interactions in 4-dimensional spacetime has been carried out since the 1960s, though as of yet no rigorous models have been constructed as viable alternatives to quantum chromodynamics, QED, or the electroweak model.

4The data-phenomena distinction was first explicated by Bogen and Woodward [1988]. This distinction is of some importance to my view, as phenomenological models, though closely connected to experiment, can only be used to describe experimental phenomena, not the data. In what follows I largely gloss over this distinction, and refer to the comparison of experimental data with phenomenological models. These statements should be understood to be shorthand for the conversion of experimental data into phenomena, followed by the comparison with phenomenological models.
Figure 2.1: A. The models as mediators view of the interplay between theory, models, and experiment; B. My proposed alternative. Models are split into two types: phenomenological and dynamical, and theory is split into a theoretical framework within which dynamical models are constructed.

cal tool is the *phenomenological model*, which serves as a mediating link between dynamical models and experiments. Given the coarse-grained view (Fig. 2.1A), phenomenological models would be most closely associated with models generally, though here I emphasize their role as contributing to theory construction. However, even in the absence of a dynamical model, phenomenological models are often constructed based on experimental results and general principles from a theoretical framework. This era of high-energy physics is exemplified by the use of phenomenological models—in the absence of an underlying dynamical model—to make sense of the rapidly advancing experimental results. As I will argue below, the mutual influence between experiment, phenomenological models, and a theoretical framework—most strikingly, the use of phenomenological models to undermine the very framework on which they were constructed—were essential to the construction of the dynamical model of electroweak interactions and the clarification of the underlying theoretical framework of quantum field theory. Before detailing the discovery of parity nonconservation in weak interactions, I will provide a brief “state of the discipline” for HEP in this era.
2.2.2 The state of high-energy physics in “the fifties”

The era commonly referred to in high-energy physics (HEP) at “the fifties” bleeds out into surrounding decades, with 1947-1963 commonly fitting within the era. This era was a period of growing pains in theoretical high energy physics; many early triumphs failed to generalize and few clear guiding principles—in the form of a rudimentary theoretical framework—were carried over from the previous era. The beginning of this era “witnessed the vindication of quantum field theory in renormalized quantum electrodynamics (QED), only to see it rejected as a theory of the strong and weak interactions. It saw the concept of symmetry emerge as a fundamental characteristic of basic physics, followed by its downfall in the parity revolution” [Brown et al., 1989b, p.3].

On the experimental side, massive increases to government funding of physics in the wake of World War II—particularly in the United States—led to a proliferation of advances in accelerator and detector technology, and rapid experimental progress. Coupled with theoretical growing pains, this meant that experimental progress greatly outpaced conceptual development. Theorists were tasked with cataloging, classifying, and systematizing the new experimental discoveries. The relative progress in HEP theory and experiment has markedly shifted since the acceptance of the standard model:

In recent years, when theory called for new particles (such as the W and Z), experiment obligingly provided them but in the fifties experiment outran theory and produced surprise after surprise. Neither the muon nor the strange particles were expected, nor were they welcomed, for the most part, for they destroyed what might have been a consensus for a new unification. [Brown et al., 1989b, p. 4]

Due to the culture of applied science within the American wartime working groups, theorists trained during or shortly after the war “were trained with close links to the experimental practice; they were instilled with a pragmatic utilitarian outlook and were taught to take an instrumentalist view of theories” [Schweber, 1989, p. 671]. Teams worked as integrated units to develop models designed to interpret and utilize experimental findings. Cost of laboratory equipment also meant that resources began to concentrate in certain labs; in order to be in close contact with cutting edge experiments, theorists in turn had to concentrate geographically.

Whatever the causes, the community of theorists marks a large shift from the traditional ways of doing theoretical physics, in which a few prominent theorists worked in relative isolation on the details of personal projects. Many factors led to an increased focus by the community on the most prima facie promising new avenues of research. In general, these were

---

5For a comprehensive historical overview of this era, including sociological and conceptual development, see Brown et al. [1989a]. The final sentence in the above quote will be the subject of some of the historical and conceptual analysis of this era below.

6The situation with experiment has worsened since 1989. With the exception of the Higgs boson discovery in 2012, experiment has failed to provide new particles in accordance with the most “natural” beyond standard model dynamical models (e.g., supersymmetry, SU(5) grand unification, technicolor, etc.). As I discuss below and elsewhere (Chapter 3), this lack of empirical evidence poses a major problem for current theory construction in HEP.

7The history of scientific development has often seen individuals working largely in isolation on projects of interest. Though physics had progressed considerably from the days of Shapin’s (1991) “gentleman scholar,” the break between pre- and post-war physics was dramatic.
phenomenological in nature; theoretical developments that had an obvious connection to new experiments were often seen as the most promising. [Feynman 1954] referred to this phenomena as the “pack effect,” in which the community quickly mobilized around the dominant approach. A consequence of this is that “[t]he community at any one time seems to be dominated by the outlook and suggestions of a single individual... Gell-Mann, Goldberger, Lee, Yang, and Chew were the dominant figures from the mid-1950s to the mid-1960s,” and were the individuals from whom the dominant approaches to theoretical particle physics post-QED originated [Schweber 1989, p.673].

There was a strong tension at the theoretical foundations of particle physics in this era. The success of QED resulted in a desire to expand the domain of quantum field theory to the nuclear interactions, while technical and conceptual problems at the heart of quantum field theory led many to distrust it as a framework even for QED, let alone the nuclear interactions. By the start of the 1960s a supposed rival program—the S-matrix program led by Chew, Goldberger, and others—emerged as a replacement for quantum field theory in hadronic physics (cf. Cushing [1990]). Though quantum field theory was still the common language in which models—both phenomenological and dynamical—were formulated in the fifties, troubles with renormalization and strong couplings led to a dismissal of quantum field theory as a framework suitable to high energy nuclear physics. Much of the effort in theoretical HEP in this era was devoted to developing a new theoretical framework for hadron physics and the strong interaction.

Symmetry principles played an important role in classifying many of the newly discovered particles. New quantum numbers were introduced, and particles were represented via group theory as simultaneous eigenstates of the Poincaré group and the additional quantum numbers—spin, flavour, strangeness, etc. Among the spacetime symmetries were the discrete symmetries of time reversal, charge conjugation, and parity conjugation. Time reversal involved replacing time variables and temporal derivatives \( t, \partial_t \to -t, -\partial_t \). Charge conjugation replaced particles with antiparticles, and switched the sign of electric charge. Parity conjugation is a left-right mirror flip of three-dimensional space. All fundamental interactions were thought to be invariant under each of these symmetry operations, and each interaction would thus conserve charge and parity.

The weak interactions were still a bit of a mystery post World War II, with only the Fermi theory of \( \beta \) decay having any success at capturing the phenomena. The Fermi [1933] interaction was governed by the following Lagrangian density:

\[
\mathcal{L} = G_F \left( \bar{u}_\nu \gamma_\mu u_\nu \right) \left( \bar{u}_\nu \gamma^\mu u_\nu \right),
\]  

(2.1)

where \( u_i \) denotes a Fermion field, \( \bar{u}_i = u_i^\dagger \gamma_0 \) its Dirac conjugate field, and \( \gamma^\mu, \mu = 0, 1, 2, 3 \) are the Dirac matrices. The subscripts on the Fermion fields denote protons, neutrons, electrons,

---

8 For a more detailed discussion about the founders of QED and their attitudes toward the shortcomings of the new theory, see [Schweber 1994]. Dyson and Schwinger were particularly critical of QED, and viewed it as a low-energy approximation to some more complete theory.

9 Though S-matrix theory was taken to be a replacement of quantum field theory as a more sparse theoretical framework—one which linked directly to phenomenological models with a rejection of dynamical models—it turned out that everything done within the S-matrix program was entirely compatible with quantum field theory, and moreover that the principles of S-matrix theory were insufficient for uniquely determining the form of phenomenological models, as was once hoped. (Again, cf. [Cushing 1990].) At the time, however, it was treated as a serious competitor to quantum field theory, with a distinct set of underlying core concepts.
and neutrinos, respectively. Each term in parentheses can be thought of as an analog of a Dirac current—the first describing the transition from a proton to a neutron, the second describing the production of the electron and neutron. The new coupling constant introduced, $G_F$, is the Fermi coupling, and was later found to be universal for weak interactions. First, note that this is a point interaction—all of the Fermions interact by contact at a point. Unlike the recently accepted dynamical model quantum electrodynamics, the Fermi weak interaction was not mediated by force carrying bosons. Neither was electric charge constant for the analogs of the Dirac currents. Fermi’s model offered success beyond $\beta$ decay, and was able to describe any processes in which a proton, neutron, electron, and (anti)neutrino were interacting at a point. The new particles discovered in the late 1940s and early 1950s (to be discussed below), however, were not easily accommodated within the model. Additionally, experiments in the fifties indicated that at higher energies, the weak interaction was likely mediated by massive Bosons, shifting the Fermi interaction from candidate dynamical model to low energy phenomenological model of a limited class of weak interactions.

In the absence of the quantum field theoretical framework, the construction of a dynamical model underlying weak interactions was difficult. As discussed in the next section, phenomenological models were being proposed to attempt to understand the exotic properties of the newly discovered particles. These were based on broadly kinematic constraints—for example, quantum statistical properties for the particles obeying relativistic momentum-energy relations, introduction of new quantum numbers, and so on—and respect for otherwise well-established symmetries and conservation laws. The rejection of parity conservation for the weak interaction seems all the more surprising given this context of impoverished conceptual resources. I hope to make clear, through analysis of this case study, the means through which a richer framework was constructed on the basis of phenomenology, experiment, and these very symmetry principles.

### 2.3 A (brief) history of parity nonconservation

The first steps toward parity nonconservation and a broadened understanding of the weak interactions began with the experimental discovery of new “V-particles” in cosmic ray experiments, so named for their characteristic V-shaped decay patterns. [Rochester and Butler] (1947) initially discovered two new types of particle, one neutral and one charged. Detection from cosmic rays limited the amount of information learned about these new particles, but it was estimated through kinematic constraints that the mass of the new particles was on the order of $300m_e - 1000m_e$, with $m_e$ the mass of the electron. One interesting feature highlighted later was the relatively long lifetime of the V-particles, compared to timescales associated with strong hadronic interactions. The original cosmic ray experiments established that these were indeed decay processes, and not scattering with matter:

> Further, very few events at all similar to these forks have been observed in the 3-cm. lead plate, whereas if the forks were due to any type of collision process one would have expected several hundred times as many as in the gas. This argument indicates, therefore, that the tracks cannot be due to a collision process but must be due to some type of spontaneous process for which the probability depends on
the distance travelled and not on the amount of matter traversed. [Rochester and Butler, 1947, p. 855]

In the following years after Rochester and Butler’s initial discovery, many other reports of V-particle production had been published, both in accelerator experiments and further cosmic ray studies establishing tighter bounds on the masses, lifetimes, and statistics of the V-particles [Serriff et al., 1950, McCusker and Millar, 1951, Hopper and Biswas, 1950, Bridge and Annis, 1951, Thompson et al., 1951, Leighton et al., 1951, Fretter, 1951, Armenteros et al., 1951]. Once the V-particles were known, refined experiments could be designed to isolate them and determine their properties. It became clear that (a) many V-particles existed, with at least five unique decay modes being detected; and (b) that the process through which V-particles were created occurred on a timescale \( \sim 10^{11} \) times faster than the decay timescale. Theorists were greatly concerned with explaining this feature, and various phenomenological rules were proposed for limiting the possible decay pathways of the V-particles. For the timescales of production and decay to be so vastly different, different coupling strengths would be required, and this implies that different forces are at work in the two processes.

The most successful of these phenomenological rules was that of associated production, first proposed in detail by Pais [1952]. He proposed a model based on the known nucleon-pion coupling, and assumed two coupling constants, one for the production of V-particles and one for their decay. The most striking features to be described by the associated production were the following:

(a) In high energy events V-particles are produced with a probability \( \gtrsim 1 \) percent of the \( \pi \)-meson production. Thus, the production is copious.
(b) These new particles have lifetimes \( \gtrsim 10^{-10} \) sec...
[(c)] if one would consider the same mechanism which produces them to be instrumental for their decay, one would estimate lifetimes \( \tau \) of the order of \( 10^{-21} \) sec. [Pais, 1952, p. 663]

The nucleon-pion coupling was used as a template, with both production and decay vertices taking the form \( \bar{\psi}\psi\phi \)—that is, a coupling term bilinear in the Fermion \( \psi \) and its adjoint, and linear in the Boson \( \phi \). Since even the statistics of the V-particles (as well as the number of distinct particles existing) was not conclusively established, Pais assumed a symmetry between charged and neutral V-particles: for each, there existed a Fermion of mass \( \sim 200m_e \) and a Boson of mass \( \sim 800m_e \). Associated production—here the requirement that an even number of V-particles enter every strong vertex interaction—was added in as a constraint by generalizing the nucleon-pion interaction; both strong and weak interaction take the form \( (\bar{\psi}_i\psi_j\phi_k) \), where \( i, j = 0 \) denotes nucleons, \( k = 0 \) denotes pions, \( i, j = 1 \) denotes Fermionic V-particles, \( k = 1 \) denotes Bosonic V-particles. The assumption was that a strong coupling can only occur when \( i + j + k = \) an even number. The idea behind this is that V-particles are produced in pairs, and quickly separate such that a strong decay can no longer occur. The much weaker interaction occurs when \( i + j + k = \) an odd number, which is why the lifetimes of decay are so much longer than would be expected if governed by the strong force. A consequence of this model is that, should large quantities of V-particles undergo collisions, interactions in which evenness was not conserved would be suppressed compared to evenness conserving interactions by a factor proportional to the ratio of the weak to strong coupling constant. Given the evidence of
the great disparity of timescales involved in the two interactions, this would imply an effective “conservation of evenness” in V-particle beam collisions.

2.3.1 Strangeness

Pais’s model, and others much like it, were explicitly phenomenological in nature. The associated production rule is simply stipulated, with no dynamical mechanism underlying it. It was, however, guided by a framework that relied heavily on symmetry principles, conservation laws, and conservative extensions of previously successful models. Both Gell-Mann [1953] and Nishijima [Nakano and Nishijima, 1953; Nishijima, 1955] continued this phenomenological modeling, by extending the associated production model to a more systematic introduction of a new quantum number, conserved in strong and electromagnetic interactions, but not in weak interactions. Gell-Mann [1956] later called this quantum number “strangeness,” summarizing the bizarre symmetry-violating state of weak interactions.

A central concept involved in both Gell-Mann and Nishijima’s analyses was that of the separability of the fundamental forces between particles, and the hierarchy of symmetries that this induced. The stronger the interaction, the more invariant quantities that interaction allows. The hierarchy picture explains why many of the conservation laws used as heuristics in constructing phenomenological models and candidate dynamical models of, for example, the strong nuclear interaction were so successful. To a first approximation, one can treat high-energy physics as formed only of the strong nuclear interaction, which is charge independent and conserves isospin. This works to a first approximation because the magnitude of the strong coupling constant is far greater than those for the other two forces. To a first approximation, then, the pion and nucleons are all stable, because electromagnetic and weak decay modes are “turned off.” The next level of approximation—found to be an approximation in light of the new weak decay processes of the V-particles— involves “turning on” the electromagnetic interaction by coupling the photon to charged matter. The electromagnetic interaction is a perturbation to the dominant strong interaction, and violates the conservation of isospin. However, isospin is still a useful approximate symmetry, since the cross-sections for isospin nonconserving electromagnetic interactions are much smaller than the isospin conserving strong interactions, provided both interactions are relevant. Charge and the third component of isospin are also conserved by the electromagnetic interaction. The strength of the interactions in the hierarchy is indicated by the characteristic lifetimes associated with each force. Strongly interacting particles decay with a lifetime around $10^{-23}$ s, typical electromagnetic decays are on the order of $10^{-17} - 10^{-15}$ s, while the new weak decays were around $10^{-9}$ s.

In analogy with the conservation of isotopic spin and baryon number—which are both only approximately conserved due to electromagnetic and weak effects—Nishijima [1955] and Gell-Mann [1956] each proposed a new quantum number, conserved for all but weak interactions.

---

10There is a slight complication to this picture, since not all particles interact via the strong force. In modern parlance, the strong interaction is a residue of the quark-gluon interaction, and only affects particles composed of colour-neutral collections of quarks. These particles—consisting of baryons like the proton and neutron, as well as mesons like the pion—are referred to as hadrons. The electromagnetic and weak interactions, in contrast, affect all matter. So the hierarchy picture is best suited to hadronic physics, though the electromagnetic and weak interactions as successive approximations still applies. Historically, hadronic physics was the template and focus in the fifties, so that this issue arises mostly retrospectively.
defined as an additional term added to the charge-isospin relation, \( q = I_3 + B/2 + (1/2)S \), with \( q \) the electric charge, \( I_3 \) the third component of the isospin vector, \( B \) the baryon number (1 for Baryons, 0 for mesons), and \( S \) the strangeness quantum number. The new quantum number allowed for useful heuristics to be developed to the second approximation (weak interactions still neglected). At this level of approximation, “the conservation of strangeness gives rise to two important qualitative effects:

(1) The law of stability: A strange particle cannot decay rapidly into ordinary ones.
(2) The law of associated production: In a collision of ordinary particles, there can be no rapid formation of a single strange particle; there must be at least two of them and the total strangeness must be zero” [Gell-Mann, 1956, p.853].

Gell-Mann and Nishijima therefore made progress in explaining the phenomenological characteristics of associated production and the long lifetimes of the V-particles, by assuming a hierarchy of independent forces, and positing a related new quantum number. Their proposal—which fit better with the minimal theoretical framework involving symmetry constraints and general principles governing scattering—altered and relaxed the current conservation laws (via the introduction of strangeness into the charge-independence equation) and helped to explain why violated symmetries were still useful as approximations. The work, however, was still largely phenomenological, since the theoretical developments of Nishijima and Gell-Mann were largely concerned with explaining and categorizing the observed phenomena. No dynamical origin for the force hierarchy or the new quantum number were given.

### 2.3.2 How many V-particles?

Associated production and the introduction of strangeness helped to explain some of the most striking qualitative features of the new V-particles, but quantitative theoretical results remained elusive. On the experimental side, more and more data were being produced, leading to a more robust experimental understanding of the properties of the new particles. To start, there were five known, distinct decay modes, the originating particles of which all with equal masses and lifetimes, up to experimental error. The particles later became known as kaons, and the possible new particles were labeled by their decay modes: \( K_{\pi^3} \), \( K_{\pi^2} \), \( K_{\mu^2} \), \( K_{\mu^3} \), and \( K_{e^3} \), with \( \pi \) indicating a decay into pions, \( \mu \) into muons, and \( e \) into electrons. The first two of these decay modes were originally labeled \( \tau \) and \( \theta \), respectively, and gave rise to what is now commonly called the \( \theta-\tau \) puzzle.

The \( \theta-\tau \) puzzle was particularly interesting in the early 1950s because these two decay modes were the first to display the striking similarities in lifetime and mass. However, the intrinsic parity of a pion is \(-1\), and since parity is multiplicative and conserved\(^{11}\), a straightforward comparison of the simplest possible decay scheme showed conflicting parity for the \( \theta \) and \( \tau \) decay modes. The overall parity for \( \theta \) was relatively straightforward: since \( \theta^0 \rightarrow \pi^0 + \pi^0 \), and the wavefunction of the right hand side had to be symmetric—recall identical Bosons must

\(^{11}\) The balance on both sides of a particle interaction is actually \( P^{j+l} \), where \( P \) is the product of intrinsic parities for each particle, \( j \) the spin of the particles, and \( l \) the extrinsic angular momentum. Thus, even though the identity of \( \theta \) and \( \tau \) would have violated parity conservation if \( j + l = 0 \) for both decays, the possibility that more complicated decay pathways would restore the parity balance remained open for a few years. It was only through Dalitz’s detailed analysis of the properties of the \( \tau \) particle that the discrepancy became pressing.
have a combined wavefunction symmetric under particle exchange, and the parity factor is multiplied under particle exchange—the $\theta^0$ decay mode had to have even parity.

The $\tau$ decay was harder to characterize, and Dalitz [1953, 1954, 1955] undertook detailed analysis of the experimentally confirmed $\tau$ events to attempt to characterize the possible spin-parity configuration for $\tau$. To this end, Dalitz constructed two-dimensional plots representing the distribution of energy from the three decay pions, which would vary according to assumptions about the spin-parity configuration of the $\tau$ meson [Dalitz, 1953]. As more events came in, the histograms of energy distributions could be more closely matched to the theoretical predictions, ruling out certain spin-parity configurations inconsistent with the evidence. By the discussion at the sixth Rochester conference—to be outlined in greater detail below—Dalitz had noted that with 600 events, the Dalitz plots were remarkably uniform, and that “the simple-minded interpretation [of the Dalitz plots] is that the distribution is uniform. This would point to a $\tau$ meson of spin-parity $0^-$ though other possibilities, such as $2^-$ are not excluded” [Oppenheimer et al., 1956, VIII. 20]. The Dalitz plots could also be made to match the observed pion distributions by supposing that $\tau$ had even higher spin values, but the higher the supposed spin, the more ad hoc the proposal, since higher spin values introduce more freedom into shaping the energy distribution curves for the decay pions. Given that no other known particles had high spin values, the basis for introducing them here amounted to little more than curve-fitting, though of course the possibility of a $\tau$ meson with high spin could not be ruled out in principle.

This led to a clear tension with the $\theta$ meson: the evidence indicated that $\tau$ had an odd parity, while $\theta$ was constrained to have even parity. A model proposed by Lee and Orear [1955] tried to resolve the puzzle with the assumption that they are two different particles, and that one decays rapidly into the other with a significant branching ratio: $\tau \rightarrow \theta + \gamma$ or $\theta \rightarrow \tau + \gamma$. The $\theta$-$\tau$ puzzle, and the issue of the number of kaons more generally, was a prime topic of discussion at the Saturday morning session of the 1956 Rochester conference. Oppenheimer opened the discussion with the following remark:

There are the five objects $K_{33}, K_{22}, K_{42}, K_{43}, K_{e3}$. They have equal, or nearly equal, masses, and identical, or apparently identical, lifetimes. One tries to discover whether in fact one is dealing with five, four, three, two, or one particle. Difficult problems arise no matter what assumption is made. It is to this problem of the identity of the K particles that a larger part of the present discussion is devoted [Oppenheimer et al., 1956, VIII. 1].

The Lee and Orear model was quickly rejected at this conference because of the lack of observed gamma rays in recent experiments reported by Alvarez [1956]. In the course of the discussion, Yang expanded on Oppenheimer’s initial comments with the following:

Of course, if they are all different decay modes of the same particle, the puzzlement would vanish…However, the situation is that Dalitz’s argument strongly suggests that it is not likely that $K^+_{33}(=\tau^+)$ and $K^+_{e2}(=\theta^+)$ are the same particle [because of opposite parity]. [Oppenheimer et al., 1956, VIII. 8]

\[^{12}\text{The notation here is } j^{\text{sign}(P)}, \text{ such that } 0^- \text{ denotes a spin zero particle with odd intrinsic parity.}\]
Perhaps inspired by the introduction of strangeness as a generalization of the charge-independence equation, Yang went on to suggest a generalized parity conjugation operation, in which some weakly interacting particles would behave as multiplets rather than invariants. This remark was the first serious suggestion that the weak interaction does not conserve parity. Yang’s haste was quickly reigned in by other participants. After summarizing Dalitz’s extensive analysis of the properties of the \( \tau \) meson, Marshak urged participants to accommodate both parity conservation and the identity of the \( \theta \) and \( \tau \) mesons, “even if one has to use a larger spin value. He felt that a last effort should be made in this direction, before introducing any startling new assumptions” [Oppenheimer et al., 1956, VIII-18]. Here we see the conviction that attempts at curve-fitting, so long as they are motivated by currently accepted and well established principles, is preferable to introducing modifications to the theoretical framework to accommodate striking new phenomena.

