BOSTON COLLEGE

GRADUATE SCHOOL OF ARTS AND SCIENCES

Department of Physics

A BROKEN SYMMETRY ONTOLOGY:
QUANTUM MECHANICS AS A BROKEN SYMMETRY

by

Jonathan E. Buschmann

Submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in the Graduate School of Arts and Sciences

July, 1988

BOSTON COLLEGE
GRADUATE SCHOOL

The thesis of Jonathan E. Buschmann

entitled A Broken Symmetry Ontology: Quantum Mechanics as a Broken Symmetry

submitted to the Department of Physics

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in the Graduate School of

Boston College has been read and approved by the Committee:

R. A. Uritam

D. A. Broido

P. M. Bakshi

Patrick H. Byrne

July 28, 1988
Date

© Copyright 1988 by Jonathan E. Buschmann

All rights reserved

A Broken Symmetry Ontology:

Quantum Mechanics as a Broken Symmetry

by Jonathan E. Buschmann

ABSTRACT

We propose a new broken symmetry ontology to be used to analyze the quantum domain. This ontology is motivated and grounded in a critical epistemological analysis, and an analysis of the basic role of symmetry in physics. Concurrently, we are led to consider non-heterogeneous systems, whose logical state space contains equivalence relations not associated with the causal relation. This allows us to find a generalized principle of symmetry and a generalized symmetry-conservation formalism. In particular, we clarify the role of Noether's theorem in field theory. We show how a broken symmetry ontology already operates in a description of the weak interactions. Finally, by showing how a broken symmetry ontology operates in the quantum domain, we account for the interpretational problem and the essential incompleteness of quantum mechanics. We propose that the broken symmetry underlying this ontological domain is broken dilation invariance.

Acknowledgments

I would like to thank Dr. Rein Uritam for his guidance as my thesis advisor and for the financial support he provided as chairman of the Physics Department. I would also like to thank the other members of my committee, Dr. Pradip Bakshi, Dr. David Broido, and Dr. Patrick Byrne for their comments and suggestions.

I especially thank my wife, Simonetta Malusa Buschmann, whose encouragement, support and love made the writing of this thesis possible. In addition, I thank her for her patience and hard work in typing this thesis.

Contents

I Interpretation of Quantum Theory

I.1 Introduction

Quantum mechanics is a theory used to describe systems in the nonrelativistic microphysical domain. It consists of a mathematical formalism: a set of primitive notions and a set of axioms involving these notions. The most important formalism from a formal and foundational viewpoint is that due to von Neumann (1955).[1] Its axioms are as follows:

This is a faithful latex-transcribed version of the original whose legacy-editor source has been lost. It includes some minor typographical corrections and two important corrections indicated also as margin notes.

Axiom I. To every system corresponds a Hilbert space $\mathcal{H}$ whose vectors (state vectors, wave functions) completely describe the states of the system.

Axiom II. To every observable A corresponds uniquely a self-adjoint operator A acting in $\mathcal{H}$.

Axiom III. For a system in state ϕ the probability $\rho_A(\lambda_1, \lambda_2|\phi)$ that the result of a measurement of the observable A, represented by $\hat{A}$, lies between λ_1 and λ_2 is given by $|(E_{\lambda_2} - E_{\lambda_1})\phi|^2$, where E_λ is a projection operator belonging to the spectral family of A.

Axiom IV. The time development of the state vector ϕ is determined by the equation $H\phi = i\hbar\partial\phi/\partial t$, where H is the evolution operator.

Axiom V. If a measurement of the observable A, represented by $\hat{A}$, yields a result between λ_1 and λ_2, then the state of the system immediately after the measurement is an eigenfunction of $E_{\lambda_2} - E_{\lambda_1}$.

The primitive notions included above are "system," "observable" and "state," which are correlated with the (assumed understood) mathematical object of a Hilbert space and its attendant entities in axioms I and II. The notions in axiom III of probability and measurement can also be taken as primitive. Their meaning in the context of axiom III is open

[1] See also Jammer (1974), Ch.l.

for interpretation and, in fact, such different meanings can yield different interpretations of the theory of quantum mechanics.