Gell-Mann proceeded to discuss the idea that \( \theta \) and \( \tau \) were now different particles in a parity multiplet, but noted that his model did nothing to explain the equality of their lifetimes. At this point the participants were encouraged to open their minds and consider somewhat more radical proposals.

Pursuing the open mind approach, Feynman brought up a question of Block’s: Could it be that the \( \theta \) and \( \tau \) are different parity states of the same particle which has no definite parity, i.e., that parity is not conserved. That is, does nature have a way of defining right [or] left-handedness uniquely? Yang stated that he and Lee looked into this matter without arriving at any definite conclusions... Perhaps one could say that parity conservation, or else time inversion invariance, could be violated. Perhaps the weak interactions could all come from this same source, a violation of space-time symmetries. [Oppenheimer et al., 1956, VIII-27-28]

After some further speculation by Gell-Mann and Michel, “[t]he chairman felt that the moment had come to close our minds,” and return to less speculative ideas. (VIII-28).

The discussion led Lee to wonder if parity would be universally violated in weak interactions, and not just in the \( \theta-\tau \) decay modes. Shortly after the conference, Lee learned from his colleague, Chien Shiung Wu, that nobody had undertaken to confirm conservation of parity in any known weak interactions [Wróblewski, 2008, p. 255][13]. Lee and Yang then quickly undertook to provide experimental means for testing the hypothesis of a handedness of weak interactions, and their paper was submitted to Physical Review on June 22, 1956.

### 2.3.3 Lee and Yang

The Lee and Yang [1956] paper begins with the hypothesis that parity is not conserved in weak interactions as a resolution to the \( \theta-\tau \) puzzle, and note that “existing experiments do indicate parity conservation in strong and electromagnetic interactions to a high degree of accuracy, but that for the weak interactions (i.e., decay interactions for the mesons and hyperons, and various Fermi interactions) parity conservation is so far only an extrapolated hypothesis unsupported

---

[13] Wróblewski provides an excellent summary of the progress of Lee and Yang in writing and disseminating their celebrated paper, as well as the rapid experimental tests of parity nonconservation by Wu and others. The following timeline is taken from that paper.
by experimental evidence” (p. 254). A consequence of parity violation is that the “true” atomic and nuclear quantum states are actually a weighted superposition of the usual state assumed and its opposite parity counterpart. The degree of parity violation can be characterized by the fractional weight of this opposite parity portion of the true quantum state. Various preexisting experimental evidence places an upper bound on the fractional weight between $10^{-2}$ and $10^{-6}$, depending on the process considered. Lee an Yang propose a few different experimental tests of parity violation, the most notable of which is the $^{60}$Co experiment.

A relatively simple possibility is to measure the angular distribution of the electrons coming from $\beta$ decays of oriented nuclei. If $\theta$ is the angle between the orientation of the parent nucleus and the momentum of the electron, an asymmetry of distribution between $\theta$ and $180^\circ - \theta$ constitutes an unequivocal proof that parity is not conserved in $\beta$ decay... The angular distribution of the $\beta$ radiation is of the form:

$$I(\theta)d\theta = (\text{constant})(1 + a\cos\theta)\sin\theta d\theta \ldots \quad (2.2)$$

if $a \neq 0$, one would then have a positive proof of parity nonconservation in $\beta$ decay. (p. 255)

Other experiments were mentioned, but rejected as being impractical. In measuring $\beta$-$\gamma$ correlation, for instance, polarization of the emitted $\gamma$ radiation would indicate violation of parity conservation, but “this polarization must be circular in nature and therefore may not lend itself to easy experimental detection” (p. 256).

Wu learned about the Lee and Yang paper from Lee, and resolved to try the $\beta$ decay asymmetry experiment with $^{60}$Co before the paper was even published. During the time of the experimental design, a preprint of the Lee and Yang paper was circulating, and still many physicists—including Landau, Pauli, and Feynman—bet against violation of parity conservation:

During the October 1956 meeting in Russia Lev Landau still maintained that parity nonconservation was an absolute nonsense. Richard Feynman bet Norman Ramsay 50$ to 1$ that experiments would prove Lee-Yang hypothesis wrong. He later paid [[Frauenfelder and Henley [1975] p. 389]]. As late as 17 January, 1957, Wolfgang Pauli wrote to Victor Weisskopf: “I do not believe that the Lord is a weak lefthander, and I am ready to bet a very large sum that the experiments will give symmetric results”. Just after sending off the letter he learned about the outcome of the experiments at Columbia. [[Wróblewski 2008 p. 258]]

The Wu et al. [1957] experiment indicated parity nonconservation in early January, 1957. Lederman’s team [1957] learned of the Wu results, and quickly undertook the pion-muon decay experiment proposed elsewhere in the same paper by Lee and Yang. The findings of Wu’s group were thus corroborated at the Columbia cyclotron, and two experimental confirmations of parity nonconservation were attained by January 8th, 1957. Further tests showed the universality of parity nonconservation, and thus the universality of weak interactions was suggested, as well as the more radical rejection of conservation of parity. This was a major step towards a dynamical model of the weak interaction, which won Lee and Yang the 1957 Nobel Prize in Physics. A major symmetry assumed in the theoretical framework was relaxed for weak interactions, and an understanding of the phenomenology of weak decays gained.
2.3.4 Toward electroweak unification

One major consequence of parity violation in weak interactions was that physicists expected charge conjugation symmetry to be violated, given the newly required two-component form of the neutrino. The CPT theorem proved that all local relativistic theories must be invariant under the combined symmetries of time-reversal (T), parity conjugation (P), and charge conjugation (C). Given that the weak interaction was known to violate parity conjugation, the simplest step forward was that one of the other two symmetries would be broken, rendering the combined action of either PT or CP an exact symmetry.

Lee and Yang [1957]—as the results from Wu et al. were still incoming—proposed a new, two-component field expression for the neutrino. In such a formalism, neutrinos are massless, and their spin and momentum vectors always align. This formalism had been considered previously, but was rejected due to the fact that the two-component formalism maximally violates parity (or “space inversion”). Under a parity transformation, the neutrino would change momentum direction, but not spin. This is not a state of the two-component representation, so the P operator does not leave the neutrino invariant. Lee and Yang also interpret the two components of the neutrino as particle-antiparticle components: the case of a spin-momentum product of 1/2 corresponds to a neutrino, while the product of -1/2 corresponds to the antineutrino. So the formalism predicts that neutrinos are not invariant under the charge conjugation operation, since this takes particles into antiparticles without changing the spin or momentum directions. However, the two-component formalism is invariant under the combined CP operation, since both momentum and particle-antiparticle are reversed, but not spin.

Eventually, this two-component formalism had to be further modified, in light of experimental results indicating that some weak processes violated CP symmetry as well [Kobayashi and Maskawa, 1973]. However, the Glashow-Weinberg-Salam electroweak model maintains a two-component neutrino field, although in this case there are separate fields for left- and right-handed neutrinos. The two-component, parity violating neutrino fields played an important part in the eventual formulation of the electroweak model. The violation of P- and CP-invariance also highlight how little is understood about the discrete spacetime symmetries. Whether CPT holds as an exact symmetry is still unknown, and has major consequences for the formalism in which HEP is cast. If CPT is violated, then an exact description of HEP cannot be in the form of a local relativistic theory.

In a less obvious route of influence, parity violation solidified the idea that the weak interactions are exceptional. The process outlined above involved introducing a new quantum number—strangeness—that was conserved in all but the weak interactions. Next, the weak interaction was shown to violate parity conjugation, thought to be a global discrete symmetry. Shortly after the discovery of parity violation, ideas about spontaneous symmetry breaking from condensed matter physics were imported to HEP (cf. Fraser and Koberinski [2016]). Nambu and Jona-Lasinio [1961] first proposed an effective model of strong interactions based of the spontaneous symmetry breaking of the BCS model of superconductivity. Here, we see the breakdown of a global, continuous chiral symmetry, which resulted in massive force mediating bosons. Though the Nambu Jona-Lasinio model was merely approximate—their model

---

14 Streeter and Wightman [1964] rigorously prove the CPT theorem in the framework of axiomatic quantum field theory, though the theorem was taken to hold well before the more rigorous proof.
predicted the presence of massless bosons that were identified with the light pions—physicists thought that spontaneous symmetry breaking could be the appropriate mechanism for introducing massive gauge bosons into a theory of weak interactions. The weak interaction seemed to break many other expected symmetries, so spontaneous symmetry breaking would fit the pattern well. However, spontaneous symmetry breaking was soon shown to result in the presence of massless Goldstone bosons [Goldstone, 1961]. Since these should appear abundantly in weak interactions, their absence was taken as evidence that the weak interaction did not employ spontaneous symmetry breaking. [Higgs, 1964], among others, showed how the Goldstone bosons could be avoided if it was a gauge symmetry—a continuous, local symmetry—that was spontaneously broken. As is well known, the Higgs mechanism plays a central role in the electroweak unification. We can draw a route from the violation of local discrete symmetries (strangeness quantum number), to global discrete symmetries (parity, charge conjugation), to global continuous symmetries (global phase) to local continuous symmetries (Higgs mechanism), all of which were fundamental to the formulation of the electroweak model. Confirmation of parity violation in 1957 indicated that the weak interaction could be understood by violating expected symmetry principles.

2.4 Analysis

The period from the discovery of the new V-particles to the consensus that parity conservation was violated in weak interactions is an illuminating example of the early stages of theory construction for a number of reasons. As mentioned in the introduction, the rapid development of HEP from the end of World War II to the cementing of the standard model as the dynamical model underlying all known high-energy phenomena in the mid 1970s is a clear example of a highly successful episode of theory construction. By studying this era, we can learn both the local facts that made it so successful, potential barriers that delayed success, and hopefully even general lessons about theory construction in physics more generally. I will tackle all three of these in the following analysis: In Section 2.4.1 I highlight the relevant local features of the discovery of parity nonconservation that led to such rapid progress. This is a clear example of the early stages of theory construction, in which there is a prominent absence of a dynamical model. Section 2.4.2 contrasts these features—especially the lack of plausible candidate dynamical models—with the wake of the new quantum theory, to highlight a potentially necessary condition for the discovery of parity nonconservation. Finally, I compare the local features of HEP in the fifties to modern theoretical physics in Section 2.4.3 and suggest that the marked lack of experimental data is a potential barrier to rapid progress in theory construction today.

2.4.1 Lessons from theory construction in high-energy physics

Given the division of a scientific discipline outlined in Section 2.2.1—into a theoretical framework, dynamical models, phenomenological models, experiment, and mathematical tools—the most striking absence from my discussion of the discovery of parity nonconservation is that of dynamical model construction for the weak interaction. There are a few reasons for this glaring absence. First, the discovery of the new V-particles was entirely unexpected on existing theoretical grounds; efforts were concentrated on solidifying their phenomenology before any
detailed dynamical models could be constructed. The work of Pais, Gell-Mann, Nishijima, and Dalitz detailed in Section 2.3 all fits into this broadly phenomenological brand of theorizing. Production and decay rules were posited to make sense of the lifetimes and products of the new unstable particles, without worry about convincing mechanisms that would lead to such rules. Pais’s model of associated production makes this rather clear: after adopting a numbering convention for Fermionic and Bosonic matter, a generalization of the accepted nuclear coupling is assumed, with coupling strength depending on the sum of the arbitrary numeric labels introduced. As the theory of weak interactions progressed past the discovery of parity nonconservation, it turned out that neither associated production nor strangeness were to play a starring role. However, their usefulness in understanding the phenomenology of V-particles should not be downplayed. Especially important was the idea of Gell-Mann and Nishijima that there was a hierarchy of three forces in HEP. The three could be treated largely independently, meaning that to various degrees of approximation, one could treat the forces as entirely separable.

Second, the absence of efforts to construct a dynamical model underlying the behaviour of the V-particles can be attributed in part to the lack of data on exotic weak phenomena. Fermi’s model of the weak interaction was specific to $\beta$ decay and variations involving the proton, neutron, electron, and neutrino. Nowhere in the Fermi model was there an easy opening in which to introduce the V-particles, especially those thought to be Bosonic. The known properties of the V-particles were consistent with their decay being governed by the Fermi coupling constant, and the simplest explanation was therefore that the V-particles were mediated by the weak interaction, but this was far from the only option, and new conflicting data would have undermined the supposition of a universal weak interaction. Related to the first point, the phenomena were so new and unexpected that it wasn’t clear what sort of phenomena they were. There was thus insufficient dynamical foundation on which to start constructing any dynamical models.

Third, the community of high-energy physicists in the fifties had a broadly experiment-focused outlook. As mentioned in Section 2.2.2, World War II had a major effect on both the social organization of the HEP community, as well as the philosophical outlook regarding the role of theory and experiment in HEP. The leading theorists—including Gell-Mann, Feynman, Dyson, and Chew—adopted a pragmatic outlook, and viewed theories as instruments through which one could understand experimental phenomena. Thus, the desire for a dynamical account of the weak interaction was of less importance; what mattered was getting a firm grasp on the phenomena. Experiment came first in the fifties, and theory was meant to categorize and explain what was observed. This social structure and philosophical outlook continued throughout the 1960s as well, and led to the rise of the S-matrix program as an explicitly phenomenological framework for hadronic physics (cf. Cushing [1990]).

Usually, new phenomena can be modeled dynamically within a discipline when the theoretical framework is robust and detailed. When the concepts underlying the discipline are...
2.4. Analysis

sufficiently detailed and formalized, they provide an important set of constraints on dynamical models. Coupled with a minimal phenomenology, these may be sufficient to start building dynamical models to explain novel theoretical results. In contrast, HEP in the fifties was in a state of theoretical flux; following the early success of QED and the underlying framework of quantum field theory, technical and conceptual issues led to major distrust in quantum field theory as an adequate theoretical framework for HEP generally. This is the final reason for the absence of efforts to construct a dynamical model in this case study. Dynamical models sit at the centre of the aspects of a scientific discipline, and require input from phenomenology, mathematical tools, and the theoretical framework. In the fifties, HEP was in the process of developing a new framework, and these beginnings relied on a minimal set of symmetry principles and scattering balance equations as guidance. These framework principles were just enough to guide the construction of phenomenological models, but were too vague to provide dynamical guidance, especially while the phenomenology of V-particles was unsettled. By the time that parity nonconservation was discovered, these tools had developed in sophistication, and dynamical models became the focus of HEP.

What the discovery of parity nonconservation highlights best is a successful first stage of theory construction in HEP. With minimal phenomenology and theoretical framework, and no dynamical model, physicists were able to mutually develop a robust phenomenological picture of the V-particles and amend the theoretical framework with which they were working, priming the discipline for the development of more and more detailed dynamical models culminating in the electroweak unification model in 1967. This is a clear case where new phenomena were discovered, and only minor revisions to the theoretical framework were required, because the framework was so underdeveloped at the time.

It may seem striking that construction of phenomenological models guided only by experiment and symmetry principles would lead to the discovery of a violation of one major symmetry principle, and that this perhaps indicates that symmetry considerations played less of a constraining role in model construction than I am claiming. Upon closer examination, however, the small steps taken along the way highlight a continuous deformation of the theoretical framework to accommodate parity nonconservation, rather than an abrupt break. It was these constraints imposed by the symmetry framework that made it clear that modifications would be needed, upon pain of conflict with a flourishing experimental and phenomenological program.

I will expand on what I mean by a continuous deformation of the theoretical framework. At no point during the history outlined above was there a model proposed that was a stark violation of the accepted minimal framework. Pais’s associated production hypothesis was built out of a conservative extension of the nucleon-pion coupling, and the modification made was the minimal necessary to account for the discrepancy in timescales associated with the production and decay of the new particles. First, a strong coupling—of the same order of magnitude as that for pion-nucleon interactions—was extended to V-particle production, for which V-particles could only be created or annihilated in pairs. Dimensional analysis based on decay lifetimes required that the coupling involving odd numbers of V-particles would have to be many orders of magnitude weaker, and that this coupling was consistent with the Fermi coupling constant. Here the continuity with accepted physics is clear: the form of interaction and the coupling constants are taken from established models, and a new selection rule is posited as an additional principle. Strong interactions obey the associated production principle,
while the weak interactions do not. Here we see continuity on the basis of analogical reasoning within the discipline of HEP.\footnote{\textsuperscript{16} Analogical reasoning gains a greater degree of importance for weak interactions in the early 1960s \cite{fraser_koberinski_2016}, and strong interactions in the early 1970s \cite{fraser_2018}, though here the analogies are formal analogies drawn between condensed matter physics and classical statistical mechanics, respectively.}

The work of Gell-Mann and Nishijima on strangeness takes the work of Pais and reformulates it to better fit the symmetry principles already in use. By generalizing the charge-independence equation to include the strangeness quantum number, associated production was explained by the fact the strong and electromagnetic interactions conserved strangeness, but the weak interaction did not. Here we see the seeds of treating the weak interaction as exceptional taking hold; though the symmetry principles were still thought to (approximately) apply phenomenologically, weak interactions were treated as perturbations to the stronger interactions and these perturbations appeared to violate the established strangeness symmetry. This is a step toward fitting associated production into the principles of the minimal theoretical framework. Quantum numbers and corresponding conservation laws were widely used to characterize HEP interactions, so the addition of a new quantum number for weak interactions fit the mold.

Related to this development—and therefore natural that both Gell-Mann and Nishijima would develop this view independently—was the separability and independence of the fundamental forces in HEP. To varying degrees of approximation we can consider only the strong interaction, or add the electromagnetic interaction, and finally consider the weak interaction as well. This hierarchy view allows one to make sense of the preservation of symmetry principles within the theoretical framework as useful heuristics, while simultaneously allowing one to acknowledge refinements to the framework required by new empirical findings regarding the weak interaction.

Though the community was now primed to accept the “strangeness” of the weak interaction, spacetime symmetry violations were still a further step. The standard view of spacetime symmetries is that they apply equally to all bodies and all dynamical interactions, and parity is naturally viewed as a mirror symmetry of spacetime.\footnote{\textsuperscript{17} There is some controversy over the correct way to view spacetime symmetries, however. \cite{earman_1989} represents a school of thought in which spacetime symmetries and dynamical symmetries are conceptually distinct, and it is a desideratum for a good physical theory that the two happen to coincide. \cite{brown_2005}, on the other hand, argues that spacetime symmetries simply encode the general dynamical symmetries we encounter in doing physics, and that it doesn’t even make sense to think of spacetime symmetries and dynamical symmetries not matching. In both cases, however, spacetime symmetries are privileged in that they apply to all dynamical models, not simply a subset.} Extension of the theoretical framework through relaxation of other symmetry principles was not sufficient to lead theorists to bet on parity nonconservation in weak interactions. The previous developments allowed physicists to “open their minds” to the possibility (cf. Section \ref{sec:2.3.2}), but overwhelming experimental evidence was required to solidify this final revision to the symmetry-based theoretical framework. The modifications of the symmetry framework were thus minimal extensions or relaxations, often argued for on the basis of analogy with other established models. Other proposed modifications—that have gone unmentioned in my brief historical account—were thrown out when shown to be inferior for interpreting experimental results. Thus the rich experimental background against which models were proposed was necessary for progressing the framework. As the electroweak model continued to develop, we see a continued trend of relaxing different symmetry requirements. With the violation of parity conservation, a discrete,
2.4. Analysis

global spacetime symmetry is broken. This trend is continued with the violation of combined charge-parity conjugation, and suspected time-reversal asymmetry as well. The introduction of spontaneous symmetry breaking into HEP led first to the concept of breaking continuous, global symmetries, but was further modified in light of the Goldstone theorem to breaking continuous, local symmetries. This final step was the key to unifying the electromagnetic and weak interactions.

I conclude this section by highlighting that this case study provides a clear example of the mutual influence that experiment, phenomenology, and a theoretical framework can have on each other. Experiment and the corresponding phenomenological models are designed and interpreted through the constraints of a theoretical framework, while at the same time highlighting limitations of the framework and suggesting ways in which it should be revised. The absence of dynamical model building here serves to highlight the unmediated influence that occurs between these components of HEP; in many other cases the direct links aren’t as clear due to the added complexity introduced when influence from a dynamical model is layered on top. Thus, we can see more clearly the utility of the disambiguation of the term “theory” into a theoretical framework and dynamical models constructed within the framework.

2.4.2 Surprising benefits of the absence of a dynamical model

The absence of a candidate dynamical model for the V-particles, beyond providing a clearer case study for the mutual interactions involved in theory construction, may also have helped theorists to remain open to the idea of parity nonconservation. As Franklin (1979, 1989) has emphasized, violation of parity conservation could have been discovered given the proper analysis of experiments carried out in 1928 and 1930, shortly after the establishment of the new quantum theory Cox et al. (1928), followed by Chase (1930), carried out experiments on double scattering of \( \beta \) radiation as a means of determining the spin and wave properties of electrons. The idea was that the first scattering would polarize the electrons, while the second would provide a uniform basis for analysis.

Given the framework provided by the new quantum theory, Mott (1929) recognized the importance of \( \beta \) double scattering as the most direct means for measuring the spin of a free electron. Basing his model on the Dirac equation for the electron, he calculated that a relativistic double scattering against heavy nuclei should lead to a higher rate of double scattering at 180° (i.e., directed back toward the source) than 0° (i.e., the opposite side of the scattering target). If parity conservation is violated in \( \beta \)-decay, one would detect instead an asymmetry of longitudinally polarized electrons as a discrepancy between 90° and 270°. Mott explicitly notes that his model does not predict a 90°-270° asymmetry.

The original experiment by Cox et al. showed an average asymmetry between 90° and 270°, though the results from each trial varied considerably, with the ratio of 90°/270° varying on either side of unity. Given that an asymmetry was not expected on theoretical grounds, these results were thought to hint at some unaccounted for systematic error in the experimental design. Chase then further refined the measurement techniques and experimental design, measuring the relative count of electrons scattered at 0°, 90°, 180°, and 270°. Chase found a statistically

\footnote{The summary provided here is based largely on Franklin’s (1989) work, though the analysis in terms of the presence of a well-established theoretical framework and dynamical model is my own.}
significant ratio of $90^\circ/270^\circ = 0.973 \pm 0.004$, and a ratio of $0^\circ/180^\circ = 0.987 \pm 0.003$. The latter result shows an asymmetry in the opposite direction of Mott’s prediction. The community did not seem to react strongly to the $90^\circ-270^\circ$ asymmetry, and focus was on resolving the disagreement between Mott’s prediction and experiment. Mott [1930, 1931] amended his calculations, and experimental checks beyond that of Chase showed a marked lack of consensus. The idea of parity nonconservation did not emerge from this era, and required the work of high-energy physicists in the fifties to become established.

I argue that two major theoretical barriers were present in the early 1930s that prevented further exploration into the possibility of $\beta$-decays violating parity conservation. First, the new quantum theory was recently established as the consensus theoretical framework for atomic physics, and the concepts and tools contained within it were the primary focus of theoretical work; this is exemplified by Mott’s papers working out the details of electron scattering as a means to measure spin. The detailed framework was much richer, and therefore more constraining, than the loose collection of symmetry principles underlying investigation into V-particles in the fifties. Further, the fact that symmetries were at the forefront of theorizing in the fifties may have made the possibility of violating those same symmetries more psychologically salient, if not an epistemically attractive idea. The more direct interplay between the framework symmetry principles and experiment also resulted in a more careful theoretical exploration of the possibility of parity violation. Here we see one way in which a discipline in its theoretical infancy is more open to fundamental shifts, because it is not constrained by an existing rich framework of concepts.

Second, and more obvious in this case, is the presence of a dynamical model guiding physicists’s expectations. The Mott prediction—based off of the Dirac equation—provides a detailed dynamical picture of the electron, and this picture was successful in predicting other phenomena related to the electron. Coupled with the inconclusive collection of experimental evidence, the natural move is to hold on to the otherwise successful model, and dismiss contrary evidence as erroneous. Though the usual idea is that there is a closer mutual interplay between dynamical models and a theoretical framework, the case of parity violation appears to be one in which the framework was more easily modified in the absence of a dynamical model. Experiment and phenomenology without dynamics led to the acceptance of parity violation, and the previous failure of discovery can be attributed in part to the contradictory mediating influence of Mott’s dynamical model. Though of course only a single example, I find this suggestive of the importance of mutually constraining and interrelating pieces of a scientific discipline. Counterfactually then, the presence during the fifties of an accepted dynamical model of V-particles, for which parity conservation was not explicitly violated, would have served to pull physicists away from the idea of weak parity nonconservation, ultimately slowing or even halting the evolution of the underlying theoretical framework.

### 2.4.3 Lessons for modern theoretical physics

One major goal of a philosophical investigation into the process of theory construction in post-World War II HEP is to extract generalizable lessons regarding successful theory construction to apply to current and future physics. I have already discussed how the interplay between experiment, phenomenology, and the theoretical framework mutually constrained each other’s evolution and led to the discovery of parity nonconservation for weak interactions. This was
an important step towards constructing a highly predictive dynamical model of the weak interaction, and eventually the standard model of particle physics, which describes the strong and electroweak forces within the framework of quantum field theory.

The situation in modern theoretical physics, though structured analogously, is epistemically very different from that of the fifties. As I have characterized the situation above, the fifties was a time of distrust in theoretical foundations, one of pragmatism and instrumentalism, as well as a time in which phenomenology dominated. These features can largely be attributed to the dominance of experiment, and the explosion of new phenomena to explain and understand. Most model building was phenomenological in nature, and dynamical details were only posited insofar as they could be quickly compared against experimental data. Dynamical models were then constructed with input from the well-tested phenomenological models, under the constraints of the theoretical framework. Theoretical speculation was held in check by the constant presence of new data, and so any proliferation of models was regularly culled.

Contrast this with the landscape in theoretical HEP today (cf. Dawid [2013, 2017]): new direct experimental data is getting more and more difficult to obtain, our current set of dynamical models has stood up to the vast majority of experimental tests19 and current issues are largely conceptual in nature. Model construction is focused more on dynamical models, and these often include predictions that are not empirically testable, either due to practical limitations (e.g., energy scales involved are many orders of magnitude higher than we can access at the Large Hadron Collider) or in principle limitations (e.g., bubble multiverses, extra compact dimensions, etc.). Direct experimental results, when they can be established at all, require massive amounts of money and time to acquire, and so far the lack of anomalies has somewhat disappointed theorists. After many years of work, the ATLAS and CMS experiments at CERN detected the Higgs boson, but so far its properties have been entirely consistent with the standard model, so no new physics is suggested by its presence [Aad et al., 2012]. Recently, LIGO has also confirmed the presence of gravitational waves, in highly sensitive experiments [Abbott et al., 2016], which serve largely as confirmation of the general relativistic framework upon which our astrophysical and cosmological theories are based.

Given the division of scientific disciplines into experiment, phenomenological models, dynamical models, theoretical frameworks, and mathematical tools, the case of weak physics in the fifties and current HEP can be seen as near polar opposites. In the former case, experiment and phenomenology dominated, while the theoretical framework was sparse and the dynamical models nonexistent. The latter case, by contrast, is one in which the theoretical framework is robust, and mathematical tools and dynamical models are constantly being developed and refined, though there is sparse experimental data and a corresponding absence of phenomenological models. Taking the place of phenomenology are toy models, constructed as simplified cases with which to test new framework principles or dynamical details.

Though a large group of physicists have been happy to work in this new era of HEP, and even claim that we have made significant progress in theory construction in such an evidence-sparse landscape [Susskind, 2007; Weinberg, 1989; Dawid, 2013], others have argued that ex-

19 There are some outstanding empirical anomalies, such as neutrino oscillations [Fukuda et al., 1998]. Evidence of flavour mixing from neutrinos emitted from the sun indicates that neutrinos must have some mass, though the standard model treats them as massless. This, however, is regarded as a minor anomaly, and one that can be fixed by finding a better mathematical representation for neutrino fields, rather than modifying the dynamics of the standard model.
periment and empirical confirmation are absolutely necessary for any real progress in physics
[Rovelli 2016] Ellis and Silk 2014[20] Though one example is certainly not definitive, I be-
lieve that the case of parity nonconservation supports the view that experiment and evidence are
necessary to drive progress in the construction of physical theories, in part because of the role
that evidence plays in weeding out potential candidates for new models and in rejecting core
framework principles[21] Though I have been speaking of the importance of the mutual
interaction and evolution of experiment, phenomenology, dynamical models, theoretical frameworks,
there is an obvious asymmetry between these interactions. In the case of parity nonconserva-
tion, dynamical models were still absent at this early stage, and the theoretical framework was
sparse. Experiment played a starring role in driving the construction and confirmation of phe-
nomenological models, though of course the framework of symmetry principles contributed
here as well. In this case, the eventual construction of dynamical models of weak interactions
and the evolution of the theoretical framework were instigated by experiment and phenomenol-
yogy. Dynamical models and a theoretical framework can in turn influence experimental design
and expectations for emerging phenomenology, but experimental data cannot be constructed
from the other pieces alone. Though the rest of the pieces needed for theory construction can
develop out of experimental data, the reverse is not true. Even the sorts of “purely theoretical”
progress that can be attained without empirical evidence—in the form of model building, foun-
dational analysis, and exploration of implications of dynamical models—are greatly aided by
constant checks against new experimental data.

To make this more clear, consider two of the three the major factors for non-empirical con-
firmation outlined in Dawid [2013]: the lack of alternatives (the No Alternatives Argument),
and the methodological continuity of new scientific disciplines (the Meta-Inductive Argument).
According to Dawid, the fact that alternatives to a new theoretical framework are difficult to
construct should increase one’s confidence that the new framework is on the right track. Addition-
ally, if it displays a high degree of methodological continuity with some other empirically
successful theoretical framework, then one should have a higher credence in the framework.
While I do not dispute that these should provide some degree of Bayesian confirmation (cf.
Dawid 2017), the case study of parity nonconservation shows that this can only provide a
weak degree of confirmation. In particular, the lack of alternatives to a model may be due to the
restrictions of enforcing a high degree of methodological continuity. Had physicists in the
fifties demanded strict methodological continuity, then a violation of a spacetime symmetry for
the weak interaction would not have been proposed, and a viable—in this case, the correct—

20 All parties agree that empirical evidence is central to progress in physics. The argument revolves around
whether theory construction can make progress in the absence of empirical evidence, and if so, the extent to
which progress can be made. In the case of Dawid, the question is whether theoretical progress without empirical
confirmation can still constitute confirmation to some extent. My point is that the theoretical side of theory
construction—the construction of new dynamical models, toy models, and framework principles—is at its best
when constantly checked against experiment. I describe below a suggestion for bringing the theoretical side of
HEP back in close contact with experiment. I am grateful to an anonymous referee for pressing me on this point.
21 My suspicion is that a thorough investigation into the history of physics further underscores this view. The
Newtonian framework of gravitation was constantly compared and refined against more and more precise ob-
servations (cf. Smith 2014), the quantum revolution was inspired by mounting anomalies within atomic and
subatomic experimental physics, among numerous other examples. Elsewhere (Chapter 3), I highlight the impor-
tance of experiment for further developments in constructing the standard model, including the confirmation of
QCD and precision tests of the anomalous magnetic moment of the electron.
alternative would not have been formed. While it is true that there is much methodological continuity in the above case, it is only in retrospect that we can distinguish the “helpful” continuity from the appropriate breaks in continuity. The retrospective reconstruction of continuity is of no help in constructing models beyond the standard model today; indeed, new alternative models might require strong methodological discontinuities that have yet to be conceived. Instead, this case study suggests that physicists should focus effort into finding new ways to bring empirical evidence to bear on HEP.

Given the state of HEP, how should we work to remedy the lack of new experimental data coming in? Focus could be shifted toward low-energy precision tests of the standard model, where the barriers to data production are much lower than direct accelerator tests. Precision testing of QED has been ongoing since the late 1940s, predominantly through measuring the anomalous magnetic moment of the electron [Aoyama et al., 2012, 2018, Hanneke et al., 2011], and precision tests of gravity have been carried out from the time of Newtonian astronomy to modern day cosmology, in the form of deviations to the cosmic microwave background and structure within the echos of gravitational waves [Troxel and Ishak, 2015]. Though it may be impossible to recreate the experimental environment of the 1950s, HEP would benefit greatly from an increased focus on precision tests. Like the parameterized post-Friedmann framework for cosmology (cf. Baker et al. [2013]), HEP could benefit from a formalism that would allow new dynamical models of high-energy phenomena to be cast in terms that would allow for direct comparison of small differences in precision measurements. If new proposed models entail that the anomalous magnetic moment of the electron, for example, deviates slightly from the prediction from the standard model, precision testing could be used to rule out models that differ in ways not measured by experiment, and lend confirmation to those that differ in the measured ways. This is one way to implement the lessons from this case study regarding the importance of constant checks from experiment on theoretical modeling.

Even when making supposed conservative extensions to empirically well-confirmed models or frameworks, experimental tests provide the only means of determining the success or failure of such extensions. When the standard model was being constructed, many supposedly conservative extensions of the existing body of knowledge ended up being discarded in favour of experimentally supported hypotheses. Gell-Mann’s model of a parity multiplet as an explanation of the $\theta-\tau$ puzzle was a more conservative extension of the known phenomenology of HEP than Lee and Yang’s proposal of parity nonconservation, but the evidence favoured the latter over the former. The same can be said of the Lee and Orear model. In general, continuity of methodology, or conservative extensions to accepted models, make great starting points for expanding the domain of a theory, or even for constructing new theories. It is in this sense that progress can be made on the side of theory, without input from experiment.22 But methodological continuity is not a reliable indicator of future empirical success, nor of some more robust notion of “getting at the reality of things.” Ultimately, physics is about the physical world, and the process of theory construction must include empirical evidence as the final arbiter. Dawid’s non-empirical confirmation is, despite his claims otherwise, different in kind from the empirical confirmation that has been instrumental to the success of physics up to and including the

---

22 This is aside from foundational work on the consistency of theoretical principles, or in understanding the full set of consequences of a dynamical model. These sorts of activities constitute progress, but I take it do not count as non-empirical confirmation of a given framework or model.
construction of the standard model.

2.5 Conclusions

The case of the discovery of parity nonconservation in weak interactions highlights the early development of a highly technical field in physics. Though not an instance of Kuhn’s (1962) pre-paradigm science—HEP was a natural breaking off from atomic physics based on quantum theory—and certainly not a period of scientific revolution, the fifties witnessed the birth of a new autonomous discipline. The discovery of parity nonconservation was a landmark step in the process of constructing the standard model, guided largely by experimental progress and construction of phenomenological models to conceptualize that progress. There was no consensus dynamical model for V-particle interaction, and little attempt to try to construct one, at least until after the discovery of parity nonconservation. Even without the presence of dynamical model construction, experiment and phenomenology led to a major revision of the general symmetry principles governing HEP. The discovery is one that was ultimately incorporated into the theoretical framework of quantum field theory, influencing the future electroweak dynamical model.

My account shows the general utility of thinking of science in terms of models, though a more fine-grained distinction was needed. Certainly for physics, and perhaps for all sciences, the division into experiment, phenomenological models, dynamical models, a theoretical framework, and mathematical tools seems to be a useful one, so long as one keeps in mind the constant mutual interactions between these pieces. In this paper we see a case of theory construction in which the construction of phenomenological models was driven by experiment—though input from the theoretical framework was still essential. In cases of mature sciences, this contrast also helps to highlight the construction of phenomenological models on the basis of approximations or modifications to a dynamical model. In that case, theory (in the standard sense of the word) drives modeling, and thus the search for phenomena. I highlight the utility of this division for further episodes in the construction and testing of the standard model in Chapter 3 and the division may prove useful for understanding theory construction in disciplines such as cosmology, condensed matter physics, and quantum information theory.

Though a modification to the coarser experiment-models-theory classification (cf. Figure 2.1), I think my account enriches and expands upon the view of science in which models play a central role. As this case study indicates, important principles that eventually come to define the theoretical framework can be suggested by models meant to link closely to experiment, and these principles can in turn drive the construction of more general dynamical models. It is this process of building models, extracting principles, and refining new models—all kept closely in check by constant experimental pressure—that forms the basis of theory construction, at least in HEP. It will be worthwhile to see how central of a role models play for theory construction elsewhere in science.
3.1 Introduction

In the mid-1960s, particle physics was in a bit of a theoretical bind. With larger and larger accelerators finding a seemingly endless supply of new resonances, the lack of a mathematically well-defined theoretical framework left theorists constructing relatively narrow phenomenological models in an attempt to systematize the new experimental findings (cf. Brown et al. [1989a]). This lack of theoretical foundation did not stop the development of more sophisticated phenomenological models—like the current algebra techniques of Gell-Mann—or the discovery of important properties of certain interactions—such as the nonconservation of parity in the weak interactions. However, in the mid-1950s quantum field theory—after a brief moment of glory in the wake of the confirmation of predictions from quantum electrodynamics (QED)—quickly fell out of favour as a framework for particle physics. By the mid-1960s, there was no longer any candidate replacement for quantum field theory either, with S-matrix theory having lost steam by this point as well (cf. Cushing [1990]).

Over the next ten years, quantum field theory made a substantial comeback. By 1975, the Glashow-Salam-Weinberg electroweak model and quantum chromodynamics (QCD) formed the foundation of the emerging standard model of particle physics. In the intervening period, a small group of theorists continued to work on the foundations of quantum field theory—with a focus on Yang-Mills theories in particular—and discovered that the representational capacity of the mathematical formalism was different than originally expected, and sufficiently rich to both account for known strong and weak phenomena. New dynamical models provided

---

1By “representational capacity” I mean the ability of a model to properly capture the relevant phenomena for its domain. In the case of QCD, for example, it is important that it exhibit asymptotic freedom and quark confinement in order to match the experimental results from deep inelastic scattering and the absence of observed free quarks, respectively. I do not mean representation in some deeper sense, implying that we should therefore be realists about the model. That, I think, requires an extra step, and is outside the scope of this paper.
concrete predictions, which were subsequently tested and largely confirmed. In this paper I will outline the major mathematical discoveries regarding Yang-Mills models of quantum field theory that led to their rapid reascendancy during this period. The majority of the work done involved constructing new mathematical techniques for exploring the space of possible models of Yang-Mills theory. This is a case where foundational analysis of the theoretical framework of a discipline led to the construction of new mathematical tools for constructing dynamical models and connecting the dynamics to experiment and phenomenology.\footnote{These terms will be defined in §3.1.1} The major techniques necessary for the acceptance of Yang-Mills models were: a full proof of the renormalizability of massless and massive Yang-Mills models, the use of renormalization group techniques to prove asymptotic freedom, and lattice quantum field theory as a tool for numerical computations in the strong coupling regime. Analysis of theory construction in the past does more than provide a philosophical depth of understanding to the history of physics; one potential contemporary use for such analysis is to learn generalizable epistemic principles that may serve as useful heuristics for theory construction today. A general lesson from this case study is that theories do not wear their consequences on their sleeve; it often takes a fair deal of analysis—including development of new mathematical tools—in order to figure out what a theory implies. These new discoveries are often more than deductive consequences of the original equations—they add to the content of the theoretical framework.

The remainder of this paper is organized as follows. In §3.1.1 I will introduce and explain the terminology I will use for discussing the various components of high-energy physics (HEP) in this case study. The terminology is not common, but I think it provides a more fine-grained distinction of the “pieces” of a scientific discipline, at least one that is highly mathematized. One goal of the present work is to provide a set of distinctions fine enough to account for the details of successful (and unsuccessful) theory construction, but general enough to apply outside of the domain of post-World War II particle physics.

Following this, in §3.2 I outline the status of Yang-Mills models circa 1965, and the epistemic context in which they were rejected. §3.3 provides a condensed history of the development of renormalization proofs (§3.3.1), renormalization group equations (§3.3.2), and lattice quantum field theory (§3.3.3). Finally, in §3.4 I discuss important lessons from this case study for theory construction in physics (§3.4.1), the use of analogies in physics (§3.4.2), and current discussions in the philosophy of quantum field theory (§3.4.3).

### 3.1.1 Terminology

My analysis below will split HEP into distinct components. The goal of this division is to clarify how these parts—often considered independently in philosophical analyses of scientific practice, or at least at a more coarse-grained level—work together in this successful era of theory construction. There are many ways to divide the practice of physics, and I do not claim this division is unique. It does, however, illuminate the high degree of collaboration and interaction between what is typically called theory and experiment in this era, and is sufficiently general to serve as a starting point for analysis elsewhere in physics.

By theoretical framework, I mean the network of principles and general mathematical constraints that serve as the common language of a research program. Currently, the theoretical
framework underlying HEP is quantum field theory. In the 1960s, however, there was distrust in quantum field theory as a general framework, and so the agreed upon theoretical framework was much more minimal. It consisted of a relativistic causal structure and conservation laws carried over from non-relativistic quantum theory. Newtonian classical mechanics is another example of a theoretical framework in physics, containing concepts such as force, inertia, mass, and so on. Within a theoretical framework, one constructs dynamical models to describe particular interactions. As a paradigmatic example, QED constitutes a dynamical model in HEP, as it describes the electromagnetic interaction between electrons, positrons, and photons. I use dynamical model in the way most would use the term “theory,” to disambiguate the particular models of interactions from the theoretical framework guiding and constraining their construction. I include the word “dynamical” to highlight the fact that in physics, these models are often encoded in some set of dynamical evolution equations. Typically, quantum field theory and QED would both be described as “theories,” though the latter is but an instance of the former. Given this distinction, it may be unclear what I mean by “theory construction.” For the purposes of this analysis, theory construction is the process by which a theoretical framework is established, and a consensus collection of important dynamical models emerges within that framework. For HEP, this is equivalent to the construction of the standard model and the working out of quantum field theory as its basis.

I further divide models into dynamical and phenomenological models, for a few reasons. First, the term “model” is ambiguous. In the first sense, we can understand the term as used in model theory. Then a model is simply an interpretation of a theory. Take, as an example, the theory of general relativity. Mathematically, any model of the theory is given by a specification of a tuple \( (M, g_{\mu \nu}, T_{\mu \nu}) \) including the manifold \( M \), a pseudo-Riemannian metric tensor \( g_{\mu \nu} \), and a stress energy tensor encoding the matter content, \( T_{\mu \nu} \). In terms of model theory, the class of these models satisfying the Einstein equations constitutes the theory of general relativity, and any particular specification is a model of the theory. Hence, an FLRW cosmological solution to the Einstein equations is a model of general relativity, though it forms the basis for the theory of cosmology. This is not usually the sense of the word “model” meant in the modeling literature in philosophy of science. This second meaning usually refers to partial constructions—with input from a the relevant theory, other auxiliary theories, and perhaps phenomenology—meant to more directly model some proper subsystem that falls under a particular theory. My terminology is distinct from these two senses, though there is overlap between my phenomenological models and partial constructions. Some model-theoretic models—like those in general relativity—would also be instances of dynamical models in my sense. However, as will become clear, dynamical models in HEP are not so rigorously or formally defined.

Experiments in high-energy physics produce data, and from these data phenomena are con-

---

3This is independent of the way in which dynamical models are interpreted. Dynamical models do not require a realist or mechanistic underlying interpretation. The dynamical models in the standard model—quantum chromodynamics and the electroweak model—are still the subject of heavy interpretive controversy, and many physicists involved in its construction take a clear instrumentalist view of the standard model. Nevertheless, the standard model is a clear case of a collection of dynamical models.

4I use the example of general relativity here because it fits particularly well with model theory. Quantum field theory, on the other hand, is nowhere near as clearly or rigorously defined, and specifying models of quantum field theory in this sense is extremely difficult.
structed. Phenomena are built from experimental data and expectations shaped by dynamical models or the theoretical framework. Mathematical methods and tools are used at every step of the process, in order to generate predictions, construct phenomena, and compare the two. As I will argue below, the mutual influence between experiment, mathematical tools, and a theoretical framework was essential to the construction and acceptance of QCD and the electroweak model. First I will provide a brief “state of the discipline” for HEP in this era.

### 3.2 Yang-Mills theories in the 1960s

Before examining the theoretical developments of the late 1960s and early 1970s, it is important to understand the epistemic situation regarding Yang-Mills theories in the mid-1960s. In this section, I outline the prevailing attitudes regarding gauge freedom (§3.2.1) and renormalization (§3.2.2), and discuss the reasons for rejecting quantum field theory as an adequate framework for HEP (§3.2.3).

#### 3.2.1 Gauge freedom

Yang and Mills [1954] created a construction procedure for local field theories, in analogy with QED. The idea is that one starts with a Lagrangian density describing a set of fields obeying some global internal symmetry. In QED this is the $U(1)$ phase symmetry, $\psi(x) \rightarrow \exp(i\alpha)\psi(x)$, $\alpha \in [0, 2\pi]$, though Yang and Mills generalize this to a transformation under a global $SU(2)$ isospin symmetry, $\psi(x) \rightarrow \exp(i\alpha_a t^a)\psi(x) = S\psi(x)$, where the $t^a$ are generators of the $SU(2)$ symmetry group, and $S = \exp(i\alpha_a t^a)$. The gauge principle is a way to elevate the global symmetry group to a local symmetry group, so that the phase parameters can vary with spatiotemporal coordinates: $\alpha_a \rightarrow \alpha_a(x)$. Standard Lagrangian densities involve derivatives of the fields, and these must be suitably modified to ensure that the new local symmetry leaves the Lagrangian invariant. One accomplishes this via the introduction of a covariant derivative

$$D_{\mu} = \partial_{\mu} - igB_{\mu},$$

such that $S[D_{\mu}\psi] = D_{\mu}\psi'$. This amounts to the introduction of a minimal coupling to a new vector field $B_{\mu} = B_{\mu a t^a}$, whose transformation properties are constrained to be

$$B_{\mu} \rightarrow S^{-1}B_{\mu}S - \frac{i}{g} (\partial_{\mu}S)S^{-1}. \tag{3.1}$$

In QED, this new vector field is naturally assigned to the photon, such that electromagnetic interactions occur between two charged particles and a photon. More generally, the new vector field will correspond to some force mediating boson, possibly possessing internal structure. The final ingredient to complete the new so-called “gauge theory” is to introduce a kinetic

---

5The data-phenomena distinction was first explicated by Bogen and Woodward [1988]. This distinction is of some importance to my view, as phenomenological models, though closely connected to experiment, can only be used to describe experimental phenomena, not the data. In what follows I largely gloss over this distinction, and refer to the comparison of experimental data with phenomenological models. These statements should be understood to be shorthand for the conversion of experimental data into phenomena, followed by the comparison with phenomenological models.

6The term “gauge” to describe the local symmetry operations comes from Weyl [1918], who sought to expand general relativity by allowing the metric field to include local variations to the length scale.
3.2. Yang-Mills theories in the 1960s

energy term for the new vector field $B_\mu$. This is given in analogy with electromagnetism as

$$\mathcal{L}_{\text{kin}} = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu}; \quad F_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu - [B_\mu, B_\nu].$$

(3.2)

Yang and Mills went through the explicit construction for a generalization of isospin invariance of the strong interaction, but the procedure is easily generalized to other internal symmetry groups. The key was generalizing the properties of the gauge field and its kinetic energy term from the Abelian $U(1)$ group of electromagnetism to general non-Abelian groups.