Axiom V, often called the "projection postulate," since in the discrete case it states that the system is projected onto an eigenstate by a measurement, is not strictly necessary for a description of quantum phenomena, and has, in fact, been altered or discarded by some interpretationists.

Other formalisms exist, most notably Dirac's, which, because of its brevity of notation and ease of calculation, is more widely used in practice. Still other formalisms were developed in order to make some quantum mechanical phenomena easier to describe, or to support a particular interpretation. Feynman's path integral approach makes clearer the wave nature of particles by emphasizing the idea of superposition in quantum mechanics and making a connection to the classical action. The S-matrix approach, by deemphasizing the time-development aspect of quantum mechanics, and concentrating on the observer-system interaction, has been used to support the similarly oriented Copenhagen interpretation. The aim of the algebraic approach and quantum logic is to avoid the interpretational problems of the standard formalisms by constructing a new non-standard algebra or logic on which to build a formalism.

We now turn to the subject of interpretation. This subject is, needless to say, a serious and complicated one for philosophers of science. Here we outline some of the basic principles that are generally accepted concerning the development of modern theories.[2]

There are several different uses of the word "interpretation." In order to connect a strictly mathematical formalism with observations, some of the primitive notions of the formalism must correspond with observations

[2] See Jammer (1974), Ch. 1.

via a set of rules. These rules are said to interpret the formalism. Most theories contain primitive notions which are interpreted in this way and some which are not. It is for this reason that such theories are called "partially interpreted" systems. In fact, von Neumann's formalism is already a partially interpreted system and not a strict mathematical formalism, by virtue of Axiom III, which provides such a rule of correspondence.

Such a partially interpreted system, however, lacks an explanatory capability. It may faithfully represent observations—and correlations that they infer—but it does not provide a genuine understanding of the ontological domain it refers to. Such a system does not posses any sort of unifying principle, and hence also lacks the ability to predict—disallowing the discovery of as yet unknown phenomena. Providing such a unifying principle to a system is also called an "interpretation."

Beyond merely a unifying principle is the construction of a model. A model can also be considered a system; however, it is a "heuristic system" as opposed to a formal system. The relation of a model to a partially interpreted system is like the relation of a computer program written in a high-level language to the corresponding assembly language program. The model provides a structure that is immediate to the mind and demonstrates in an obvious way the self-consistency of the theory; in other words, it provides a "picture."

It may turn out that a particularly appealing model (i.e., one that demonstrates a convincing cohesiveness and explanatory value) exhibits many of the characteristics of the corresponding formalism, but not all of them. We may then wish to change the formalism instead of changing the model. The formalism is more amenable to such changes, whereas the model would simply have to be discarded. This process is also given the name interpretation.

We finally note one particular way in which an interpretation may be found for a theory. It may be noticed that the mathematical formalism of a theory is equivalent or very similar to another well-established theory with an established model. The model of the established theory may then be proposed as a model for the uninterpreted theory. This was, in fact, the idea behind the early semi-classical interpretations of quantum mechanics.

The "theory of quantum mechanics," as we described it above, is a partially interpreted system. Quantum mechanics is unique in the history of physics in that this formalism was developed prior to, and independently of, a unifying principle or model. This is the sense, then, in which quantum mechanics lacks an interpretation. Many different interpretations have been put forward since the introduction of the quantum mechanical formalism. The Copenhagen interpretation, originally formulated by Niels Bohr, is often today called the "orthodox interpretation," but neither it nor any other interpretation has gained general acceptance by investigators of the foundations of physics.

What this means for the microphysical realm is that we have no generally accepted picture or "physical understanding" of this domain. The necessity of a model is an open question, and, as we will see, the Copenhagen interpretation explicitly denies the possibility of constructing a model for quantum mechanics. This would, however, be a unique case, and prior to the advent of quantum mechanics, the possibility of the non-existence of a model for a physical theory was never considered.