The classical form of the Yang-Mills Lagrangian does not contain an explicit mass term for the gauge boson, since this term would violate gauge invariance. However, at the time of publication, Yang and Mills were uncertain of the implications for boson mass in a fully renormalized quantized theory. Due to the difficulties of renormalization, they “[had] therefore not been able to conclude anything about the mass of the $[B]$ quantum” (p. 195). Yang and Mills argued that mass for the gauge boson was an important concern for the viability of the Yang-Mills theory as a quantum field theory of strong interactions. “[I]t is inconsistent with present experiments to have their mass less than that of the pions, because among other reasons they would then be created abundantly at high energies and the charged ones should live long enough to be seen” (p. 195).

Rather than awaiting a solution to renormalization questions for Yang-Mills theories—which wouldn’t come about until about 1970—many began adding mass terms in for gauge bosons by hand. Glashow’s (1961) early model of electroweak unification focused on a Yang-Mills type theory with an $SU(2) \times U(1)$ gauge group. He proposed an idea of partial Lagrangian symmetries, where all terms in the Lagrangian except mass terms obeyed the partial symmetry. Symmetry concerns were important for Glashow in order to understand partially conserved currents, such as strangeness and what he called “isobaric spin.” Gauge invariance and renormalizability weren’t brought into Glashow’s discussion. He had hoped that there would be some mechanism that would ensure renormalizability and generate mass, but this wasn’t discovered until ’t Hooft proved that massless Yang-Mills theories undergoing spontaneous symmetry breaking were renormalizable, discussed below in §3.3.1. Gauge invariance was explicitly violated by the mass terms, and Glashow’s argument involving partial symmetries did not demonstrate that gauge invariance would be restored upon quantization. Given the analogy with QED—where gauge freedom represented mathematical redundancy and all physical quantities had to be gauge invariant—it was hard to see how mass terms that vary with gauge could be considered physical. In the next section I will discuss the status of renormalization, and how this further influenced the rejection of gauge theories.

3.2.2 Renormalization

Renormalization was initially thought of as a means to “cure” relativistic electrodynamics of its divergences. One major problem with quantizing the electromagnetic interaction was that the formalism was plagued with divergent quantities. Arguments were given justifying the straightforward neglect of some of these infinite quantities, and subtraction procedures could

---

7Divergences were also rampant in the classical relativistic theory. Unlike for nonrelativistic models of the atom—for which quantization introduced stability—quantization did not solve the divergence problem. Quantum electromagnetism suffered from logarithmic divergences instead of the steeper linear divergences of the classical
be used to remove certain divergences. For example, the energy of the ground state of the Hamiltonian operator diverges in quantum field theory, but physicists at the time argued on physical grounds that only differences in energy from the ground state were measurable, and so one could effectively set the ground state energy to zero without changing physical predictions. In a certain sense, this “renormalizes” the vacuum energy for the theory. However, further divergences occur within the theory, leading to two distinct problems. First, since electrodynamics is inherently relativistic, a relativistically invariant renormalization procedure was needed. If a subtraction procedure could only be carried out under a special foliation of spacetime, it was unclear if the new renormalized theory would be Lorentz invariant in the right ways. A manifestly Lorentz invariant procedure would ensure that one did not rely on a privileged reference frame. Second, the renormalization procedure could only be effective if there were a finite—and preferably a small—number of divergences to be removed from the Lagrangian or Hamiltonian. These problems were solved in tandem, independently by Tomonaga [1946], Schwinger (1948a, 1949), and Feynman [1949]. Dyson (1949a, 1949b) proved that the three approaches were equivalent, and strengthened the proof of renormalizability for QED. A brief account of the solutions to these problems is outlined below.

Tomonaga [1946] developed methods for generating manifestly Lorentz-invariant generalizations of the canonical commutation relations of nonrelativistic quantum theory, and his later work uses this formalism to formulate a relativistic theory of the photon-electron interaction. His idea was to use a hybrid of the Heisenberg and Schrödinger representations for the field operators—now known as the interaction picture—to develop four-dimensional commutation relations. In a manifestly Lorentz-invariant formulation of field theory, invariant subtraction procedures for removing the infinite quantities were therefore easier to formulate. Schwinger (1948a, 1949) showed that the difficult divergent quantities in QED could be absorbed into the physical electric charge—due to vacuum polarization effects—and the electron mass—due to electron self-interaction effects. He further showed that there were no analogous divergences in the photon self-energy.

Though the divergences are still problematic, the fact that they can be limited to two sources is promising for the predictive power of QED. Schwinger supposed that some future theory would cure these divergences. In the absence of this successor theory, if the divergences are limited to a small number of physical quantities, then QED can still be successful so long as they are properly renormalized away, replacing the divergent mass and charge with those “true” empirically determined values.

Feynman [1948] then developed a physical motivation for modifying QED to include the

---

8In more modern terms, renormalization is a process which occurs after a regularization procedure. Regularization is a process by which divergent quantities are replaced by finite quantities depending on some arbitrary regularization parameter. Renormalization, on the other hand, is a process in which one takes the regularized theory and determines the “physical” form of the relevant parameters in the theory—usually masses and coupling constants—in such a way that they do not depend on the value of the regularization parameter. If this can be done, then the theory is renormalizable. A straightforward removal of the ground state energy value is therefore not a renormalization in this modern sense, but earlier views regarding “renormalization methods” were closer to “removing divergences from a theory.” In this older sense of the term, subtracting the ground state energy was a renormalization of the Hamiltonian.

9For a more comprehensive account of the history of the development of QED, see Schweber [1994].
3.2. Yang-Mills theories in the 1960s

relativistic cutoff, at least in cases involving virtual photons. In a following paper, Feynman [1949] introduced the now well-known path integral formalism for QED, and showed that a relativistic cutoff procedure would yield results equivalent to Schwinger’s Hamiltonian approach when the quantities calculated could be shown to be independent of the cutoff value (i.e., were well-defined in the limit taking the cutoff to infinity). The great advantage of Feynman’s path integral approach was that it handled collision interactions in a particularly simple way. Feynman diagrams could be used to visualize the processes, and simple rules for moving from the diagrams to scattering amplitudes for collision processes were almost algorithmic in their simplicity. Dyson [1949a] showed that the Tomonaga, Schwinger, and Feynman formalisms are all equivalent when calculations can be carried out in all three, and introduced a new renormalization procedure for Schwinger’s formalism. A few months later, Dyson [1949b] demonstrated that the formalisms of Feynman and Schwinger express the same underlying theory, insofar as their $S$ matrix elements agree (p. 1736).

More importantly, the paper showed that QED was renormalizable—that is, it yields finite $S$ matrix elements to all orders. The divergences are absorbed into the physical mass and charge parameters using a relativistically invariant cutoff procedure to separate out the divergent parts of a given $S$ matrix element, and absorb them into the physical charge or mass order-by-order. Though Dyson remarks on the utility of the $S$ matrix for prediction, he is puzzled by the structure of QED:

The surprising feature of the $S$ matrix theory, as outlined in this paper, is its success in avoiding difficulties. Starting from the methods of Tomonaga, Schwinger and Feynman, and using no new ideas or techniques, one arrives at an $S$ matrix from which the well-known divergences seem to have conspired to eliminate themselves. This automatic disappearance of divergences is an empirical fact, which must be given due weight in considering the future prospects of electrodynamics. Paradoxically opposed to the finiteness of the $S$ matrix is the second fact, that the whole theory is built upon a Hamiltonian formalism with an interaction-function which is infinite and therefore physically meaningless. (p. 1754)

In order to reconcile the seeming paradox, Dyson chose to interpret the cutoff-dependent QED as representing a physical limitation to the possible measurements in the theory. If we idealize to an observer limited in measurement precision only by the principles of relativity and quantum theory (i.e., $\hbar$ and $c$ as the fundamental limiting constants), then the original Hamiltonian picture would be accurate, and all physically meaningful quantities would diverge. The cutoff-dependent renormalized theory, in contrast, represents the fact that there are other physically relevant limitations on the precision of our measurements, including atomic scales and electromagnetic couplings, since our measuring equipment is composed of matter whose chief interactions are electromagnetic. Dyson shared Schwinger’s hope for a more complete theory as a successor to QED, in which the divergent Hamiltonian formalism will appear as a limiting

---

10 The resulting integrals, however, are not guaranteed to be amenable to analytic solution. It is still the case that diagrams with a higher order in the coupling constant $\alpha$ lead to integrals that are enormously difficult to compute. For example, one major empirical success of QED is the high degree of precision match between the experimentally determined anomalous magnetic moment of the electron and the value predicted as a perturbative expansion of $\alpha$ in QED. The current state of the art calculation cannot be carried out analytically, and provides a prediction to fifth order in $\alpha$, or tenth order in $e$ [Kinoshita, 2013].
case when infinitely precise measurements are allowed. So already we see a dissatisfaction with the cutoff-based renormalization procedure, even by those who invented it. Though a successful basis for calculation, the lack of principled basis for the cutoffs was thought to represent an incompleteness to QED.

Stueckelberg and Petermann [1953] showed that the different theories one arrives at when introducing different cutoff procedures are all related, and a theory with cutoff \( \Lambda_1 \) is related to one with cutoff \( \Lambda_2 \) by a group transformation. This is the origin of the term “renormalization group.” Gell-Mann and Low [1954] utilize their independent formulation of the renormalization group to examine the asymptotic behaviour of QED. They start by introducing a regularization parameter \( \lambda \) in such a way that the physical electric charge is a function of the bare charge and \( \lambda \), \( e = e_\lambda \). The new electric charge serves to interpolate between the long-distance physical charge and the bare charge, such that \( \lim(\lambda \to 0) e_\lambda = e \) and \( \lim(\lambda \to \infty) e_\lambda = e_\infty \). Gell-Mann and Low find that the family of parameters \( e_\lambda \) obey a differential equation of the form

\[
\lambda^2 \frac{d^2 e_\lambda}{d\lambda^2} = f(e_\lambda^2, m^2/\lambda^2),
\]

(3.3)

where the function \( f \) has a power series expansion, and the high energy values \( \lambda \gg m \) are approximated by the function \( f(e_\infty^2, 0) \). Using this approximation, they show that the bare charge \( e_\infty \) is either infinite, or a root of the equation \( f(e_\infty^2, 0) \); in both cases this is independent of the physical charge \( e \). This is the first time that renormalization is quantitatively connected to the behaviour of QED at high energy scales, and the indication here is that the electric charge displays asymptotic divergence.

It is worth noting that most of quantum field theory in the 1950s was largely done within the canonical Hamiltonian formalism. It turns out that the choice of formalism is very important for proving renormalizability, and Yang-Mills theories are easiest to renormalize either within the path integral formalism or directly at the level of Feynman diagrams.

By the mid-1960s, those thinking about renormalization in quantum field theory would have had in mind the subtraction procedures of Tomonaga, Schwinger, Feynman, and Dyson, and perhaps the scaling behaviour of QED as investigated by Gell-Mann and Low. The reservations regarding the physical meaning of subtraction renormalization would have also been prevalent, and this was one of the reasons for a general distrust of quantum field theories as a basis for the strong and weak interactions.

### 3.2.3 Rejection of local field theories

Despite the empirical success of QED, by the mid-1960s most physicists were convinced of the futility of using quantum field theory as a foundation for the rest of HEP. Though the Yang-Mills procedure provided a recipe for generating field theories involving gauge bosons and nearly arbitrary internal symmetry groups—a huge advantage given the prevalence of group theory and symmetry considerations in particle classification—it appeared to only be suitable for massless gauge bosons. Further, its status as a renormalizable theory was unknown. Renormalization methods were in their infancy in the mid-1960s, and most physicists were skeptical of renormalization even in the successful domain of QED.

Both the strong and weak nuclear interactions are short range interactions, and this means
that, if they are mediated by force-carrying bosons, then these bosons must be massive. As mentioned above, adding mass terms to a Yang-Mills Lagrangian spoils gauge invariance and adds further complications to the renormalizability question. The developments in renormalization techniques also seemed to suggest that the electric charge in QED really was divergent at high energies. Most physicists more-or-less shared the views of Dyson, Feynman, and Schwinger, and thought of QED as a useful tool for predictions, but unsuitable as a standalone fundamental theory. The simple cutoff and subtraction renormalization methods were viewed as a pragmatic way to conceal this defect of QED.

Other groups were more vocal about the rejection of quantum field theory as an inherently incoherent framework for particle physics. Landau and his collaborators in Russia were also investigating the structure of QED, and found an ultraviolet pole that has since come to be known as the Landau pole. The argument is similar to the scaling behaviour investigated by Gell-Mann and Low. They show that the coupling constant for theories like QED diverges as the cutoff momentum is taken to infinity, and they interpreted this (incorrectly, as it turns out) to be a general feature of quantum field theories. Unlike the more conservative distrust of QED expressed by the American physicists who created it, Landau et al. thought their result showed that no quantum field theory could be a candidate for a complete theory of fundamental particle interactions. They took this as reason to abandon the quantum field theory formalism altogether, and this view was influential for Chew and the development of the S-matrix program (cf. [Cushing, 1990, pp.129-30]).

The S-matrix theory emerged as a rival research program for hadronic physics in the late 1950s, and its early phenomenological successes were taken by Chew, Goldberger, and others as a further sign that quantum field theory had run its course. The S-matrix theory was inspired by the mathematical problems with quantum field theory as well as the surprising successes of dealing directly with constraints on the properties of a scattering matrix. Heisenberg started an S-matrix project in the 1940s, which was largely forgotten with the rise of QED. However, Chew, Gell-Mann, and Goldberger, initially inspired by Dyson’s S-matrix treatment of QED, used the S-matrix formalism as a self-consciously phenomenological framework with which to make predictions for hadronic physics (cf. [Cushing, 1990]). The principles of S-matrix theory became more complex, and by the mid-1960s even this rival to quantum field theory was treated with skepticism.

So, while the tools were in place that would eventually be used as the foundation of the standard model of particle physics, their properties were poorly understood. Over the next decade, however, new mathematical and phenomenological developments would lead to the rapid reemergence of Yang-Mills theories. These will be outlined in the next section.

---

11The heuristic argument for the connection between range of interaction and mass relies on a limit based on Heisenberg’s uncertainty principle. The energy-time version of the uncertainty relation is \( \Delta E \Delta t \geq \frac{1}{2\hbar} \). The timescale on which a boson can exist is related to the rest mass as \( t \approx \frac{\hbar}{2mc^2} \). So a particle traveling near the speed of light would have a range \( R \approx \frac{\hbar}{2mc} \). This argument is initially due to Wick [1938], explicating the Yukawa model of nuclear forces.

12Even today, the mass gap problem in Yang-Mills theories is a topic of interest among mathematically inclined physicists. The Clay Institute has offered up $1 million as a reward for solving the mass-gap problem as one of their seven millennium problems.
3.3 The reemergence of Yang-Mills theories

In the late-1960s, the constituent quark model was accepted as a useful fiction for classifying new hadronic particles, and Gell-Mann had begun a program of current algebra as an extension of the spirit of the S-matrix program (cf. Cao [2010]). Weinberg [1967] published a paper on a plausible field-theoretic model of electroweak unification, but this was largely ignored, for the reasons given in the previous section. There was a general distrust of quantum field theory, with the renormalizability of Yang-Mills type models still in major doubt—especially models with massive bosons. Phenomenology regarding the weak interaction was handled with a combination of current rules and the Fermi-model. Very quickly, however, experimental discoveries and mathematical advances led to the widespread acceptance of Yang-Mills theories as the foundation for both the strong and weak interactions, and quantum field theory regained its place at the foundations of particle physics.

In this section I will outline some of these developments, highlighting the ways in which the refinement and development of new mathematical tools led to discoveries about properties of Yang-Mills theories that fit well with the emerging experimental evidence, especially for the strong interaction. §3.3.1 discusses the developments in understanding renormalization, leading to a proof of such for both pure and spontaneously broken Yang-Mills theories. §3.3.2 outlines the development of renormalization group methods, used to analyze the asymptotic behaviour of quantum field theories underlying the strong interaction. The renormalization group methods, as emphasized by Wilson, are generic tools for handling systems undergoing behaviour for which a wide range of energy scales are not only all relevant, but highly nonseparable. Finally, §3.3.3 outlines the lattice field theory developments, which allow for analysis of the low-energy (and therefore strong coupling) regime of the strong interaction, and provide a plausibility argument for quark confinement.

3.3.1 Proof of renormalizability

While the majority of physicists working in America and Western Europe were focused on current algebra, a few physicists retained an interest in quantum field theories. Most notable for this paper was the work of Martinus Veltman, who worked on renormalizing Yang-Mills theories in relative isolation in the late 1960s, until his student Gerard ’t Hooft joined him in the early 1970s. The work of ’t Hooft ended up with a proof of the renormalizability of massless Yang-Mills theories, including ones with spontaneously broken symmetries.

Veltman’s work

By the mid-1960s, the dominant formalism in which to do quantum field theory, or to calculate S-matrix elements, was the canonical operator formalism. Within this formalism, one begins with a manifestly Poincaré-invariant Lagrangian density and uses it to construct a Hamiltonian operator in terms of the fields and their corresponding canonical field momenta. Canonical quantization proceeds by imposing (anti)commutation relations between the fields and their...
conjugate momenta (fermion) boson fields,
\[ [\phi(x, t), \pi(y, t)]_\pm = i\hbar \delta^3(x - y), \quad [\phi(x, t), \phi(y, t)]_\pm = [\pi(x, t), \pi(y, t)]_\pm = 0, \] (3.4)

where \( \phi \) is a field operator, \( \pi = \partial_t \phi \) is its canonical momentum, and \([\cdot, \cdot]_\pm\) represents the anticommutator (plus) and commutator (minus), respectively. The canonical operator formalism has the virtue of being highly similar to non-relativistic quantum mechanics, and guarantees the unitarity of the resulting S-matrix. However, in moving from working with a Lagrangian density to a Hamiltonian, Poincaré invariance becomes highly obscure. The canonical formalism also turns out to be ill-suited to scattering problems, where the path integral formalism excels.

In the path integral formalism, one starts with the same Lagrangian density for a classical field theory and inserts this into the classical action. A partition function—analogue to a partition function in statistical mechanics—is then constructed as a functional integral over field configurations in the action,
\[ Z[\phi] = \int \mathcal{D}\phi \exp \left[ \frac{i}{\hbar} \int d^4x \mathcal{L}(\phi(x), \partial_\mu \phi(x)) \right], \] (3.5)

where the term inside the exponential is \( i/\hbar \) times the classical action, and \( \mathcal{L} \) is the classical Lagrangian density. The functional integral “quantizes” the classical action by including non-extremized field configurations, i.e., “paths” for which \( \delta S[\mathcal{L}] \neq 0 \). The classical limit corresponds to focusing on the extremized action, where effectively only this one field configuration has a measurable contribution. In the path integral formalism, Poincaré invariance remains explicit, and scattering amplitudes are easily related to the functional integral. However, unlike the canonical operator formalism, unitarity is not guaranteed.

Finally, non-Abelian gauge theories complicate both formalisms significantly. Considering the fact that Yang-Mills type theories involve non-Abelian gauge fields, these complications are highly relevant for the epistemic environment of the mid-1960s. Many Russian physicists worked on modifying the path integral formalism to account for non-Abelian gauge freedom [Faddeev and Popov [1967]]. Veltman, on the other hand, had the insight to work directly with the Feynman rules for a theory. Rather than trying to prove renormalizability directly from the Lagrangian, or even from one of the two dominant formalisms, Veltman found it much easier to work directly with the diagrammatic representation of a theory [Veltman [1997]]. As he later recounts, however, there is an additional step when working directly with the Feynman diagrams:

a simple canonical transformation of fields may turn a perfectly reasonable set of Feynman rules into an unrenormalizable mess. Let me emphasize: unrenormalizable. An example of that is a gauge theory in the physical (or unitary) gauge. That is an unrenormalizable theory. Even if you subtract the known (that is, known from the renormalizable version) infinities, you do not wind up with a finite theory. Green’s functions have infinities all over the place. Only when you pass to the

\[ For \) quantum field theories involving fermions, the field theory has to be somewhat artificial in that Grassmann fields are used in place of classical real-valued fields. This is to ensure the appropriate anticommutation relations upon quantization. \]
S-matrix do these infinities go away, assuming that your regularization method is quite perfect. [Veltman, 1997, p.149]

One can arrive at many, prima facie distinct sets of Feynman rules from the same Lagrangian through simple canonical transformations prior to deriving the rules. And even worse, renormalizability will only be provable for a small subset—perhaps a singleton set—of Feynman rules. Most sets of Feynman rules will actually be provably nonrenormalizable! This was another epistemic hurdle that Veltman had to clear: a proof of renormalizability is an existence proof in this paradigm. One must show that there exists a set of Feynman rules for a theory that are renormalizable. Proofs of the nonrenormalizability of a particular set of Feynman rules, which physicists at the time thought amounted to proofs of the nonrenormalizability of the dynamical model as a whole, do not actually tell one anything about the renormalizability of the model as a whole. As a first step to the renormalizability of Yang-Mills type theories, Veltman recounts one of the main tricks that he employed, the Bell-Treiman transformation:

Introduce a free scalar field, not interacting with the vector bosons. Now replace the vector field with some combination of vector field and scalar field; at the same time add vertices such that the scalar field remains a free field. Surely then the physics remains the same. But the Feynman rules for the new theory were different: the propagator for the W-field was replaced by the propagator for the combination, and that combination could be chosen so as to lead to less divergent Feynman rules. The price to be paid were the new vertices, and the new particle entered as a ghost (remember that is was a free particle). That is how ghosts entered my scheme. I called the technique the Bell-Treiman transformation. Neither Bell nor Treiman was responsible. [Veltman, 1997, p.155]

I reemphasize that this was only a first step towards a proof of the renormalizability of Yang-Mills theories. In effect, the strategy Veltman took towards renormalization was as follows:

1. Understand the relationship between a Lagrangian and its possible sets of Feynman rules.

2. Use canonical transformations to manipulate the degrees of freedom such that gauge-varying terms end up as free fields.

3. Find the “correct” set of such Feynman rules for a given Lagrangian, with which to demonstrate renormalizability.

---

Note that canonical transformations are distinct from gauge transformations. A canonical transformation is a change of field variables, leading to (anti)commutation relations involving different field operators. Though this leads to the problem of unitarily inequivalent representations in quantum field theory, fields related by canonical transformations are generally thought to represent the same physical situation.

In the context of the above quote, Veltman was actually working on renormalizing explicitly massive Yang-Mills theory, which was ultimately a failure. What Veltman accomplished was to renormalize massive Yang-Mills theory up to one loop. This was an important feat, especially in light of the more modern view of the quantum field theories as being effective theories of matter; a theory that is renormalizable to one loop can be used to generate low-energy predictions on scattering. Fermi’s theory of the weak force was one-loop renormalizable, and Veltman further showed that Yang-Mills theory with an explicit mass term was equally useful. The steps Veltman took here can also apply to the massless Yang-Mills case, or the case where mass is obtained for the vector bosons through spontaneous symmetry breaking.
4. Tame divergences in groups, such that renormalizability is demonstrated in steps (i.e., prove renormalizability to one-loop, then prove general renormalizability).

These techniques were then coupled with a more powerful regularization scheme by Veltman’s student, Gerard ’t Hooft.

’t Hooft’s work

While Veltman was working on renormalizing Yang-Mills theories, a topic he “avoided dragging students into,” his student ’t Hooft expressed interest in the field. Veltman conceded, but the compromise was that, “[f]or at least part of their thesis work, I insisted on more phenomenologically oriented work” [Veltman, 1997, p.166].

’t Hooft was initially inspired by the sigma model, constructed by Gell-Mann and Lévy [1960], which treated pions as the fundamental fields, and included a scalar sigma field whose vacuum state symmetry was spontaneously broken, giving mass to the pions. ’t Hooft suspected that spontaneous symmetry breaking may be the appropriate mass generation mechanism to ensure the renormalizability of Yang-Mills theories. This suspicion led to a particular strategy for working on the problem of renormalization: start by proving the renormalizability of massless Yang-Mills theory, and then show that the mechanism for spontaneous symmetry breaking does not spoil renormalizability. This was a departure from Veltman’s work, as Veltman was focused on explicitly massive variants of Yang-Mills theory.

The major obstacle for proving the renormalizability of massless Yang-Mills theory in general turned out to be finding a gauge invariant regularization scheme. The problem with contemporary cutoff procedures, or lattice regularization, is that gauge invariance is spoiled, and gauge invariance is required in order for the S-matrix determined from the theory’s Feynman rules to be unitary. Further, explicitly gauge invariant regulators could be constructed, but their complexity past a one-loop correction was unwieldy, and their complexity obscured the unitarity and causality of the theory. What ’t Hooft discovered was that a new regularization trick would solve the problem in a way that allowed for easier order-by-order regularization, and which manifestly preserved the unitarity and causality of the theory: dimensional regularization.

’t Hooft started with something much like the Gell-Mann and Lévy sigma model, prior to spontaneous symmetry breaking. In effect, this is a Yang-Mills theory with an additional scalar field introduced. The first hint of the new regularization procedure came in an intermediate proof of the one-loop renormalizability of the sigma model. In order to provide appropriate counterterms for contributions to the Feynman diagrams internal to the loop, ’t Hooft moved to five dimensions during the regularization process. The one-loop renormalizability proof of the symmetric sigma model transferred over rather easily to the case in which the symmetry is spontaneously broken. “[T]he step remaining to be taken was a small one. As I knew from Cargeèse, the actual nature of the vacuum state has little effect upon renormalization counterterms” [’t Hooft, 1997, p. 191]. The transition was relatively easy because ’t Hooft realized that the regularization only needed to preserve gauge invariance of the total set of terms (plus counterterms) in the Lagrangian. A gauge fixing leading to spontaneous symmetry breaking
breaking amounts to the introduction of individual gauge-varying terms, though the model as a whole can remain gauge-invariant.

Figure 3.1: A single fermion loop contribution to the self-energy of the photon. From this diagram, one can easily understand the terminology of one-loop renormalization: terms like this must be properly renormalized to prove one-loop renormalizability of a dynamical model.

One can see the new regularization and renormalization processes as follows in an example from quantum electrodynamics (cf. 't Hooft [1971a]). Consider the contribution to the photon self-energy from a single fermion loop, as in Figure 3.1. At a first pass, the integral associated with this diagram diverges quadratically, but can be regularized by replacing the propagator \((m + iyk)^{-1}\) with a series of terms \(\sum_j c_j (m_j + iyk)^{-1}\) such that

\[
c_0 = 1, \quad m_0 = m, \quad \sum_j c_j = 0, \quad \sum_j c_j m_j = 0, \quad \sum_j c_j m_j^2 = 0,
\]

effectively adding a series of terms similar to the original propagator to the integral. For finite \(m_j\), the new integral converges, and can be solved explicitly with a change of variables. Then one takes the limit of \(m_j \to \infty, j \neq 0\) (keeping the \(c_j\) constant), which is necessary in order to neglect terms in the integral of order \(q^2/m_j^2\). The resulting expression, which I will refer to as \(\Pi_{\mu\nu}\), is rather complicated, but importantly the resulting term does not satisfy gauge invariance, and the renormalized photon mass term is not zero. The gauge condition is of the following form:

\[
q_\mu \Pi_{\mu\nu}(q) = 0,
\]

(3.6)

where \(q_\mu\) is the photon 4-momentum.

The offending portion of the expression is a rank one polynomial in \(q^2\), and can simply be cancelled by the introduction of a counterterm in the Lagrangian. As 't Hooft notes, “[t]hese terms [introduced to the Lagrangian] are local and have dimension less than or equal to four, so that causality and renormalizability [respectively,] are not destroyed” ['t Hooft, 1971a, p.178]. Though a simpler renormalization scheme exists for this term in QED, 't Hooft’s procedure turns out to be generalizable beyond this particular propagator and beyond QED to Yang-Mills theories; rather than imposing fully gauge-invariant constraints on the renormalization procedure, one can replace the propagator with one of revised form and add counterterms to the Lagrangian to cancel out the resulting terms that spoil gauge invariance. Arbitrary constants can then be fixed by imposing conditions of the form of equation (3.6), which 't Hooft calls generalized Ward identities. So long as the counterterms are local, and of dimension less than or equal to four, this procedure preserves explicit locality and renormalizability of the overall Lagrangian.

The advantage of this procedure for massless Yang-Mills theories is that the regularization is easier to define than procedures for which total gauge invariance is manifest, while the desirable properties of locality and causality are preserved. The trick in moving from QED to
massless Yang-Mills is that an additional term must be added to the denominator of the propagator to regulate the Lagrangian in a gauge invariant manner. ’t Hooft motivated this additional term (for one-loop) by assuming that the in-loop momenta actually have five components, and that the fifth component for all had a fixed magnitude $M$. This results in an effective replacement of the gauge boson and scalar particle propagators,

$$\frac{\delta_{ab}\delta_{\mu\nu}}{k^2} \to \frac{\delta_{ab}\delta_{\mu\nu}}{k^2 + M}$$  \hspace{1cm} (3.7)

$$\frac{\delta_{ab}}{k^2} \to \frac{\delta_{ab}}{k^2 + M}$$  \hspace{1cm} (3.8)

Internal to the loop, the gauge boson will have a fifth polarization direction, and this is treated as a new particle with its own Feynman rules. Imposing the generalized Ward identities on a theory like this ensures renormalizability, and the new $M$ dependence serves as an effective regulator. The case of spontaneous symmetry breaking relies on the same renormalization procedure, and ’t Hooft [1971b] showed this in a follow-up paper using the example of a partial breaking of SU(2) with a scalar isospin-1 boson.\textsuperscript{18}

Getting beyond one-loop renormalizability turned out to require a different approach to regularization, the hint of which is to be found in the above regularization procedure. Moving from four to five dimensions internal to one-loop was successful, but beyond one-loop renormalizability the trick was inadequate. ’t Hooft and Veltman [1972] generalized the procedure into a process now known as dimensional regularization.

The procedure suggested in ’t Hooft [1971a] was based on the observation that the Ward identities do hold irrespective of the dimension of the space involved. By introducing a fictitious fifth dimension and a very large fifth component of momentum inside the loop suitable gauge invariant regulator diagrams could be formulated. This procedure breaks down for diagrams containing two or more closed loops because then the “fifth” component of loop momentum may be distributed over the various internal lines. It was guessed that more dimensions would have to be introduced, and thus the idea of continuation in the number of dimensions suggested itself. This is the basic idea employed in this paper. (p. 190)

‘t Hooft and Veltman define an analytic continuation of S-matrix elements involving some number of loops in the complex $n$-plane, where positive integer values of $n$ correspond to a spacetime dimension of that integer. The continuation is defined such that elements involving finite diagrams at $n = 4$ agree with the conventional results. Divergences in the perturbative expansion can be shown to be poles in the complex plane at $n = 4$, and the generalized expressions are analytic in $n$. Then renormalization is just a subtraction of the poles, along with

\textsuperscript{18}In the course of proving renormalizability of Yang-Mills theories, this was a minor step. However, this second paper was hugely influential in the development of the Standard Model, as it proved at least one-loop renormalizability of the Glashow-Salam-Weinberg electroweak model. The model was therefore proven to be usable for first order predictions, which were later tested and confirmed the adequacy of the model. By this time, ’t Hooft and Veltman [1972] had demonstrated full renormalizability using the dimensional regularization procedure (see below), and the electroweak model was accepted as the appropriate description of the newly unified electromagnetic and weak interactions.
the proof of unitarity and causality of the elements for all \( n \) given by ’t Hooft and Veltman. This amounts to the claim that, at a given order in perturbation theory, the terms introduced to subtract off the pole are real and local. Unitarity uniquely determines the imaginary part of the Lagrangian from that of lower orders, so the new terms introduced at a higher order cannot contribute in unexpected ways. The requirement that the new terms are local is necessary to ensure causality.

Since the new dimensional regularization method could be applied equally well to massless Yang–Mills theory and Yang–Mills with massive gauge bosons from spontaneous symmetry breaking, ’t Hooft and Veltman proved the renormalizability of the electroweak model and what would become quantum chromodynamics. The missing ingredients for the latter—asymptotic freedom, confinement, and the presence of massless mediating bosons—were developed in parallel, and will be discussed in the next two subsections.

### 3.3.2 Development of the renormalization group

The work of Gell-Mann and Low [1954] on the scaling behaviour of electric charge in QED was not picked up in the particle physics community. By the mid-1960s, work on scaling was done primarily in the realm of condensed matter physics or classical statistical mechanics, where physicists’ efforts were focused on understanding critical behaviour in phase transitions.

Near the critical point of a large system, some quantity related to the system—called the order parameter—will abruptly change. At the critical point, the susceptibility of the order parameter usually diverges. The main theoretical example studied in the 1960s and 1970s was the Ising model of a ferromagnet. In the simplest version of this model, one has a cubic lattice of spin degrees of freedom, for which the Hamiltonian includes only nearest neighbour interactions between spins (causing their direction to correlate) and the influence of an external magnetic field. At some critical temperature, the spins will spontaneously align, causing a global magnetization of the material, i.e., \( M(T > T_C) = 0 \rightarrow M(T_C) ≠ 0 \). As the system approaches the critical temperature from above, the correlation length—parameterizing the average size of “blocks” of spins that are aligned—diverges. Kadanoff [1966] developed a technique to quantitatively analyze the spin-spin correlations for the Ising model near the critical temperature. In effect, he iterated a coarse-graining procedure by which one would take \( 2 \times 2 \times 2 \) blocks of spins, and treat these as contributing a single effective magnetic moment in the Hamiltonian. These blocks would then enter into the Hamiltonian with simple nearest-neighbour interactions, as in the original Ising model. The form of the Hamiltonian would remain invariant under the iterated coarse-graining, while certain physical parameters—the effective temperature and external magnetic field—might be distinct from the original model. Thus there was a set of relationships in which the temperature and external field at lattice spacing \( 2L \), \( T_{2L} \) and \( B_{2L} \) were given in terms of \( T_L \) and \( B_L \). Kadanoff found that, at the critical point, \( T \) and \( B \) would have fixed values, independent of the particular lattice spacing \( L \). If the lattice spacing is allowed to change continuously, rather than in integer multiples, one can derive differential equations governing the change of \( T \) and \( B \) in terms of lattice spacing, which

\[ \text{This is assuming an infinitely extended ferromagnet. In general, true critical phenomena in statistical mechanics require the thermodynamic limit be taken as an idealization: the volume and number of particles both go to infinity such that the density } N/V \text{ of particles remains constant.} \]
reach a fixed point as \( L \to \infty \). These are formally similar to the differential equation governing the scaling behaviour of the electric charge in Gell-Mann and Low’s analysis of quantum electrodynamics. Kadanoff was one of the first to systematically treat the scaling behaviour of classical systems in the language of quantum systems. He reformulated the Onsager solution to the 2D Ising model in terms of Green’s functions. This was a key step in the eventual use of scaling methods in quantum field theory, as Green’s functions are a particularly natural tool in quantum field theory.

Kenneth Wilson, a former graduate student of Gell-Mann’s, became interested in the scaling of physical constants in quantum field theory, and began with a close analysis of the Gell-Mann and Low paper. As mentioned above in §3.2.2, this work used a dummy momentum index to interpolate between the long-range, physically measured electric charge \( e \), and the short range, bare charge \( e_0 \). The renormalization group equation (3.3) governs the change of the electric charge \( e_\lambda \) with the change in momentum parameter \( \lambda \), and Gell-Mann and Low show that \( e_\lambda \) increases as \( \lambda \to \infty \), though it was unclear whether the bare charge reached a fixed point or diverged. This work inspired Wilson to consider the scaling behaviour of coupling “constants” in quantum field theory, particularly in the high-energy domain of candidate dynamical models of quantum field theory.

After studying the renormalization group treatment of Gell-Mann and Low, Wilson developed a set of rules for short distance expansion for products of operators in a quantum field theory, which behave singularly in the limit of point products. These were based on high-momentum Feynman diagrams, transformed into position space. Wilson [1965] used a similar method of analysis on the fixed source meson theory, where

I realized that the results I was getting became much clearer if I made a simplification of the fixed source model itself, in which the momentum space continuum was replaced by momentum slices. That is, I rubbed out all momenta except well separated slices, e.g., \( 1 \leq |k| \leq 2, \Lambda \leq |k| \leq 2\Lambda, \Lambda^2 \leq |k| \leq 2\Lambda^2, \Lambda^n \leq |k| \leq 2\Lambda^n \), etc. with \( \Lambda \) a large number.

This model could be solved by a perturbation theory very different from the methods previously used in field theory. The energy scales for each slice were very different, namely of order \( \Lambda^n \) for the \( n \)th slice. Hence the natural procedure was to treat the Hamiltonian for the largest momentum slice as the unperturbed Hamiltonian, and the terms for all lesser slices as the perturbation. In each slice the Hamiltonian contained both a free meson energy term and an interaction term, so this new perturbation method was neither a weak coupling nor a strong coupling perturbation. [Wilson [1982] pp.115-6]

The usual perturbation methods used in quantum field theory involve expanding in a power series about the coupling constant, while here one treats lower energy effects as perturbations to a high energy Hamiltonian. Further, Wilson generated an iterative renormalization procedure for the Hamiltonian of free meson theory. Starting with \( n \) momentum slices and using the ground state for the unperturbed \( n \)th slice Hamiltonian, the next term was an effective Hamiltonian for the remaining \( n - 1 \) slices, with the coupling constant renormalized. This was the first

---

20 This work on operator product expansions would eventually be published as Wilson [1969], after Wilson resolved some issues with the expected behaviour of expansion coefficients in the strong coupling domain.
practical use Wilson found for the renormalization group formalism. In this way, one could isolate momentum scales from the theory, solve them, and iterate to the next momentum stage.

Wilson’s focus on the Hamiltonian in the fixed source meson theory transferred over to the Kadanoff picture of the Ising model, and Wilson [1971b] ended up reformulating and generalizing the Kadanoff picture of the Ising model near a critical point. As with much of his work on the renormalization group and critical phenomena after this point, Wilson refitted the Kadanoff picture to allow for continuous scaling, and investigated the asymptotic behaviour of differential equations relating the dependence of temperature $T_L$ and magnetic field $B_L$ on the scaling length $L$. These take the following form,

$$\frac{dT_L}{dL} = L^{-1} u(T_L, B_L^2)$$

$$\frac{dB_L}{dL} = L^{-1} B_L v(T_L, B_L^2),$$

with the assumption that $u$ and $v$ were analytic at the critical temperature for a phase transition. Wilson was able to rederive the Widom-Kadanoff scaling laws for the Ising model from a more general, parallel set of assumptions, beginning with these differential equations.

In analogy with the analysis of Gell-Mann and Low, Wilson found the differential form of the renormalization group relations to be most useful. One major benefit of thinking of the renormalization group in terms of differential equations is that qualitative analyses of scaling behaviour are easily obtainable. In the case of Wilson’s analysis of the Kadanoff model, the particular functions $u$ and $v$ that one derives are actually singular at the critical temperature. Wilson argues that the physical picture underlying the Kadanoff block spin formalism implies that the differential equations should not be singular:

In the block picture one should be able to construct $[T_L]$ and $[B_L]$ just by adding up interactions of individual spins within a block or across the boundary between two blocks; it is hard to see how this simple addition over a finite region can lead to singular expressions for $[T_L]$ and $[B_L]$, as a function of $[T]$ and $[B]$, if $L$ is fixed... in the spirit of the Kadanoff approach one does not try to get specific forms for $[u(T, B^2)]$ and $[v(T, B^2)]$ because this would require that one take literally the idea that all spins within a block act as a unit. [Wilson, 1971b, p. 3177]

In part II, a replacement form is found for the equations $u$ and $v$ to remove the singularities in the differential equation, but part I is dedicated to the qualitative features of scaling that can be determined from placing general constraints on the form of $u$ and $v$. The main point is that the analytic differential equation is capable of recovering critical point singularities—most notably the Widom-Kadanoff scaling law—as asymptotic divergences at $L = \infty$.

Importantly for the reemergence of Yang-Mills theory, Wilson [1971a] also applied the renormalization group analysis to plausible candidates for a theory of the strong interactions.

---

21Wilson uses an analogy with a simple classical mechanical system—a ball at the crest of a hill—to argue that the singularities inherent in a particular form of differential equation may be an artifact of the variables chosen to represent the equation. This is also familiar in the context of solutions to Einstein’s field equations, where coordinate singularities can arise, and an appropriate transformation of coordinates must be done to remove the false singularity.
3.3. The reemergence of Yang-Mills theories

In this paper renormalization group methods are applied to the strong coupling constant, to determine how it would scale with momentum and the properties that would follow. Experiment and accompanying phenomenological models from the late 1960s indicated that scale invariance was an approximate symmetry of deep inelastic scattering—where the strong interaction was thought to play an important role—and the logarithmic corrections to scale invariance were well known (cf. [Cao, 2010 Ch.6]). Wilson demonstrated that broken scale invariance would follow if the strong interaction was described by a renormalizable theory, and if the renormalization group flow asymptotically approached a fixed point at high energy [Wilson, 1971a, Sec. III.D]. The Gell-Mann and Low analysis also indicated that the electromagnetic coupling strength would increase in strength at high energies, making it plausible that at some energy scale $\Lambda$ the strength of the strong and electromagnetic forces would become equivalent. Wilson showed that, if this were true, the fixed point for strong interactions would additionally be infrared stable, allowing for an iterative “bootstrap” procedure to determine renormalized coupling constants as a function of energy scale (Sec. III.F). The importance of these results is that they provided some concrete connection between the Lagrangians describing models of the strong interaction—for which the standard perturbative procedures employed in QED would not work—and the phenomenology of strong interactions.

Conjectured in the conclusion of this paper, and confirmed by later developments, was the idea that renormalization group methods would be essential for the solution of strongly interacting relativistic fields. The benefit of the renormalization group equations is that they allow one to determine the dynamics at a particular energy scale, under the assumption that dynamics at higher energy scales have already been solved. “In order to solve the infinite number of energy scales which exceed energies of practical interest, one must iterate the renormalization-group transformation an infinite number of times, thus making asymptotic behavior of this transformation of crucial importance in solving the model” [Wilson, 1971a, p.1842].

One final scaling property of the gluon model for strong interactions—later christened QCD—was discovered by 't Hooft and later Gross and Wilczek [1973] and Politzer [1973]: asymptotic freedom. The scaling of a coupling constant in massless Yang-Mills theory is parametrized by a function $\beta(g)$ that depends on the coupling constant. Analysis of $\beta(g)$ indicated that the coupling constant would decrease with increasing energy scales to the point where $g \rightarrow 0$ as $\Lambda \rightarrow \infty$, and so a massless Yang-Mills theory is asymptotically free. For an $SU(3)$ Yang-Mills theory, up to 16 fundamental fermion fields could be introduced without spoiling asymptotic freedom. Asymptotic freedom is important because at very high energies, the strong coupling constant would become small, and so perturbative methods could be used to extract predictions from these models. Earlier testing of scaling in the deep inelastic scattering regime [Bloom et al., 1969] [Breidenbach et al., 1969]—where asymptotic freedom becomes relevant—showed that hadrons behaved as composites of point particles, and vindicated the predictions generated from the QCD Lagrangian.

\[ \text{\footnotesize \textsuperscript{22}}} \text{'t Hooft presented the final equation for the scaling of the beta function at a conference in 1972, but never published the results. Gross, Wilczek, and Politzer would eventually win the 2004 Nobel prize for the theoretical prediction of asymptotic freedom.} \]
3.3.3 Lattice quantum field theory

A final piece of the puzzle for connecting the QCD model with strong interaction phenomenology was required. In QCD quarks are the fundamental fermion fields, with gluons being the gauge bosons mediating the interaction between quarks. However, no free quarks or gluons had ever been detected experimentally; instead, only baryons and mesons—composed of 3 quarks or a quark-antiquark pair according to QCD, respectively—had been observed interacting via the strong force. If QCD was to be the foundational model for the dynamics of the strong interaction, it had to give rise to quark confinement in a natural way.

One problem in determining whether or not quark confinement would arise in QCD is the inverse of the asymptotic freedom property: at low energies—where baryons and mesons exist, and at which collider experiments are conducted—the strong coupling constant is large. The majority of the methods for generating dynamical predictions from quantum field theories depended on a perturbative expansion in powers of the coupling constant, and it is precisely in the most easily empirically accessible regime that these break down for QCD.

Wilson’s work on lattice QCD (1974)—inspired by, but not directly related to his work on the renormalization group analysis—provided a plausible dynamical account of quark confinement, and was convincing enough to remove this worry.

The inspiration for pursuing lattice quantum field theory began from Wilson’s earlier work on the fixed source meson model:

However, I learned from this picture of the Hamiltonian that the Hamiltonian would have to be cutoff at some large but finite value of momentum $k$ in order to make any sense out of it, and that once it was cutoff, I basically had a lattice theory to deal with, the lattice corresponding roughly to the position space blocks for the largest momentum scale. More precisely, the sensible procedure for defining the lattice theory was to define phase space cells covering all of the cutoff momentum space, in which case there would be a single set of position space blocks, which in turn defined a position space lattice on which the field $\phi$ would be defined. I saw from this that to understand quantum field theories I would have to understand quantum field theories on a lattice. [Wilson 1982, p. 117]

In order to place a quantum field theory on a lattice, one must first move from a Minkowski spacetime metric to a Euclidean spatial metric. High-momentum cutoffs in the original quantum field theory then correspond to the spatial separation of lattice sites. The biggest “trick” for developing a quantum field theory on a lattice was to make the lattice theory explicitly gauge invariant. This is important because the renormalization procedure used to make lattice quantities finite would spoil the restoration of gauge invariance afterward, so the quantities must be gauge invariant before renormalization.

In the lattice formulation of non-Abelian Yang-Mills theory coupled to fermions—of which lattice QCD is a particular model—confinement is demonstrated in the strong coupling ($g \to \infty$) limit. First, one starts with a position space path integral formalism. Associated with each classical path is a contribution to the path integral which is weighted by a gauge field term.

---

\(^{23}\)This change is central to the physical disanalogies between models in quantum field theory and condensed matter physics (cf. Fraser and Koberinski 2016). Causal and modal structures change dramatically when time is converted to a spatial dimension.
with an overall weighting for all paths that includes the free gauge field action. The total weight factor is averaged over all quark paths and all gauge field values. In the strong coupling limit Wilson showed that the lattice version of the gauge field average for a closed path is exponentially suppressed by its area, meaning that the dominant contribution to the propagator is the path with least area for a given perimeter. This leads to a suppression of large area quark-antiquark loops (there is a term proportional to the gauge field average in the quark propagator) in position space; large area loops are necessary for the separation between bound quarks to be sufficient for individual detection as effectively free particles.

Though not a conclusive proof of quark confinement, Wilson provided a compelling reason to think that quarks would be confined in QCD using lattice methods. The further advantage of lattice quantum field theory is that it allows for small QCD systems to be solved using numerical methods on a computer. Even today, most results in QCD that do not depend on the use of perturbative methods in the high-energy, weak coupling limit still rely on a lattice formulation of the model.

With this last epistemic barrier removed for QCD, the path was clear for the full acceptance of the Glashow–Salam–Weinberg electroweak model and QCD as the basis for describing the fundamental forces governing the interactions in HEP.