This question obviously raises epistemological issues which need to be addressed.

The EPR Paradox

Conceptual problems in physics are often clarified and sharpened by the construction of a so-called paradox. There have been many such paradoxes put forward concerning quantum phenomena. The most famous and probably the most important is the Einstein, Podolsky, Rosen (EPR) paradox (Einstein, Podolsky, and Rosen (1935).)

The EPR paradox involves a two-level quantum system. The original and subsequent formulations of this situation involve two particles produced in a zero-momentum state. EPR originally considered two microscopic particles produced such that

$$x_1 + x_2 = 0, \; p_1 + p_2 = 0,$$

where x_1, and x_2 are the positions of the two particles, respectively, and p_1 and p_2 are their momenta, and considered measurements of these variables after the particles had become spatially separated. Later constructions more commonly used today to describe the paradox and which have also been realized in experiment concern either two spin-$^1/_2$ particles produced in the singlet state or two photons produced in a similar state.

We consider here a two-electron system (first considered by David Bohm (1951),) produced as described above and allowed to become spatially separated. The spin part of the state vector of this system is

$$\psi = {}^1\!/\!\sqrt{2}[\hat{n} \uparrow (1) \otimes \hat{n} \downarrow (2) - \hat{n} \downarrow (1) \otimes \hat{n} \uparrow (2)], \tag{I.1.1}$$

where $\hat{n} \uparrow\downarrow (i)$ describes a state in which particle i has spin "up" or "down" along the $\hat{n}$ direction. This state is spherically symmetric so $\hat{n}$ can be any direction. After the particles have become separated their spins are measured with Stern-Gerlach apparatuses set up at locations A and B. If apparatus A is set up to measure spin along the $\hat{a}$ direction and apparatus B along the $\hat{b}$ direction, then the quantum mechanical expectation value for the observable $A_{\hat{a}} \cdot B_{\hat{b}}$, where $A_{\hat{a}}$ and $B_{\hat{b}}$ are the results of the measurements in units of $\hbar/2$, is

$$E(\hat{a}, \hat{b}) =< \psi|\sigma_1 \cdot \hat{a}\sigma_2 \cdot b|\psi >= -\hat{a} \cdot \hat{b}. \tag{I.1.2}$$

If $\hat{a}||\hat{b}$ we get the expected result $E(\hat{a}, \hat{a}) = -1$; i.e., the spins of the two electrons are anti-correlated. Since the state (I.1.1)—a state of superposition of the states spin up at A, spin down at B, and spin down at A, spin up at B—is considered to be a complete description of our knowledge of the system, we can never predict the exact results at both apparatuses but merely their anti-correlation. Consequently, if we measure the spin of particle 1 along the $\hat{a}$ direction, we will be able to predict with certainty the result of a measurement made immediately afterward on particle 2 along the $\hat{a}$ direction. According to the projection postulate, this is because after the first measurement the system immediately enters an eigenstate associated with the result of this measurement, which is either of the two superposed states.

This may be a startling observation—that a measurement whose result is strictly undetermined becomes determined by a remote measurement—but it does not by itself indicate a paradox: that is, a self-inconsistency in the formalism of quantum mechanics. However, had we chosen to measure the spin of particle 1 at A not along the $\hat{a}$ direction, but along a direction orthogonal to $\hat{a}$, we would have determined the result of the measurement of the spin of particle 2 at B along this direction. To be able to predict the spin of a particle along orthogonal directions does conflict with the quantum mechanical formalism, however, since these are non-commuting observables and cannot be simultaneously determined. This is the apparent paradox. The value of this paradox is that it compels one to further interpret the formalism.

EPR did not formulate their original argument as a paradox. It was their intent, instead, to demonstrate that the quantum mechanical description of the microphysical domain is an incomplete one. They defined a "complete" theory by requiring that for such a theory "every element of the physical reality must have a counterpart in the physical theory."