3.4 Analysis

The rapid shift in perception of quantum field theory—from mathematically inconsistent and useless for strong interactions, to forming the basis of the standard model of particle physics—between the mid-1960s and mid-1970s has been outlined above. It is largely a story of a deepening understanding of the properties of quantum field theories (particularly of the Yang-Mills type) and their suitability as a theoretical framework for recovering particle physics phenomenology. This is a case study in developing mathematical tools to understand the properties of a theoretical framework, and then using the newly understood properties as a basis for justifying the construction of dynamical models of both the strong and (electro)weak interactions.

In this section, I will outline what I take to be the important lessons this era provides about theory construction in physics. I will start from specific epistemic and methodological breakthroughs for particle physics (§3.4.1 and §3.4.2), and finally address the relevance of looking at the process of theory construction for modern views on effective field theories (§3.4.3).

3.4.1 Implications for theory construction

The reemergence of Yang-Mills theory as a theoretical foundation for particle physics—forming the two dynamical models at the heart of the standard model—highlights the importance of developing mathematical techniques for investigating the properties of a theoretical framework. In the mid-1960s theorists occupied a limited epistemic perspective on quantum field theories, particularly Yang-Mills theories. Until it was established that renormalizability was an existence claim for a given model of quantum field theory (i.e., one needed to find the appropriate set of Feynman rules), physicists strongly suspected that Yang-Mills theories were not renormalizable, and the possibility of a mass generation mechanism for Yang-Mills gauge bosons was not conceived. Further, the strong interaction was not well understood, and it was unclear
if baryons and mesons were supposed to be the fundamental fields, or if the non-relativistic quark model could be turned into a fully relativistic field model. All of these misunderstandings are related to the limited understanding of the quantum field theoretical framework and its representational capacity.

By better understanding the representational capacity of the formal system, physicists were able to enrich their theoretical framework—by expanding the space of possible Lagrangians consistent with the core concepts of relativistic quantum field theory—and more easily construct candidate dynamical models—such as QCD and the electroweak model. Further, one could prove the compatibility of particular dynamical models with new experimental discoveries by employing the new mathematical tools. These tools were used to understand the representational capacities of the theoretical framework (quantum field theory). ’t Hooft and Veltman’s dimensional regularization scheme was used to show that Yang-Mills type quantum field theories were in fact renormalizable, which meant that the class of Yang-Mills models could be candidates for consistent dynamical models of strong and weak interactions. In the case of weak interactions, the Weinberg-Salam model of electroweak unification had already been worked out, so focus shifted to finding decisive empirical signatures—like weak neutral currents. For the strong force, more work was needed. As Cao [2010] has already detailed, there were multiple candidate relativistic quark-gluon models, but it was unknown how they connected to the weak binding constituent quark model—useful for group-theoretic classification of hadrons—and the relativistic current algebra.

Renormalization group analysis of the beta function for non-Abelian Yang-Mills theories showed that pure Yang-Mills theories were asymptotically free, though the addition of fermions would introduce terms in the beta function that increased with increasing energy. It turns out that the larger the Yang-Mills internal symmetry group, the more fermion generations could be accommodated without spoiling asymptotic freedom. This helped to rule out the simpler quark-gluon models, in which only a single gluon field existed (internal $U(1)$ symmetry), and allowed QCD (internal $SU(3)$ symmetry) to explain the success of the simple constituent quark model, in which hadrons behaved as composites of point quarks. In the deep inelastic regime, scattering results indicated that constituent quarks were effectively free, vindicating the high energy freedom of QCD. The $SU(3)$ symmetry group matched the $SU(3) \times SU(3)$ symmetry of current algebra as well.

The last question to be answered for QCD was whether the coupling of quarks and gluons prohibited free quarks at low energies. Wilson’s lattice methods made a convincing argument for low energy quark confinement, and QCD was accepted as the consensus dynamical model of strong interactions.

Importantly, the sigma and $U(1)$ gluon models were not asymptotically free, and previously proposed phenomenological models (constituent quark models, current algebra) could be connected to QCD in appropriate limiting cases. Without the new mathematical tools, these properties of candidate strong interaction models could not have been elucidated.

All three major developments outlined in §3.3—renormalizability, the renormalization group, and lattice quantum field theory—were tools with which to analyze the framework of quantum field theory. Renormalizability was demonstrated for large classes of quantum field theories, with the relevant models being the class of massless (or spontaneously broken) Yang-Mills theories. The knowledge that this class of models was renormalizable led to further investigations of candidate dynamical models—the electroweak model and QCD were the prime targets. The
renormalization group analysis of QCD was essential to its acceptance, since asymptotic freedom made the substructure of hadrons accessible via deep inelastic scattering. Crucially, the number of fundamental fermion fields was limited in order to ensure asymptotic freedom, and this theoretical limit was consistent with the known number of fermions. The close interplay of mathematical investigation into the theoretical framework with experimental tests allowed for the emergence of consensus dynamical models within a matter of a few years.

The importance of better understanding the framework is highlighted well in Veltman’s reflections on the importance of renormalizability proofs:

Personally I have always felt that the proof was much more important than the actual construction of a model, the Standard Model. I felt that, once you knew the recipe, the road to a realistic description of Nature would be a matter of time and experiment... The proof of renormalizability also provided detailed technical methods such as, for example, suitable regularization methods, next to indispensable for any practical application of the theory. In longer perspective, the developments in supersymmetry and supergravity have been stimulated and enhanced by the renewed respectability of renormalizable field theory. [Veltman, 1997, p.145]

Though it may seem obvious in retrospect, one lesson to keep in mind for contemporary theory construction in physics is that it takes time and innovation to discover the consequences and representation capacities of a theoretical framework. Much of the hard work in theory construction comes when trying to understand the consequences and representational capacities of a theoretical framework. In the case of particle physics, the theoretical framework of quantum field theory—mathematized in terms of Lagrangians, action functionals, canonical quantization procedures, Green’s functions, scattering amplitudes, etc.—required a broader set of mathematical tools beyond perturbative expansions in coupling, and a more systematic treatment of renormalization. It turned out that the successful application of perturbative expansions and non-systematic treatments of renormalization were permissible for QED due to some peculiar features of the electromagnetic interaction: first, the dimensionless coupling constant for QED is relatively small ($\alpha \approx 1/137$), making a perturbation about the coupling constant accurate after only a few orders; second, QED is a special case of an Abelian gauge theory, and did not require more sophisticated regularization and renormalization techniques. The failure of techniques that worked for the comparatively simple QED did not mean that quantum field theory in general would fail, and it took the work of the few physicists who remained interested in quantum field theory to prove this.

One important feature of quantum field theory in the mid-1960s was that there was already a useful dynamical model in existence: QED. Though it seemed to be the case that QED was not easily extended to accommodate the strong or weak interactions—for example, the renormalization procedures were not easily generalized—it provided a clear example showing that quantum field theory was at least suitable for a limited domain of particle physics. In the context of quantum field theory, we can now see why straightforward extensions of QED would have to be unsuccessful. One needed to move from Abelian gauge theory to non-Abelian gauge theory, and the tools required to handle the latter turned out to be much more complicated.
3.4.2 The success of formal analogies with statistical mechanics

An understanding of the full representational capacity of Yang-Mills theories required the development of novel mathematical techniques to explore their renormalizability, scaling behaviour, numerical solutions, and mass generation mechanisms. All of these methods were either developed within or originated from condensed matter physics, and were carried (back) over to particle physics due to the formal similarities between the structure of models in the two disciplines.

As mentioned in §3.3.2 and §3.3.3, much of Wilson’s work regarding the renormalization group and lattice QCD was inspired by strong formal analogies with statistical mechanical systems. Though the initial application of renormalization group equations to particle physics was through Gell-Mann and Low’s (1954) treatment of scaling in QED, condensed matter physicists (like Kadanoff) did much of the work on the renormalization group in the 1960s, in the context of simple statistical mechanical models—both quantum and classical. Many of Wilson’s papers on the renormalization group deal variously with particle physics and statistical physics. The three landmark papers published in 1971 deal with the applicability of renormalization group methods to the strong interaction (1971a), and to the Kadanoff scaling picture of the Ising model (1971b;1971c). In the more systematic paper on renormalization group methods, Wilson and Kogut [1974], both statistical mechanical and quantum field theoretic systems are treated. A few sections at the end of their paper—particularly Section 10—outline the formal connection between statistical mechanics and quantum field theory. In particular, the Feynman diagrams for a \( \phi^4 \) lattice quantum field theory—when converted to a Euclidean metric—are identical to the spin correlation functions for the generalized Ising model.

Wilson and Kogut [1974, Sec 12.2] present the details of renormalization group flow and its applicability to quantum field theory in the case of a four-dimensional \( \phi^4 \) model. One starts with the model defined on a lattice, which is implied by the introduction of a momentum cutoff scale. First, one must regularize the model, and then introduce a nonzero coupling constant at infinite “correlation length”—the analogue of which in quantum field theory is a continuum relativistic model. The reasons that a tight formal correspondence can be set up between (classical, in this case) statistical mechanics and quantum field theory are complex, but the application of the renormalization group analysis to both is no accident. In later works, Wilson explicitly emphasizes the expected generality of renormalization group methods as applicable to numerical solutions of physical situations. Unlike the simple problems with a few degrees of freedom that are amenable to simple numerical approximation schemes,

There is a fourth class of problems which until very recently lacked any convincing numerical approach. This fourth class is a subclass of problems involving a large or infinite number of degrees of freedom. The special feature of this subclass is the problem of “renormalization.” Originally, renormalization was the procedure for removing the divergences of quantum electrodynamics and was applied to the Feynman graph expansion. The difficulties of renormalization prevent one from formulating even a Monte Carlo method for quantum electrodynamics. Similar

\(^{24}\)Mass generation—in the form of spontaneous symmetry breaking—was not discussed in this paper. For a detailed analysis of the formal analogy between spontaneous symmetry breaking in the Higgs mechanism and in superconductivity, see Fraser and Koberinski [2016].
difficulties show up in a number of problems scattered throughout physics (and chemistry, too). These problems include: turbulence (a problem in classical hydrodynamics), critical phenomena (statistical mechanics), dilute magnetic alloys, known as the Kondo problem (solid state physics), the molecular bond for large molecules (chemistry), in addition to all of quantum field theory. In this paper the problem of renormalization will be shown to be the problem of many length or energy scales. [Wilson] 1975, p.171

It is clear from this quote that Wilson, at least, saw the renormalization group as a mathematical technique for systematically treating problems where energy scales are not cleanly separable. So the treatment of renormalization, though initially developed in order to make QED a predictive model, is really a quite general phenomena in physics, and should perhaps be likened more to techniques such as Taylor expansions and Sturm-Liouville theory. These are techniques for finding solutions to formal problems, that have little directly to do with the physical situation at hand.

So if the renormalization group methods are meant to be so generally applicable, why was the analogy with statistical mechanics important for their development beyond QED in particle physics? It’s because the trick with the renormalization group is to find a way to express the problem that makes it clear how to use the techniques.

[I]t is rather difficult to formulate renormalization group methods for new problems; in fact, the renormalization group approach generally seems as hopeless as any other approach until someone succeeds in solving the problem by the renormalization group approach. Where the renormalization group approach has been successful a lot of ingenuity has been required: one cannot write a renormalization group cookbook. [Wilson] 1975, p.185.

After Wilson’s success in formulating the Ising model in such a way that a renormalization group analysis could be performed, he looked for ways to transform quantum field theory problems into the same form as the classical statistical mechanical Ising model. As Fraser [2018] has emphasized, the analogy between the classical Ising model and four-dimensional $\phi^4$ model is facilitated by transforming the $\phi^4$ model into a Euclidean metric by Wick rotating the temporal part of the metric $t \rightarrow -it$, and discretizing the spacetime by introducing a minimum length scale. Then, one must establish a formal correspondence between the spin-spin correlation function $\Gamma_{m,n}$ in the Ising model and the propagator $D_m(n\tau)$ in the $\phi^4$ model,

$$\Gamma_{m,n} = \zeta D_m(n\tau),$$  (3.11)

where the Ising model is defined on a lattice with spacings $m$ and $n$, the propagator is defined on a lattice with spatial spacing $m$ and temporal spacing $n$, $\tau = -it$, and $\zeta$ is a constant of proportionality. Given this formal identification of models in statistical mechanics, the renormalization group formulation of the Ising model can be applied straightforwardly to quantum field theory.\[25\]

\[25\]There is a bit more work to be done to establish that the quantum field model reaches a critical surface when the continuum limit is taken, and this will vary from model to model within quantum field theory. See Wilson and Kogut [1974], Fraser [2018] for the remaining details.
To summarize, Wilson developed methods for solving models in physics for which energy scales are all highly linked, and qualitative behaviour is sensitive to the interactions spanning large ranges of energy. In order to apply these methods to non-Abelian quantum field theory—particularly Yang-Mills models—formal analogies with soluble models in statistical mechanics were essential. Wilson thus used the successes in statistical mechanics as a basis for formal analogies with quantum field theory, and found ways to apply the renormalization group analysis across both domains. In setting up the formal correspondence, putting quantum field models on a lattice was an important intermediate step as well. The importance of the inspiration from statistical mechanics was that renormalization problems had already been successfully solved there; Wilson was able to more-or-less carry over those results once he had established the identity in equation (3.11).

### 3.4.3 Renormalizability and effective field theory

Given the predominant viewpoint in modern physics that quantum field theories—and the standard model in particular—form an adequate framework only of effective field theories, the importance of renormalizability proofs for the acceptance of Yang-Mills theories might seem a bit odd. If we view quantum field theories as effective theories with a built in cutoff scale [Weinberg, 1979, Wallace, 2011, Williams, 2017, Fraser, 2017], then full renormalizability cannot be a necessary mathematical property of a physical model of quantum field theory, however nice it may be. So why was the HEP community so dismissive of candidate Yang-Mills models for the strong and (electro)weak interactions until ’t Hooft published a proof of their renormalizability?

The most important reason that a full proof of the renormalizability of Yang-Mills theories was essential to their acceptance is that the view of the standard model as a collection of effective field theories depends critically on the use of renormalization group methods to demonstrate that non-renormalizable terms become negligible at low energies. Weinberg [1979] outlined the utility of the use of what he called “phenomenological Lagrangians” as a tool for both understanding properties of interactions not captured within known models, and justifying the addition of nonlinear, nonrenormalizable terms. Though fully nonrenormalizable terms in a Lagrangian can only be used to make predictions at low orders—higher order contributions from such terms introduce a growing number of arbitrary parameters that must be fixed—one can study their scaling behaviour using the renormalization group methods. In this way, high energy properties—both qualitative and quantitative—of nonrenormalizable terms can be explored. Further, a general lesson from the renormalization group analysis of critical phenomena

---

26 There is not a consensus that an effective field theory view of the standard model is the best way to interpret the utility of quantum field theoretic models. Many people working in axiomatic and/or algebraic quantum field theory, for example, aim to provide an exact model for realistic interactions, to which the standard perturbative methods of conventional quantum field theory create an asymptotic expansion (e.g., Streater and Wightman [1964], Buchholz and Verch [1995], Halvorson and Müger [2007], Feintzeig [2017]). These may still be effective theories in the sense that they have a limited domain of applicability, but they would then be candidates for a more standard philosophical interpretation. Others have criticized the limited utility of a realist interpretation of effective field theories based on the renormalization group (Fraser [2018], Ruetsche [2018]). Though these are important philosophical issues, they are orthogonal to the discussion here. For the purposes of the main discussion, I will uncritically accept the effective field theory view, and attempt to explain why a proof of renormalizability is still epistemically important in HEP.
in statistical mechanics is that, at large distances (equivalent to low energy-momenta) many interaction terms become entirely irrelevant to the behaviour of the system in question. That is, the relative contributions from certain interaction terms “die off” at low energies, and many different theories (in the sense of having different terms in their Lagrangian) will all reduce to a critical surface in the space of Lagrangians on which only renormalizable terms have nonzero coefficients. The same is suspected to hold true for models in quantum field theory, such that a hierarchy of increasingly complicated Lagrangians is expected to obtain at higher and higher energies. These Lagrangians will not contain all and only renormalizable terms, and therefore would require some high-energy cutoff in order to generate empirical predictions. This is why many physicists and philosophers now view quantum field theory as a framework only for effective field theories, and that some new class of theory will be required for a “complete” model of fundamental physics applicable at all energy scales.

This modern view was not developed until after the construction of the electroweak model and QCD, and in fact could not have been convincingly argued for without the help of renormalization group techniques. Furthermore, the most convincing argument for a hierarchy of effective field theories—that low energy models of quantum field theory will retain only renormalizable terms—depends crucially on an understanding of the flow from general Lagrangians down to the critical surface of renormalizable interaction terms. The process of proving renormalizability is a two part process: first, one selects a general class of models with similar sets of terms (e.g., Yang-Mills models); second, one must develop appropriate novel techniques for actually proving that this particular class of Lagrangians is actually renormalizable, such as dimensional regularization. This is clearly how Veltman viewed the importance of renormalizability (cf. quote in §3.4.1), though he would presumably not subscribe to the effective field theory view.

I have shown that the process of arriving at the standard model of particle physics required the coincident development of mathematical tools for dealing with non-Abelian gauge theories and experimental discoveries corroborating the newly discovered properties of the framework. The mathematical developments connected non-Abelian Yang-Mills theories to experiment by demonstrating that renormalizability, asymptotic freedom, and confinement were all properties of QCD, fleshing out the bare Lagrangian form of the candidate model of strong interactions. This case study has provided lessons for the detailed process of theory construction in physics, highlighting the fact that theories are rarely axiomatic systems with a neat set of deductive consequences. Theories like the standard model are instead modular, and rely on key conceptual and mathematical tools that can largely be treated independently. In the case of HEP, the tools were often constructed by analogy with condensed matter physics. These lessons for theory construction can inform the process of constructing new theories, in particular those attempting to quantize gravity.
Chapter 4

QED, Q.E.D.

4.1 Introduction

Quantum electrodynamics (QED)—as part of the standard model of particle physics—stands as one of the major pillars of fundamental physics. The standard model is our best theory of the subatomic constituents of the universe, but is also widely regarded as merely an effective theory—an approximation to some more fundamental theory that would unify gravity with the atomic forces. For several decades, physicists have been hopeful that new physics will be discovered that falls outside the scope of the standard model, but to date the standard model has proven to be sufficient for all new data from the Large Hadron Collider. Many high-energy physicists have instead turned their attention to early universe cosmology as a way to test beyond standard model theories. The early universe is a window to physics well beyond our current standard model due to the extremely hot and dense conditions thought to obtain in the early stages after the big bang. However, the evidence one can gather from the early universe is highly mediated by our best theory of gravity and its model of the universe, which are thought to be theoretically incompatible with the quantum field theories at the foundation of the standard model.

A second option for insight into physics beyond the standard model comes from precision tests within the standard model. Unlike the early universe, precision testing does not allow for energy scales that go well beyond our current best theories; instead, one looks for minute discrepancies between measured and predicted low-energy phenomena within the standard model. One hope is that precision tests of the standard model will begin to reveal discrepancies that cannot be resolved by factoring in more detail from known physics. Failure to reconcile precision measurements with the standard model also provides hints as to what phenomena will become crucial for testing future theories. Since we are in a position to expect that the standard model will be succeeded by a new, more fundamental theory, it is useful to look to the details of the current generation of precision tests. We compare the current precision testing of QED with the best known example of precision testing playing a significant role in theory change—Newtonian astronomy—in order to: (1) outline the research program dictated by the structure of QED (2) demonstrate the ways in which successful precision predictions confirm QED and the standard model; and (3) outline how precision tests can eventually lead to discrepancies and hint at seeds of a new theory. In this paper we focus on precision testing of QED centred
How predictions are made in QED

around the fine-structure constant $\alpha$, since it is the fundamental coupling constant on which all QED predictions rely.

Smith [2014] has argued that research programs in physics generate knowledge by assuming that the theoretical framework is correct, and using it to search for the dominant causal factors affecting a system. Progress is made when careful experimental work precisely determines the values of certain parameters; the slight discrepancies that emerge between theory and experiment are then resolved by maintaining the essential correctness of the framework, and adding more detail into the model of the phenomenon. This can be accomplished through (1) improving the mathematical tools of analysis; (2) creating more realistic models of the known causal factors; or (3) including sub-dominant causal factors that were neglected on the first analysis. A familiar example of the latter would be accounting for friction in the motion of a ball rolling down an inclined plane. Trust in the framework as guiding this process only fails when there is a persistent failure to resolve these slight discrepancies. Nevertheless, the framework has generated genuine knowledge by identifying the dominant causal factors behind the phenomena. Smith’s work focuses on Newtonian astronomy, but as we will argue here, a similar structure holds for precision tests of the standard model.

The remainder of this paper is organized as follows. In §4.2 we outline some history regarding predictions in QED, and the structure of the relationship between theory and evidence. The relationship is one of ever-increasing precision in measurements of electromagnetic phenomena and their corresponding theoretically predicted values. Converging, sometimes independent lines of evidence have tightly constrained the low-energy value of the fine-structure constant over the past several decades. Next, in §4.3 we detail the current state-of-the-art in experimental (§4.3.1) and theoretical (§4.3.2) tests of QED. Right now, the most precise measurement is of the anomalous magnetic moment of the electron. The theoretical prediction already signals a small discrepancy between a pure-QED prediction and the measured value of the anomalous magnetic moment. In this case, however, the discrepancy is resolved within the standard model, and so no new physics is yet required. In §4.4 we briefly discuss the quantum Hall effect, one of a few QED-independent methods for determining the fine-structure constant. Independent precision tests like these provide confidence that “closing the loop” through QED-mediated phenomena is not viciously circular. In §4.5 we argue that this sort of precision testing could be highly useful in constructing models of physics beyond the standard model. The argument proceeds largely by analogy with the transition from Newtonian gravity to the theory of general relativity. Finally, in §4.6 we conclude by claiming that, even if the standard model is replaced, it has still generated real advances in scientific knowledge.

4.2 How predictions are made in QED

When one thinks about the relationship between theory and experiment in particle physics, quantum electrodynamics (QED) is often the paradigm example of close agreement and interplay between abstract formalism and experimental phenomena. Indeed, QED is often paraded as ‘the most rigorously tested theory ever’, or as ‘having the most precise agreement between theory and experiment’. This is the result of many iterations of precision tests of QED from the 1940s up to today. The most famous—and currently the most precise—test of QED is a determination of the anomalous magnetic moment of the electron.
The electron’s spin was discovered experimentally by Uhlenbeck and Goudsmit [1925], and the effort to include spin in a quantum mechanical description of the electron’s interaction with the electromagnetic field led to Dirac’s (1928) equation. According to Dirac’s equation, the magnetic moment of the electron due to its intrinsic spin is given by

\[ \mu_S = -g \frac{e}{2m_e} S \]  

where \( e \) is the charge of the electron, \( m_e \) its mass, \( S \) its spin, and the factor of \( g \) has the value of 2.

QED’s initial triumph came shortly after its full construction, in the form of Schwinger’s calculation of the first few significant figures in the anomalous magnetic moment of the electron, so called because it departs from Dirac’s value: \( a_e = (g - 2)/2 \). A pair of experiments conducted in 1947-48—by Nafe et al. [1947] via hyperfine splitting in hydrogen and deuterium, and Kusch and Foley [1948] via Zeeman splitting in various elements—measured a value of \( a_e(\text{experiment}) = 0.00119(5) \). Schwinger [1948b] developed techniques to handle radiative corrections, taming the divergences that had plagued earlier attempts to calculate quantities such as the self-energy of the electron. Using these techniques, Scwhinger found a correction to Dirac’s value for the electrons dipole moment to first order in the fine structure constant \( \alpha \), \( a_e(\text{theory}) \approx 0.00162 \). The result was calculated by renormalizing the one loop contribution to the electron vertex function. The value—which is inscribed on Schwingers tombstone—was in agreement with the contemporary experimental results, and this important success inspired confidence in QED. Today, as we will discuss below, the precision of agreement between theoretical and experimental determinations of \( a_e \) extends to \( \Delta a_e = 0.91 \times 10^{-12} \).