EPR's argument proceeded as follows. They assumed the correctness of the predictions of quantum mechanics as given by equations (I.1.1) and (I.1.2). They assumed a criterion for the existence of an element of physical reality: "If, without in any way disturbing a system, we can predict with certainty (i.e., with probability equal to unity) the value of a physical quantity, then there exists an element of physical reality corresponding to this physical quantity." In addition, they implicitly assumed no action at a distance.

EPR's conclusion that quantum mechanics is incomplete now follows logically from their premises. Since we can predict with certainty the spin of particle 2 along any direction without disturbing it (i.e., by measuring the spin of particle 1,) there must exist an element of physical reality corresponding to the spin of particle 2 along every direction. Since quantum mechanics does not allow the specification of the spin of a particle along orthogonal directions, these elements of physical reality have no counterpart in the theory, and quantum mechanics must be an incomplete theory.

There have been many proposed resolutions of the EPR paradox and/or rebuttals of their argument for the incompleteness of quantum mechanics. We will consider some of the more important such responses when we discuss some of the particular interpretational attempts of quantum mechanics, since, as we indicated above, answering EPR's argument requires one to take an interpretational stance.

I.2 The Copenhagen Interpretation

As mentioned earlier, the "orthodox interpretation" of quantum mechanics is the Copenhagen interpretation, also called the complementarity interpretation. It was originally formulated by Bohr,[3] and that version that exists today is, for the most part, identical to Bohr's ideas. Complementarity was put forward as a general principle by Bohr and was suggested by him and by others to be applicable to many other disciplines. We shall be mostly concerned, however, with the application of the principle to quantum mechanics as a possible interpretation. Briefly, the abstract principle of complementarity applies to a situation which can admit two descriptions which completely contradict and exclude one-another. Of course, these two descriptions cannot be applied "simultaneously" to the

[3] See, for instance, Jammer (1974), Ch.4.

situation at hand. Rather, only one description can be chosen, but which one can be used is indeterminate until chosen.

The need for this principle in quantum mechanics, according to Bohr, is due to the breakdown of the classical ideal of explanation in the microphysical realm—as exemplified by the indeterministic character of the predictions of quantum mechanics and the consequent wave-particle dualism—and the simultaneous need to express observations in classical terms. Thus different "complementary" classical concepts need to be applied to the same quantum phenomenon at different times. Generally, these complementary classical descriptions are a causal description and a space-time description. This accounts for the necessarily different forms of time development in quantum mechanics as given by the fourth and fifth axioms of von Neumann; i.e., the causal time development as described by Schrödinger's equation and the space-time description as given by the projection postulate. The Schrödinger equation affords a causal description of the time development of the state ψ, but ψ is not an observable object. To obtain a space-time description of the system associated with ψ, we must make a measurement on the system, but by doing so we introduce an "uncontrollable element" of disturbance, thereby destroying the causal description.

It should be noted that it is often claimed that there is also a kind of complementarity between the concepts of position and momentum and between the concepts of wave and particle. Although this position is held by some members of the Copenhagen school,[4] Bohr rejected these claims, since for him it is not the formalism or concepts of quantum mechanics

[4] C.F. von Weizsacker called this "parallel complementarity." See Jammer (1974), Ch. 4.

that stand in complementary relation, but only phenomena—requiring mutually exclusive classical "pictures"—can be complementary.

As we will discuss now, Bohr's reply (Bohr, 1935) to the EPR argument provides us with a bit clearer view of his complementarity interpretation. In his criticism of EPR's argument, Bohr explicitly rejected one of their premises, namely the reality criterion. In particular, he considered ambiguous EPR's claim that there existed an element of physical reality if one "can predict" the value of a quantity. For EPR the *actual* choice of the particular quantity to measure in a particular experiment is inconsequential to the physical character of the system at hand. In the quantum realm, however, Bohr claimed, it makes no sense to talk about the state of a system without reference to an experimental setup. For Bohr, the non-commutivity of the observables chosen by EPR to measure, and, hence, quantum theory's inability to specify values for both simultaneously, is just a direct reflection of the complementary descriptions needed to express each; consequently, different and mutually exclusive experimental procedures are required to measure these observables, since it is by these procedures and their results that we communicate these descriptions. Furthermore, any definition of physical reality in the microphysical domain must take this into account by acknowledging that it is only the object under investigation plus the measuring apparatus which can be considered the essential system in any ontological or epistemological analysis; any further "dissectional" analysis is necessarily ambiguous.

It is clear now from the rejection of the EPR reality criterion and the claim of an inseparable object-instrument description of quantum phenomena that the Copenhagen interpretation does not interpret quantum mechanics in the sense in which we claimed earlier quantum mechanics was lacking an interpretation; i.e., in the sense of a model. At the same time, however, these same positions obviously also prohibit the possibility of constructing a model for the microphysical domain. The Copenhagen interpretation, instead of interpreting in this sense tries to make more palatable "strange" non-classical phenomena and attempts to synthesize classical descriptions with these phenomena; nevertheless, it asks us to accept these phenomena prima facie.

As a consequence of this aspect of the Copenhagen interpretation, it is impossible to criticize it on physical grounds. It can only be criticized on the basis of its epistemological foundations or lack thereof. This we will take up in the next chapter.

I.3 Hidden-Variable Theories

One interpretation follows directly from the acceptance of the conclusions of EPR's incompleteness argument. This idea is that quantum mechanics needs to be completed—that is, supplanted by a theory which logically includes the quantum mechanical formalism but also satisfies EPR's criterion for a complete theory. Such theories (theories, since they propose to alter (i.e., add to) the formalism of quantum mechanics) are called hidden variable theories, or, more specifically, local hidden variable theories (LHTV), if they reject action at a distance. The idea behind these theories is that the state description of quantum mechanics needs to be supplanted by additional variables, which will complete the state description (in the sense in which EPR showed it to be incomplete), but which are hidden from observation and can in principle remain so.

The significance of these theories, even though, as first proposed, they could be experimentally indistinguishable from quantum mechanics, is that they allow one to retain EPR's reality criterion and ascribe our inability to measure and know, for instance, the simultaneous value of the spin of an electron along different directions, as merely that, a lack of knowledge. We are, then free to reject the Copenhagen interpretation's position that these values of spin are only definable with respect to an experimental arrangement and the subsequent measurement made with this apparatus.

Here, we will not consider any specific LHVT, although there have been such theories put forward.[5] Rather, we will present the startling result of Bell (1965), that the whole class of LHVT can be shown to have experimentally measurable differences with quantum mechanics.

Bell's Theorem

Consider, once again, the Bohm-EPR experiment with two electrons. The quantum mechanical description and predictions for measurements made on this system is given by equations (I.1.1) and (I.1.2).[6]

We consider now the description and predictions made by LHVT. Such theories will be defined as assuming the realism criterion[7] and a locality (no action at a distance) condition. The realism criterion is satisfied by replacing ψ, the quantum mechanical state, by a local realistic state λ (i.e., one that provides a complete description as explained above) with a distribution function ρ over a space Λ, so that

[5] See, for instance, Belinfante (1973) for a review.

[6] ~~Chapter and section numbers on equations will only appear when the equations are in different sections or chapters.~~

[7] See Section I.1.

$$\int_{\Lambda} d\rho = 1 \tag{I.3.1}$$

The locality assumption is implemented by assuming measurements made on the two individual particles yield results that are independent of the measurement made on the other particle; i.e.,

$$(A_{\hat{a}} \cdot B_{\hat{b}})(\lambda) = A_{\hat{a}}(\lambda) \cdot B_{\hat{b}}(\lambda). \tag{I.3.2}$$

The expectation value complement of equation (I.1.2) is then

$$E(\hat{a}, \hat{b}) = \int_{\Lambda} A_{\hat{a}}(\lambda) B_{\hat{b}}(\lambda) d\rho. \tag{I.3.3}$$