Central to this test—and all other precision tests—of QED is the fine-structure constant, \( \alpha = e^2/\hbar c \), the dimensionless parameter characterizing the coupling strength between electrons, positrons, and photons. The fine structure constant was initially proposed by Sommerfeld [1921] in a relativistic extension of the Bohr model of the hydrogen atom. This was the motivation for introducing a relativistic spin quantum number for the electron; \( \alpha \) was then interpreted as the ratio of minimum angular momentum allowed by relativity to that allowed by quantum theory. With the advent of QED in the 1940s, \( \alpha \) was reinterpreted as the fundamental dimensionless coupling constant for pure electromagnetic systems (i.e., consisting of positrons, electrons, and photons only).

Famously, quantum field theories do not predict the numerical values of their coupling constants. After renormalization, the physical charge of the electron in QED is strictly an empirical input. So prediction of quantities like \( a_e \) depend on a measured value of \( \alpha \) (or \( e \)). Additionally, predictions for most QED effects depend on a perturbative expansion of the generating functional, in powers of the coupling constant \( \alpha \). Some observable quantity \( F \) is expanded as a power series

\[ F(\alpha) = \sum_{n=0}^{\infty} A_n \left( \frac{\alpha}{\pi} \right)^n, \]  

where Feynman diagrams for the interaction are used to calculate the \( \{A_n\} \) to a given order. Predictions have constantly been improved by calculating effects to higher orders in the perturbative expansion.\(^1\) This complicates the picture of “confirming” QED beyond a simple

\(^1\) At some point, including more terms in the expansion will actually lead to decreased accuracy of predictions;
4.2. How predictions are made in QED

hypothetico-deductive case of deriving a value of $a_e$ then comparing to experiment; instead there is a continued process of refining the measured value of $\alpha$, using it as input for calculating higher order perturbative expansions, and comparing more precise predictions to a new generation of precision experiments.

If the determination of $a_e$ were the only QED effect to enter into this cycle, one might worry that the converging results were circular and therefore doing little to confirm QED. However, there are many experimental effects which can, when combined with the theoretical apparatus of QED, be used to either determine the value of $\alpha$ or compare to the QED prediction. Importantly, there are also ways to measure $\alpha$ that do not depend on relativistic effects of QED. These independent measurements of $\alpha$ come from effects in condensed matter physics, and provide a tight consistency check on the converging QED results. As one can see from Table 4.1, the various means for determining $\alpha$ show remarkable agreement to very high levels of precision. Many of these values are mediated by QED, though the phenomena vary. The most precise are the low-energy QED effects and the condensed matter measurements. The latter do not depend on the details of QED.

Table 4.1: Measurements of $\alpha^{-1}$. Low-energy QED values are taken from articles cited below. All other values are from Peskin and Schroeder [2018, p. 198]. The most precisely measured values come from low-energy QED tests and condensed matter.

<table>
<thead>
<tr>
<th>Low-energy QED</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$e^-$ anomalous magnetic moment</td>
<td>137.035 999 149 1 (33 0)</td>
<td>137.035 999 049 (90)</td>
</tr>
<tr>
<td>Atom recoil measurements</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Spectroscopic Measurements</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutron Compton wavelength</td>
<td>137.036 010 1 (5 4)</td>
<td></td>
</tr>
<tr>
<td>Muonium hyperfine splitting</td>
<td>137.035 994 (18)</td>
<td></td>
</tr>
<tr>
<td>Lamb shift</td>
<td>137.036 8 (7)</td>
<td></td>
</tr>
<tr>
<td>Hydrogen hyperfine splitting</td>
<td>137.036 0 (3)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Condensed Matter</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Quantum Hall effect</td>
<td>137.035 997 9 (3 2)</td>
<td></td>
</tr>
<tr>
<td>AC Josephson effect</td>
<td>137.035 977 0 (7 7)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scattering</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cross sections for $e^+e^-$ reactions</td>
<td>136.5 (2.7)</td>
<td></td>
</tr>
</tbody>
</table>

QED has thus generated a strategy for precision tests of electromagnetic phenomena: first, determine a simple leading order prediction of some effect, using the best available value of $\alpha$. The perturbative expansions in quantum field theory are actually divergent series, and are assumed to be asymptotic expansions of some deeper theory. Asymptotic expansions approximate a function up to some finite order, at which point additional terms take one further away from the true value of the function (cf. Peskin and Schroeder [2018], Zee [2010]).
This is the \( n = 1 \) term from Eq. (4.2). Next, this value is compared to the best available measurements; if the two values agree within their respective uncertainties, one aims to improve precision on both sides thereby reducing the uncertainties. If a discrepancy exists between theory and experiment, one should first aim to improve the theoretical prediction, including any additional details that may be relevant. In both cases, improved theoretical precision comes initially from including higher-order terms in the expansion. Difficulties arise at higher orders for a few reasons. First, the number of Feynman diagrams included in the determination of \( A_n \) increases exponentially with \( n \). Second, each individual diagram contributes more and more complicated integrals at high \( n \), meaning that numerical methods are needed to solve the integrals. Finally, the series expansions are known to diverge, and are therefore thought to be asymptotic expansions. Asymptotic expansions only provide good approximations up to some finite order, after which the prediction gets worse and worse. There is also no way to determine the order \( n \) at which this occurs, without knowing the underlying function to which the expansion is asymptotic. If it were possible, in practice, to calculate arbitrarily high order contributions to the expansion, at some point the theoretical prediction would diverge from the experimental value. Luckily, this last issue is only one of principle, since the practical difficulty of determining higher-order contributions means that the state-of-the-art is an expansion up to \( \alpha^5 \).

If discrepancies exist between theory and prediction at high precision, the standard model suggests that non-QED forces may be relevant. The strong, weak, and electromagnetic interactions are all unified under the standard model, and all contribute to interparticle interactions in highly complex ways. A system of electrons, positrons, and photons is the closest to a “pure QED” system, since these are stable under the electroweak force, and the strong force only acts directly on quark composites. However, the self-energy of the electron includes contributions from all virtual particles in principle, so some residual effects from the strong and weak sectors may play a noticeable role as precision of measurement increases. Finally, if one cannot account for discrepancies by including strong and weak interaction effects, one has reason to believe that some new physics is playing a role. We will see this process in action in the next section.

4.3 State of the art precision measurement and derivation of \( \alpha \)

4.3.1 Determining \( \alpha \) through the electron’s magnetic moment

The first experimental measurement of the magnetic moment of the free electron, showing that it deviates from Dirac’s value \( (g = 2) \), came from spectroscopic measurements of bound electrons. Kusch and Foley [1948] subjected beams of gallium, indium, or sodium atoms to an oscillating magnetic field, and determined the frequencies required to induce Zeeman splitting. These measured frequencies are related to the \( g \)-factor of the bound electron as \( \hbar \omega = g \mu_B B_0 \), where \( \mu_B = \frac{e \hbar}{2m_e} \) is the Bohr magneton and \( B_0 \) is the magnetic field strength. They avoided the challenge of determining \( B_0 \) to high precision by considering the ratio of frequencies associated with different transitions. This experiment—and all subsequent studies of bound electrons—provide only indirect measurements of the anomalous magnetic moment of the electron. Bound
electrons couple to external magnetic fields through their total angular momentum $J = L + s$, which includes both spin ($s$) and orbital ($L$) angular momentum. Spectroscopic experiments measure the full $g$-factor ($g_J$), and further assumptions are needed to extract the value for the free electron. Kusch and Foley evaluated $a_e$ based on a particular assumption about the electronic coupling. Uncertainties in the theoretical description of the spin-orbit coupling, and of the atomic nucleus itself, pose fundamental limits to precision in measurements of $a_e$ from bound electrons.

Measurements of free electrons are a promising avenue to attain higher precision: in principle, such measurements could directly determine $a_e$, avoiding the complications arising from atomic binding and the beyond-QED physics governing the constituents of the atom. Bohr discouraged this idea in the early days of quantum mechanics. He argued against the viability of free-electron measurements of the anomalous magnetic moment, using a Stern-Gerlach apparatus, on the grounds that the separation of an electron beam into distinct classical trajectories based on spin would violate the uncertainty principle [Garraway and Stenholm, 2002]. As [Louisell et al., 1954] emphasized, this line of argument was often taken to imply a much more sweeping prohibition of measurements based on free electrons than was warranted. They designed an experiment that determined the value of $a_e$ based on the precession of electron spin as a beam passed through a uniform magnetic field, between two scatterings. (There is then a simultaneous measurement of one component of $S$ and the position.) Within a decade, experiments based on spin precession in a static magnetic field eclipsed the precision attained by spectroscopic study of bound electrons [Rich and Wesley, 1972].

The highest precision measurements achieved to date are based on what Dehmelt called a “geonium atom”: electrons held in bound states in an earthbound apparatus. A device called a Penning trap effectively replaces the binding forces of an atomic nucleus with an adjustable combination of electromagnetic fields. This is a purely QED system, with the essential physics fully described in terms of leptons and interactions with the electromagnetic field. Stripping the nucleus out of the system removes the complications and sources of imprecision in atomic measurements, enabling incredibly high precision direct measurements of $a_e$.

A charged particle in a uniform magnetic field (with field strength $B$) moves in a circular cyclotron orbit, with a cyclotron frequency $\omega_c = \frac{e}{mc}B$. A Penning trap confines particles radially, roughly within a plane, using a uniform magnetic field along the $z$-axis. The particles are further prevented from moving away from the plane along the field lines by an electric

Figure 4.1: Scale drawing of a Penning trap, from Brown and Gabrielse [1986].
quadrupole field, produced by three electrodes—two end caps shaped as hyperboloids, and one ring electrode. These features of the trap are illustrated in Figure 4.1. The addition of the electrostatic field modifies the simple cyclotron motion in two ways. First, the quadrupole potential confines the particles, and also leads to simple harmonic motion with a frequency \( \omega_z \) along the \( z \)-axis. Second, the cyclotron frequency is reduced slightly (to \( \omega'_c \)) and the center of the cyclotron motion “drifts”. This slow drift is called the “magnetron” motion and has a much lower frequency \( \omega_m \).

The energy levels of the “geonium” atom consist of cyclotron energy levels (Landau levels), with further line splitting for spin, axial harmonic motion, and the magnetron motion. For an electron moving in a uniform magnetic field, there is a spin precession frequency \( \omega_S \) in addition to the cyclotron frequency \( \omega_c \). If these frequencies were identical, the energy levels of the atom would be degenerate: the \( s = +1 \) state at a given cyclotron level \( n \) would have the same energy as the \( s = -1 \) state at the cyclotron level \( n + 1 \). Yet due to the anomalous magnetic moment of the electron there is a small difference between these frequencies: \( \omega_n = \omega_S - \omega_c \). The anomaly \( \alpha_e \) can then be expressed in terms of this frequency, \( \alpha_e = \frac{\omega_n}{\omega_c} \).

Current measurements incorporate a number of ingenious experimental techniques for controlling the geonium atom and measuring the frequencies of transitions between different states to extremely high precision [Brown and Gabrielse 1986]. The axial resonance is particularly important, as it can be directly detected and experimentally manipulated by monitoring the voltage between the endcap and ring electrodes. The number of particles in the trap can be controlled, and a single particle can be maintained in a stable state for extremely long periods of time. Dehmelt famously kept one positron, which he named “Priscilla,” in a Penning trap for 3 months. The coupling between spin and cyclotron motions to axial motion is enhanced by introducing an additional inhomogeneous magnetic field (a “magnetic bottle”). Dehmelt introduced a technique based on the “continuous Stern-Gerlach effect”: the magnetic field induces a coupling between the axial motion of the electron and its spin orientation. Rather than the familiar spatial separation of different spin states, there is a separation in axial frequency, which is continuously monitored. Schematically, the experiment proceeds by driving the electron into a higher energy level, which simultaneously changes the spin orientation and cyclotron level. After the cyclotron motion returns to thermal equilibrium, the continuous Stern-Gerlach effect
is used to measure the spin state and determine the anomaly frequency \( \omega_a \). Since the cyclotron frequency is also simultaneously measured, this leads to a direct determination of \( a_e \). The crucial feature of this experimental design is the exquisite precision that can be attained in these frequency measurements.

[Hanneke et al., 2008] represents the state-of-the-art in precision measurements of the electron’s anomalous magnetic moment. A more detailed discussion of the experimental apparatus is found in [Hanneke et al., 2011]. The major advances in precision come from a few avenues. First, [Hanneke et al., 2008] use a cylindrical Penning trap, rather than the previously used hyperboloid geometry. The cylindrical cavity can be treated analytically, such that fringe effects are known and counteracted. This leads to a greater stability of the trapped electron, since “shifts of the electron’s oscillation frequencies are avoided” [Hanneke et al., 2011, p.3]. Avoiding shifts in oscillation frequency within the cavity allows for a more precise measurement of \( \frac{\omega_a}{\omega_c} \). Further, a surrounding electron plasma is used to determine frequency shifts of the cavity itself; knowing these cavity shifts eliminates a major source of uncertainty compared to previous measurements.

Next, Hanneke et al. employ quantum nondemolition measurements of the cyclotron and spin energy levels. A quantum nondemolition measurement leaves the measured state intact, so repeated measurements can be made without altering the state of the system. Formally, a quantum nondemolition measurement requires the Hamiltonian for the measured system—in this case the Penning trap—to commute with the Hamiltonian describing its interaction with the measuring system—here a one-particle self-excited oscillator. Additionally, the trapped electron “serves as its own magnetometer, allowing the accumulation of lineshape statistics over days” [Hanneke et al., 2011, p. 1]. The repeated measurements improve the accuracy of frequency measurements, and reveal that a major source of error remaining is the broadening of expected lineshapes over time.

All of these improvements lead to a measurement of the electron’s magnetic moment to a precision of 0.28ppt

\[
g/2 = 1.001 159 652 180 73(28) \quad (4.3) \\
a_e(HV08) = 115 965 218 0.73(28) \times 10^{-12}. \quad (4.4)
\]

The physics of Penning traps does not depend in any close way on the details of QED; non-relativistic quantum theory with classical electromagnetic fields is almost entirely sufficient to understand and control the single-electron stored in the Penning trap. In places, small relativistic corrections are needed, but these are low-order effects that do not depend on the details of QED.

### 4.3.2 Precision determination of \( \alpha \) through the standard model

Aoyama et al. (2012, 2018) performed a calculation of theoretical contributions to the anomalous magnetic moment of the electron, up to the 10th order\(^2\). The primary novel contribution in this paper is the calculation of Feynman diagram amplitude contributions at the 10th order, as well as an improved precision calculation of 8th order terms. For the current state of precision

\(^2\)Their result is 10th order in the electric charge—the expansion is taken to \((\alpha/\pi)^5\).
measurements, as discussed above, Aoyama et al. had to compute contributions to the anomalous magnetic moment that go beyond QED to include other parts of the standard model. The contributions to the anomalous magnetic moment of the electron can roughly be broken down additively as follows:

\[ a_e(\text{theory}) = a_e(\text{QED}) + a_e(\text{Hadronic}) + a_e(\text{Electroweak}), \tag{4.5} \]

with \( a_e(\text{QED}) \) being the “pure QED” contribution, \( a_e(\text{Hadronic}) \) the contribution from low energy quantum chromodynamics in the form of hadronic radiative corrections, and \( a_e(\text{Electroweak}) \) the contribution from higher-order electroweak diagrams. The QED contribution can be calculated as shown in equation (4.2), replacing \( F(\alpha) \) with \( a_e(\text{QED}) \) and \( A_n \) with \( a_e^{(2n)} \):

\[ a_e(\text{QED}) = \sum_{n=0}^{\infty} a_e^{(2n)} \left( \frac{\alpha}{\pi} \right)^n. \tag{4.6} \]

Each \( a_e^{(2n)} \) can further be broken down into terms independent of lepton mass, terms depending on the electron mass \( (m_e) \), the muon mass \( (m_\mu) \), and the tau mass \( (m_\tau) \):

\[ a_e^{(2n)} = A_1^{(2n)} + A_2^{(2n)}(m_e) + A_3^{(2n)}(m_e/m_\mu) + A_4^{(2n)}(m_e/m_\tau) + A_5^{(2n)}(m_e/m_\mu, m_e/m_\tau). \tag{4.7} \]

The simplest of these terms are the set of \( A_1^{(2n)} \), and the first three terms are known analytically. \( A_1^{8} \) depends on contributions from 891 Feynman diagrams, while \( A_1^{10} \) depends on 12 672 vertex diagrams. These terms must be evaluated numerically—\( A_1^{10} \) requires grouping of diagrams into distinct families, running separate solvers on each group. The bulk of the Aoyama et al. [2012] paper is dedicated to a detailed account of solving the largest set of diagrams contributing to \( A_1^{10} \), while other terms are taken or slightly improved from previous work.

Hadronic contributions to the anomalous magnetic moment can be broken down to contributions to vacuum polarization and light-light hadron scattering, while the electroweak term is solved analytically for one- and two-loop weak effects on the self-energy of the electron. Aoyama et al. [2012] are cognizant of the need for an independent input value of \( \alpha \) in order to determine the QED contribution to \( a_e \).

To compare the theoretical prediction with the measurement, we need the value of the fine-structure constant \( \alpha \) determined by a method independent of [the anomalous magnetic moment]. The best \( \alpha \) available at present is the one derived from the precise value of \( h/m_{Rb} \), which is obtained by the measurement of the recoil velocity of rubidium atoms on an optical lattice. (p. 3)

Using the rubidium recoil input value of

\[ \alpha^{-1}(\text{Rb10}) = 137.035 \, 999 \, 049(90) \, [0.66 \, \text{ppb}], \tag{4.8} \]

the total combined contributions to \( a_e \) give a value of

\[ a_e(\text{theory}) = 1 \, 159 \, 652 \, 182.032(13)(12)(720) \times 10^{-12}, \tag{4.9} \]

The separation is not purely additive, since \( a_e(\text{Hadronic}) \) will contain contributions from QED at higher orders, and \( a_e(\text{Electroweak}) \) will contain hadronic and QED corrections. However, the non QED corrections to \( a_e \) are not of sufficient precision for these nonlinearities to affect the prediction.
where the first and second error terms are due to the 10th order QED and hadronic contributions, respectively, and the third (and largest) uncertainty is due to the experimental uncertainty in the fine-structure constant $\alpha(Rb10)$. As mentioned above, the experimental value of the anomalous magnetic moment is $a_e(HV08) = 1.159652 \times 10^{-12}$, such that the agreement between the two is very high:

$$a_e(\text{theory}) - a_e(\text{HV08}) = (1.30 \pm 0.77) \times 10^{-12}.$$  \quad (4.10)

Agreement to a precision within the uncertainties of both the $a_e(\text{theory})$ and $a_e(\text{HV08})$ requires that $a_e(\text{hadronic})$ and $a_e(\text{electroweak})$ are included in the total calculation of the electron’s anomalous magnetic moment. The contribution to $a_e$ from hadronic and electroweak effects is $a_e(\text{Hadronic}) + a_e(\text{Electroweak}) = 1.735 \times 10^{-12}$. The increased precision in measurement from Hanneke et al. [2008] would have led to a slight discrepancy from a pure-QED prediction, and additional factors from the other forces of the standard model are therefore required to fully account for the best experimental value. At this point, we already see the QED research program pointing to physical details outside of pure QED.

Looking at the uncertainties in $a_e(\text{theory})$ and $a_e(\text{HV08})$, we see that the theoretical uncertainties are far smaller than the experimental uncertainties, with the largest source of uncertainty in $a_e(\text{theory})$ coming from the original input value of the fine-structure constant. Aoyama et al. [2012] make the same observation:

The intrinsic theoretical uncertainty ($\sim 38 \times 10^{-15}$) of $a_e(\text{theory})$ is less than $1/20$ of the uncertainty due to the fine-structure constant [Eq. (4.9)]. This means that a more precise value of $\alpha$ than [Eq. (4.9)] can be obtained assuming that QED and the standard model are valid and solving the equation $a_e(\text{theory}) = a_e(\text{experiment})$ for $\alpha$. (p. 3)

This leads to a new value for $\alpha$ of:

$$\alpha^{-1}(a_e) = 137.035\,999\,1491(15)(14)(330),$$  \quad (4.11)

where the uncertainties are due to the tenth-order QED prediction, the hadronic correction, and the experiment, respectively.

One might wonder why the discrepancy between the pure QED value of $a_e$ and $a_e(\text{HV08})$ is not an anomaly posing a problem for QED as a theory. A major reason for this is the history of increasing convergence between measured and theoretically determined values of the anomalous magnetic moment, beginning at the inception of QED. The ever more precise agreement between higher-order expansions of $a_e(\text{theory})$ and higher precision methods for determining $a_e(\text{experiment})$ has led to a fruitful research program. Further, we know that real-world particle interactions are highly complex and nonlinear, such that even relatively “clean” systems like the electron self-energy cannot be described with QED alone. The success in getting more and more precise agreement between theory and experiment with QED alone provides support to QED as correctly describing the anomalous magnetic moment of the electron to a high degree of precision. The presence of a discrepancy then indicates that the idealization of a pure QED system no longer holds. Since the discrepancy is resolved by introducing contributions from the strong and weak sectors of the standard model, we actually end up learning more about the


nature of $a_e$. The discrepancy serves as an indicator of the limits of the highly idealized model, and the theoretical framework of the standard model gives us the additional physics needed.

Compare this case to the prediction of a new planet—Neptune—within the framework of Newtonian gravity (cf. Brookes [1970]). In 1845, Le Verrier concluded that the irregularities observed in the orbit of Uranus could be resolved if there was an eighth planet orbiting the sun, outside of the orbit of Uranus. Neptune—the predicted planet—was observed one year later, in what was considered a major triumph of Newtonian theoretical astronomy. The success of the Newtonian framework led to precise agreement between observed and predicted planetary orbits. Small discrepancies in the orbit of Uranus were not seen as refuting the Newtonian system—instead Le Verrier looked for new physical effects compatible with the framework. The presence of a new planet resolved the discrepancy, and its eventual observation meant that the original discrepancy led to a more accurate understanding of the solar system.

The second interesting feature of Aoyama et al.’s [2012, 2018] work is that it serves two purposes: first, they predict a value of the anomalous magnetic moment of the electron to compare with the experimental results of Hanneke et al. [2008]. Second, the low theoretical uncertainty associated with their calculation coupled with the low experimental uncertainty of the experimental result allows for a new determination of the fine-structure constant. To date, this value remains the most precise determination on record.

As mentioned above, $\alpha(a_e)$ is calculated by assuming $a_e$(experiment) is exact, and that the 10th order QED expansion—plus the other standard model factors—exhaust the relevant theoretical factors to include. So, rather than $\alpha(Rb10)$ being used as the empirical input to determine $a_e$(theory), one uses $a_e(HV08)$ as empirical input to calculate $\alpha(a_e)$. It is important to remember that this result is independent of the prediction of $a_e$(theory). The values of $\alpha(a_e)$ and $a_e$(theory) cannot both be precisely correct, though the agreement between $a_e$(theory) and the experimental value $a_e(HV08)$ gives license to the new predicted value of $\alpha(a_e)$, since agreement (within error) between the predicted and measured value of the anomalous magnetic moment makes plausible that the Aoyama et al. calculation captures all relevant physics within the precision of the Hanneke et al. experiment. When a new, more precise measurement is made, the process will have to continue. More detail from within the standard model will be needed, and if successful, the constraints on the maximum deviation from the standard model will be tightened once again.