We further assume the strict anti-correlation between the measurements made on the two electrons when their spin is measured along the same direction:

$$A_{\hat{a}}(\lambda) = -B_{\hat{a}}(\lambda). \tag{I.3.4}$$

Now consider measurements made on the two electrons along different directions. Consider three different orientations of the Stern-Gerlach apparatuses $\hat{a}, \hat{b},\ and\ \hat{c}$. We can write

$$\begin{aligned}E(\hat{a},\hat{b}) - E(\hat{a},\hat{c}) &= \int_\Lambda [A_a(\lambda)B_{\hat{b}}(\lambda) - A_{\hat{a}}(\lambda)B_{\hat{c}}(\lambda)]d\rho \\ &= -\int_\Lambda [A_{\hat{a}}(\lambda)A_{\hat{b}}(\lambda) - A_{\hat{a}}(\lambda)B_{\hat{c}}(\lambda)]d\rho,\end{aligned}$$

using equation (I.3.4), and

$$E(\hat{a},\hat{b}) - E(\hat{a},\hat{c}) = -\int_\Lambda A_{\hat{a}}(\lambda)A_{\hat{b}}(\lambda)[1 - A_{\hat{b}}(\lambda)B_{\hat{c}}(\lambda)]d\rho,$$

since $|A_{\hat{b}}| = 1$. We next take the absolute value of both sides of this equation and, by taking the absolute value inside the integral, obtain an inequality:

$$\begin{aligned}&|E(\hat{a},\hat{b}) - E(\hat{a},\hat{c})| \leq \int_\Lambda [1 - A_{\hat{b}}(\lambda)B_{\hat{c}}(\lambda)]d\rho \\ &|E(\hat{a},\hat{b}) - E(\hat{a},\hat{c})| \leq 1 + E(\hat{b},\hat{c}), \qquad \text{(I.3.5)}\end{aligned}$$

where we have used equations (I.3.1) and (I.3.3). This is "Bell's inequality." Quantum mechanics can yield results in conflict with this inequality. In particular, choose $\hat{a}\cdot\hat{b} = \hat{b}\cdot\hat{c} = 1/2$ and $\hat{a}\cdot\hat{c} = -1/2$. From equation I.1.2 we find

$$|E(\hat{a},\hat{b}) - E(\hat{a},\hat{c})| = 1 \qquad Q.M.$$

and

$$1 + E(\hat{b}, \hat{c}) = 1/2, \qquad Q.M.$$

in obvious violation of the inequality (I.3.5).

This result—that all local realistic theories are in conflict with quantum mechanics, or that quantum mechanics cannot be subsumed as part of a local realistic theory—is known as Bell's theorem. Other inequalities—also called Bell inequalities—which allow for non-100% efficient detectors, and so are more useful for comparison to actual experiments, have been derived.[8] Experiment has overwhelmingly vindicated quantum mechanics and has thereby eliminated LHTV as viable alternative theories.[9] It also indicates that we must give up either locality or realism in the microphysical domain.

The newer inequalities were also necessarily derived under a broader assumption of locality. Whereas Bell's original assumption was that the state λ determined exactly the outcome of any measurement on the system (therefore, λ is to be considered a state of a "deterministic hidden variable theory"), the broader locality conditions assume that the state λ can evolve stochastically (thereby defining a "stochastic hidden variable theory") and/or that measurements can depend locally on random variables associated with the measuring apparatuses ("contextual hidden variable theory.") Since it is these inequalities that have been tested against, it is worthwhile to investigate this assumption behind them more closely to determine exactly what one must give up in light of the ex-

[8] See Clauser and Shimony (1978) for a review.

[9] See Clauser and Shimony (1978) and Aspect et al. (1982).

perimental results. We consider then a clarifying examination of this question by Jon Jarrett.

Jarrett's Work

Jarrett (1984) showed that the locality condition used in deriving most Bell inequalities is equivalent to two simpler conditions. He also carefully explicated what accepting or rejecting each of these conditions entails.