### 4.4 Independent measurements of $\alpha$: The quantum Hall effect

One might worry that, if all of the various precision determinations of $\alpha = e^2/\hbar c$ depend on QED-mediated calculations, then the constant increase in precision for determining $\alpha$ is merely a consistency check on QED. We have tried to show in the previous section that this is not the case. Over and above the convergence and increased precision offered by different theory-mediated measurements, as well as the multi-purpose theoretical predictions, we have access to independent methods for determining $\alpha$. These are not theory-independent methods, but rely on theories other than QED. Most directly, one can use the independent measurements of the electric charge from classical electrodynamics, and $\hbar$ and $c$ from non-relativistic quantum
mechanics and relativity, respectively. This historically provided an initial value with which to start the process of ever increasing precision described above. Even this fact lends some support to the legitimacy of high-precision QED-mediated convergence. However, the limits of precision—most notable for classical electrodynamics—of rudimentary measurements mean that this provides only a rough initial starting point. Advances from condensed matter physics allow us to perform high-precision independent measurements of $\alpha$. Though these tests don’t quite meet the precision of the best QED-mediated measurements, consistency within error between QED- and condensed matter-mediated measurements provide further support that the standard model is correctly describing the properties of the electron, and that these properties are stable regularities in nature. The most precise means of determining $\alpha$ outside of QED is via the quantum Hall effect.

4.4.1 Determining $\alpha$ using the quantum Hall effect

When a magnetic field is applied perpendicular to the 2D plane of the material, the electrons move in a cycloid pattern of radius $r = \frac{mv}{eB}$, with $v$ the velocity of the electron, and $B$ the magnetic field strength. Upon quantization, the allowed cyclotron orbits become discretized, with energy levels $E_n = \hbar\omega_C (n + 1/2)$, where $\omega_C = \frac{eB}{m}$ is the cyclotron frequency. In the quantum Hall effect, large magnetic fields are applied, so that $\omega_C$ is large. Each energy level—called a Landau level—is highly degenerate. The quantization of Hall conductance occurs when magnetic fields are sufficiently large that effectively all free electrons within the material occupy a single Landau level. At this point, the material’s resistivity is attributable to the resistivity associated with a single Landau level, making high-precision measurements of resistivity as a function of $B$ possible. This is why materials exhibiting the QHE are taken to be macroscopic quantum states. The high degeneracy of the Landau level at high $B$ and the low $T$ used to suppress thermal fluctuations effectively make all of free electrons behave in sync as a single quantum state.

In the regime relevant for the quantum Hall effect, the conductance $\sigma$ is measured as a function of $B$, and a plot of the resistance $\rho = \frac{1}{\sigma}$ shows a stepwise increase with magnetic field strength, while plateaus in the resistance increase in width for higher magnetic fields.

The precision with which the differences in resistance at each plateau can be determined indicate that $k$ is an integer to a precision of approximately 1ppb, which leads to a highly 

---

4For more details on the quantum Hall effect, see Yennie [1987], Prange and Girvin [2012], Tong [2016].
accurate determination of the ratio \( e/h \) at low energies \(^5\) Since \( \alpha = \frac{1}{4\pi\epsilon_0} \left( \frac{e^2}{\hbar c} \right) \), while \( c \) and \( \epsilon_0 \) are exactly defined quantities, the ratio determined through precision tests of the quantum Hall effect can be used to calculate \( \alpha \).

In practice, measurements of the quantum Hall effect are made to precisely determine the von Klitzing constant \( R_K = \frac{h}{e^2} \), as this is a phenomenological standard for the fundamental unit of resistivity in materials. Experimentalists seek precision in this measurement for reasons of metrology, independent of the measurements of \( \alpha \) from QED (cf. [Trapon et al. 2003]). However, the increased precision on \( R_K \) allows for increased precision in determining \( \alpha = \frac{4\pi\epsilon_0 c R_K}{e^2} \). The current state-of-the-art gives:

\[
\alpha^{-1}(\text{QHE}) = 137.035\,997\,9(32),
\] (4.12)

which—though not in exact agreement with Eq. 4.11—agrees to five decimal places.

The quantum Hall effect provides an important “independent” measurement of \( \alpha \), and is one of the most precise means for measuring \( \alpha \)—aside from the measurement of \( a_e \), atom recoil, and indirect calculation from the neutron Compton wavelength. Once consistency has been established between values measured using QED and independent tests like the quantum Hall effect, the worry about circularity from using a QED-mediated result to test QED is assuaged, and one can treat these precision measurements as direct tests comparing theory to experiment.

### 4.5 Using precision tests to go beyond the standard model

The predominant view of QED—and the rest of the standard model—is that it is an effective field theory. What this means is that it is an approximation to a more fundamental underlying theory, and that under suitable approximations, the effective theory can be found within the more fundamental theory. An effective field theory is just an effective theory that employs fields in its effective description. In most effective field theories—such as condensed matter quantum field theory and hydrodynamics—the fields are usually interpreted as a continuum approximation of the entities in the underlying theory—atoms and molecules, respectively. For the standard model, the fundamental underlying theory is, as yet, unknown. If the standard model is simply an effective theory, and therefore ultimately incorrect, why would physicists assume the validity of QED when designing and conducting precision tests of the electron’s properties?

The standard model forms a framework in which a research program of precision testing can be carried out. As [Smith 2014] notes for the Newtonian framework, which set a research program for astronomers for nearly 200 years:

[T]he primary question astronomers addressed when they compared calculations with observations is, What, if any, further forces are at work? The preoccupation of their research has not been with testing the theory of gravity, but with identifying further forces at work. To this end, their research presupposes the theory of

---

\(^5\)Given the scaling behaviour of coupling “constants”, one does not expect \( e \), and therefore \( \alpha \), to be constant at all energy scales. The low energy precision tests of \( \alpha \) should therefore give different results from high-energy tests—the \( \beta \)-function for QED indicates that \( \alpha \) should increase with increasing energy. According to the particle data group, \( \alpha \approx 1/128 \) when \( Q^2 = m_W^2 \), and the precision value quoted is for \( Q^2 = 0 \).
4.5. Using precision tests to go beyond the standard model

In much the same way, the back and forth process of measuring and determining from theory the anomalous magnetic moment—with ever more precision at each iteration—is predicated on QED, and the standard model more broadly. However, as mentioned in Section 4.2, there are complications to viewing QED and Newtonian gravity as exactly analogous. First, discrepancies between measured and predicted values of the anomalous magnetic moment are not necessarily a sign of new forces at work. Due to the fact that QED is described in terms of an asymptotic series, at some (unknown) point higher-order contributions to the expansion will diverge from the “true” value. We cannot use perturbation theory to get arbitrarily close to some exact value predicted by the theory; such a value would require a more rigorous formulation of QED, to which the generating functional provides an approximation. Given the current structure of QED, there is no “true” underlying value it predicts, to which the generating functional is an approximation. While complex approximations used to make predictions of $n$-body systems in Newtonian astronomy could be asymptotic series, in principle the Newtonian theory gives some complicated but exact expression for the evolution of the system. A better approximation scheme might not have the same fault, and since the “true” equation is known, bounds can be placed on the amount of error expected from an asymptotic approximation scheme.

Second, for Newtonian astronomy, the further forces at work are (almost) always gravitational forces. The deviations come from the complicated effects of smaller, nearby planets. These effects may be impossible to calculate analytically, and so require novel approximation procedures to include, but the working hypothesis is almost always that one has not taken into account the full details of other gravitating bodies. As we have seen for the anomalous magnetic moment, discrepancies between the measured value and the state-of-the-art “pure QED” required additional terms involving the strong and weak forces. So the methods of reasoning are slightly more complex in the case of the standard model.

Finally, Smith [2014] makes much of the robustness of the new forces at work. For example, the hypothesis that Neptune existed and explained the anomalies in Uranus’ had to be verified through the observation of Neptune itself, and other effects it has on the solar system. This is a crucial aspect of the success of Newtonian astronomy, and these causal dependencies survived the transition from Newtonian gravity to the theory of general relativity. This is heavily lacking in the precision testing of QED. Right now, the various tests converge to a high degree of precision, but there is little else that new features of the theoretical predictions are used to explain. For example, the Feynman diagrams from the tenth order (in electric charge) contribution to the electron’s anomalous magnetic moment are virtual processes, and therefore don’t have any independent causal powers. At most, the diagrams will have a similar contribution in other QED processes, and the numerical value determined from Aoyama et al. [2012] could be used in calculating these other effects.

All this said, precision testing of QED exhibits the same presupposition of the standard model that precision tests of astronomy did for Newtonian gravity. Importantly for the effective field theory view of the standard model, presupposing Newtonian gravity did not impede the development of general relativity. As Smith [2014] notes,

Strikingly, the transition from Newtonian to Einsteinian gravity, as a matter of historical fact, left all the previously identified details of our solar system that make
a difference and the differences they were recognized as making intact. In other words, the details that make a difference in our solar system and the differences they make proved more robust in the transition to Einsteinian theory than the Newtonian theory that had provided the basis for identifying them. This collection of difference-making details therefore has the strongest claim to knowledge produced by the two centuries of research predicated on Newtonian theory. But Newtonian theory also has a claim to knowledge, namely as a theory that, while holding only approximately over a restricted domain, still was adequate to establish many details that make a difference and the differences they make within that domain. (p.266)

In this way, the effective field theory view of the standard model is not simply compatible with presupposing the validity of the standard model, but is necessary for precision testing. But, more than this, precision testing can be our best guide to breakdowns in the standard model, and may indicate discrepancies to be remedied by some “beyond the standard model” theory.

Consider again the case of Newtonian gravity and general relativity. One crucial prediction for Einstein was the remaining 43″/century of Mercury’s perihelion precession. Over the previous century or so, astronomers were able to come up with more and more detailed models of Mercury’s orbit, and were able to account for much of the observed 575″/century precession. Most of this precession is due to gravitational “tugs” from near-solar celestial bodies. However, there were repeated failures to account for the further 43″ discrepancy, and this signalled that perhaps some modification to the inverse square law was required. Were it not for the work of successive additions of additional forces at work highlighting the missing 43″/century, it is likely that Einstein would not have had this crucial piece of evidence for his new theory. First, the lack of attention to the anomaly would have made it less likely that Einstein would have placed any significance on resolving it with general relativity. Second, even if Einstein calculated Mercury’s orbit with general relativity, the 43″/century makes up only a tiny fraction of the observed precession. The fact that general relativity predicts some precession is very different from predicting the amount needed to close the gap between predicted and observed precession after over a century of precision testing.

In much the same way, precision tests of QED—and the standard model more generally—may be crucial for testing new physics beyond the standard model. As we have seen with the case of the anomalous magnetic moment of the electron, precision testing has been a constant process of refining both observations and predictions. The standard model is presupposed in these tests, since QED effects need to be exploited in the experimental design. Predictions from the theoretical side are constantly taking into account finer and finer details of QED interactions in order to derive more precise results. So far theory has been able to stay in precise agreement with observation, though the results of Aoyama et al. (2012, 2018) have marked an important

6Some may be uncomfortable using with the term “prediction” to describe this episode, since the anomalous perihelion precession was known before Einstein derived it from the weak field limit of general relativity [Einstein 1915]. Many distinguish between predictions, which occur before an effect is known or measured, and retrodictions, which occur after. We believe that what is epistemically relevant for a prediction is the independence of the construction of the theoretical apparatus from the “predicted” phenomenon. In this case, Einstein constructed the general theory of relativity without factoring in the anomalous precession of Mercury’s perihelion. So the fact that Einstein’s derivation made up for the missing factors counts as a prediction on this reading. However, this is not a point that is essential to what follows. The reader can substitute “retrodiction” for “prediction” in the text here without altering our point.
step in that forces outside of QED are now required to maintain the agreement. QED alone is no longer sufficient to account for the total known self-interaction of the electron. At this stage, known interactions (low-energy weak effects and virtual hadronic effects) make up the difference. The ideal situation for those working on theories involving new physics would be a persistent discrepancy to emerge in the next round of precision tests—one that could not be remedied by including effects from the weak or strong forces.

Though the prospects of significant deviation from the standard model at low energies is an outside possibility, the prospects look dim that more direct, traditional methods in particle physics will reveal new physics any time soon. Many physicists fully expected evidence of naturalness at the LHC when the Higgs was discovered—in the form of new particles or evidence of new forces[7]. However, up to this point, no new physics has been discovered. The simple fundamental scalar Higgs seems to be the best supported by data from the LHC, meaning that the discovery of the Higgs has not provided any insight into physics beyond the standard model. Further, guided searches using simplified models have not found any new particles beyond the standard model [McCoy and Massimi 2018]. The LHC is operating at its maximum energies, so the lack of new physics discovered is concerning. Moreover, the LHC is entering a phase of testing, after years in the previous “search” phase. During the search phase, experiments were focused on finding evidence of the existence of new particles. Despite a few anomalies, search only found strong enough evidence for the Higgs boson to claim a discovery. The testing phase primarily involves mass production of the Higgs, to better determine its properties[8]. For this purpose, total centre of mass energy will be sacrificed for beam intensity near the Higgs threshold. If some new theory predicts new physics at an energy scale above the current threshold of the LHC, the only direct testing method would involve building a new, bigger accelerator. This would be an expensive, time consuming undertaking. Given that the Superconducting Supercollider was rejected by US congress, and the missing Higgs boson has been discovered, most high-energy physicists are pessimistic about the odds of a new accelerator being funded. If funding is approved, a new accelerator would take years to complete; for the time being, indirect methods of testing are the only window into physics beyond the standard model.

QED is the prime candidate for precision testing for a few reasons. First, most of the easily manipulable stable particles interact primarily via the electromagnetic force. Electrons are particularly useful for study, as they do not experience the strong force, and are stable under weak interactions. Second, the electromagnetic force is the weakest of the three forces described by the standard model[9] and is best suited to perturbative expansion. The process of refining predictions outlined in Section 4.2 depends on an increased precision coming from adding new terms from the perturbative expansion in powers of $\alpha$. For nonperturbative effects in the standard model—most notably in the low-energy quark regime—precision is possible in experimental detection, but difficult from the side of theory[10].

[7] Giudice [2017] is a prominent example of this sort of thinking. As the head of the theoretical team at CERN, his thoughts on the role of naturalness have evolved heavily since the dawn of the LHC [Giudice 2008].

[8] Though the precision here will ultimately be far lower than precision testing of the properties of the electron, the goal of testing the properties of the Higgs boson is to better understand what sort of particle it is. One hope is that increased precision on its properties will reveal that it is incompatible with the simple scalar description provided by the standard model; this would hint at new physics to be explained by future theory.

[9] At least when ordered by coupling constant; timescales associated with weak decays can often be longer than electromagnetic decays.

[10] A notable “precision test” of the standard model currently being conducted by numerous groups is the search...
There is precedent elsewhere in physics for using precision tests as a means to place tight bounds on the deviation of parameters predicted by current theory. Precision testing of general relativity was sought as early as the 1960s, and precise deviations from the theory were formalized in the parameterized post-Newtonian framework [Will, 1971]. This framework represents a limitation of general relativity to the weak gravitational (Newtonian) regime, and elevates certain key quantities that would parameterize deviations from general relativity into variables. Given this framework, one can characterize alternative theories of gravity by the value they give to these variables in a “theory-space” of post-Newtonian theories. Precision measurements come in to place bounds on the variables characterizing the post-Newtonian theory space. A similar formalism has been developed more recently to characterize deviations from the ΛCDM model of cosmology, and the Friedmann-Lemaitre-Robertson-Walker spacetime on which it is predicated [Baker et al., 2013]. The benefits of exploring—and ruling out—large areas of theory space with precision testing are clear. One can constrain future theories that go beyond our current best models, even in the absence of concrete proposals. Further, when new models are proposed, they can quickly be parameterized to fit within the theory-space, and the parameters from the reduced theory compared to experimental bounds.

One should not overstate the power of this form of testing, however. Any so-called “model-independent” framework—such as the parameterized post-Newtonian framework—must still make substantial assumptions in order to have any quantitative power. The parameters chosen will encode the community’s expectations regarding the ways in which future theory will deviate from current theory. This sort of project is not exempt from Stanford’s (2006) problem of unconceived alternatives. The assumptions made in constructing the parameterized theory space might still miss important alternatives. Fundamental changes to the concepts of the current theory—the standard model, general relativity, or the ΛCDM model of cosmology—are unlikely to be captured by these rather conservative extensions of the current theoretical framework. Simply replacing select constants in the standard model with unconstrained parameters does little to alter the conceptual framework currently in place. Further, the map from possible future theories to the parameterized theory space could be many-one, and there is no natural measure on the latter. This means that there is no well-defined sense in which one is ruling out large domains in the parameterized space, or that doing so would entail that realistic candidate theories are thereby heavily constrained.

That said, placing bounds on deviations from current theoretical expectations is a highly systematic method of constraining future theory, especially when little is known about the contours of that future theory. Whatever shape the future theory takes, the history of theory change suggests that it should be possible to limit its domain to the appropriate regime to compare the value of, say, the anomalous magnetic moment of the electron to that predicted by the standard model. Though the issues mentioned in the previous paragraph cannot be avoided, we must acknowledge that theory construction is essentially inductive, and therefore never infallible. Until a new theory is constructed whose restriction to low energies does not fall meaningfully outside of the standard model range in parameter space, we cannot say that for evidence of nonzero neutrino masses. Mixing of solar neutrinos suggests that at least 2 of the three neutrino flavours from the standard model have nonzero mass [Fukuda et al., 1998, Battye and Moss, 2014]. Neutrinos are currently treated as massless within the standard model, but models that factor in nonzero mass have been constructed [Gonzalez-Garcia and Maltoni, 2008, Ma, 1998]. Though outside the scope of this paper, these precision tests could also point to new physics beyond the standard model.
important alternatives are missed in theory space. Precision testing has an important role to play in the current landscape of theoretical physics, and precision testing of QED is the current gold standard in testing the standard model.

4.6 Conclusions

Precision testing of QED is a subtle process that relies on a mutual interaction between experimentalists and theorists. The fine-structure constant $\alpha$ is the key input needed for making predictions in QED, and can be experimentally determined in various ways. We have discussed the anomalous magnetic moment of the electron—arguably the most famous QED prediction—and the quantum Hall effect as QED-mediated and independent methods of measuring $\alpha$, respectively. As more and more precise measurements become possible, more sophisticated techniques are required to extract more detailed predictions from the generating functional of QED. As testing becomes more and more precise—supposing the QED and the standard model are able to account for the results within experimental error—the possible deviations from the standard model are constrained. This is useful in the process of constructing theories that go beyond the standard model, since these will ideally predict deviations from certain expected quantities within the standard model.

One can see the precedent for using precision tests of theories to aid future theory construction throughout physics. We discussed the example of the precession of the perihelion of Mercury in some detail, and mentioned more contemporary examples from cosmology. But there is value to the knowledge generated through precision testing, beyond its use for constructing future theories. To paraphrase Smith [2014], the standard model has so far proven adequate to establish details discovered in the process of precision testing QED. These details make a difference to our measurements, and the standard model establishes the nature of the differences they make.
Chapter 5

Conclusions

The goals of this dissertation were to gain insight into the epistemology and methodology of theory construction in high-energy physics, and to develop a new division of scientific disciplines suited to elaborating on the process of theory construction. We learn from Chapter 2 how an emerging discipline can refine its theoretical framework with the help of experiment and phenomenological models; in the case of the discovery of parity nonconservation for weak interactions, this was done in the absence of any dynamical model. Refinements to the theoretical framework were needed, as it was too sparse in the 1950s to construct a convincing model of the weak interaction. However, in Chapter 3 we learned that many of the key mathematical ingredients for the standard model—Yang-Mills theory, massive force-mediating bosons, and renormalization techniques for quantum electrodynamics—were in place in the 1950s, albeit in an underdeveloped form. Foundational work by Veltmann, ’t Hooft, and Wilson, among others, led to a deeper understanding of the representational capacity of quantum field theory as an appropriate theoretical framework for high-energy physics. In order to better understand that representational capacity of Yang-Mills gauge theories, new mathematical tools were needed. These include dimensional regularization, renormalization group flow of coupling constants, and lattice quantum field theory. Importantly for the construction of the standard model, these mathematical developments coincided with experimental advances, so that Yang-Mills gauge theories were quickly accepted as appropriate for modelling the strong and weak interactions. Finally, Chapter 4 illustrates the constant back-and-forth between the dynamical model and experimental techniques that occurs in the precision testing of quantum electrodynamics. In this final case, the theoretical framework and dynamical models for high-energy physics are matured and well-developed; precision testing is an important way to place constraints on future theories that go beyond the standard model.

What all of these cases show is close contact between experiment, phenomenological modelling, dynamical models, and the development of new mathematical tools. Theory construction in high-energy physics required the close interaction of physicists working in all areas of the field; I have argued that this is a feature that should be emulated in modern fundamental physics. As alluded to above, this process is being closely followed in the current era of precision cosmology. Those working on modelling physics beyond the standard model—particularly on quantum theories of gravity—should work closely with experimentalists to devise important experimental tests of their models, and to build phenomenological bridges between dynamical models and experiment.
I have also shown how historical case studies can play a role in informing foundational philosophical issues in fundamental physics. Symmetries play an important role in all areas of quantum theory, and were essential in the early development of high-energy physics as described in Chapter 2. However, as discussed in Section 2.4, symmetries served a largely heuristic purpose; the development of the electroweak model is a story of violating expected symmetries time and time again. Chapter 2 also illustrates how important experimental results are for streamlining theory construction. The construction of the standard model was so successful because experimental results were in plentiful supply. Model building was either trying to catch up to new experimental results, or was held in check by new results. Models that didn’t fit with new experimental results were abandoned quickly, and attention was focused to the successful models. In Chapter 3 this lesson was reinforced; though the focus of that chapter was the new mathematical techniques that were developed to show the representational capacity of Yang-Mills models, features such as confinement, asymptotic freedom, and weak neutral currents needed to be experimentally verified for quantum chromodynamics and the electroweak model to be accepted, respectively.

Chapter 3 also highlights the historical importance of dimensional regularization for proving renormalizability. Though the modern view of the standard model as an effective field theory downplays the importance of renormalizability proofs, they were essential to the initial acceptance of the standard model, and dimensional regularization is a useful technique for analyzing foundations of quantum field theories today. One recent dissolution of the cosmological constant problem uses dimensional regularization to renormalize the divergent vacuum energy away [Mostepanenko and Klimchitskaya, 2019]. In Chapter 4 I argued that the way forward today is via precision testing of the standard model. Quantum electrodynamics is the prime candidate for precision testing, due to the long range of the force and the stability of photons and electrons. Chapter 4 also gives reason to trust the results of the standard model and the methodology of testing by taking the theoretical framework for granted, even if we discover that there exists interesting physics beyond the standard model. High-energy physics is the twentieth century’s most precisely tested theory, and the standard model forms one of the two pillars of modern fundamental physics. Though it is almost certain to be succeeded by a new theory, the knowledge generated about the strong and electroweak forces will persist.
Bibliography


Curriculum Vitae

Name: Adam Koberinski

Post-Secondary Education and Degrees:

- University of Waterloo, Waterloo, ON
  2010-2014 B.Sc.

- University of Western Ontario, London, ON
  2014-2015 M.A.

- University of Western Ontario, London, ON
  2015 - 2019 Ph.D.

Honours and Awards:

- SSHRC Bombardier CGS D 2015-2018
- Western Doctoral Excellence Research Award 2016-2017
- Beyond Spacetime Essay Prize 2018
- Rotman Institute Graduate Book Prize 2018
- Richard A. Harshman Scholarship 2019

Related Work Experience:

- Teaching Assistant
  The University of Western Ontario
  2014 - 2018

- Graduate Research Assistant
  The University of Western Ontario
  2015-2019
Publications:


A. Koberinski. “Problems with the cosmological constant problem,” Forthcoming in *Beyond Spacetime, Oxford University Press*.