We, once again, consider the Bohm-EPR setup. We assume that any theory which correctly describes this experiment assigns a state description to the two electrons which yields a unique joint probability function $(d_1, x_1; d_2, x_2)$[10] for the result x_1 from a measurement of the spin of electron 1 along the direction d_1, and the result x_2 from a measurement of the spin of electron 2 along the direction d_2. As usual, we express the results of spin measurements in units of $\hbar/2$ so that $x_i = \pm 1$. This probability function is subject to the following obvious normalization requirements:

$$\sum_{x_1} (d_1, x_1; 0, 0) = 1 \tag{I.3.6a}$$

$$\sum_{x_2} (0, 0; d_2, x_2) = 1 \tag{I.3.6b}$$

$$\sum_{x_1, x_2} (d_1, x_1; d_2, x_2) = 1, \tag{I.3.6c}$$

where zeros indicate no measurement is made.

Jarrett defined the condition of "locality" by the conditions

[10] We use a much briefer notation than Jarrett. Wc do not exhibit those characteristics of the joint probability function due to other possible variables.

$$(d_1, x_1; 0, 0) = \sum_{x_2} (d_1, x_1; d_2, x_2) \tag{I.3.7a}$$

$$(0, 0; d_2, x_2) = \sum_{x_1} (d_1, x_1; d_2, x_2). \tag{I.3.7b}$$

Any theory which satisfies these conditions is said to be "local." These conditions state that the result of a measurement on electron i can depend only on the state it is in and on the state of the measuring device i (where the state of the measuring device is allowed to be specified by other variables in addition to d_i.) More to the point, such a measurement cannot depend on the state of the other, remote, measuring device. For this reason (as Jarrett explicitly proved) the locality condition prohibits the transmission of any information superluminally with an EPR setup. Hence, we may also call this condition "Einstein locality." It is also important to note that, although this condition does not allow a measurement to depend on the state of a remote measuring device, it need not be stochastically independent of the outcome of a measurement at the other measuring device.

This last observation leads to the next condition Jarrett defined. "Completeness" is defined by

$$(d_1, x_1; d_2, x_2) = \sum_{x'_2} (d_1, x_1; d_2, x'_2) \cdot \sum_{x'_1} (d_1, x'_1; d_2, x_2). \tag{I.3.8}$$

A theory is a "complete" theory if and only if it satisfies equation (I.3.8). In words, equation (I.3.8) says that our joint probability can be written as the product of two separate probabilities, each of which has the results at one detector "summed out." This condition demands exactly what Einstein locality does not—it demands the stochastic independence of the outcomes of measurements at the two Stern-Gerlach apparatuses. This condition of completeness does allow a measurement to depend on the state of the two-particle system and on the states of both measuring devices. We will refer to this condition later as Jarrett completeness.

"Strong locality" is defined by

$$(d_1, x_1; d_2, x_2) = (d_1, x_1; 0, 0) \cdot (0, 0; d_2, x_2). \tag{I.3.9}$$

A theory is said to be "strongly local" if it satisfies this condition. Condition (I.3.9) is just a statement of ordinary probabilistic independence—i.e., the joint probability can be written as a product of the two separate probabilities. This condition can be seen to reduce to Bell's original locality condition for deterministic theories, equation (I.3.2). Strong locality is also equivalent to the usual locality condition used in deriving the general Bell inequalities.

Jarrett next showed that equation (I.3.9) is logically equivalent to the conjunction of conditions (I.3.7) and (I.3.8). First, assume strong locality; then,

$$\begin{aligned}\sum_{x_2}(d_1, x_1; d_2, x_2) &= (d_1, x_1; 0, 0) \cdot [\sum_{x_2'}(0, 0; d_2, x_2')] \\ &= (d_1, x_1; 0, 0)\end{aligned}$$

by equation (I.3.6b). Similarly,

$$\sum_{x_1}(d_1, x_1; d_2, x_2) = (0, 0; d_2, x_2).$$

So, all strongly local theories are local. Also,

$$\begin{aligned}
&\sum_{x_2'}(d_1, x_1; d_2, x_2') \cdot \sum_{x_1'}(d_1, x_1'; d_2, x_2) \\
&= (d_1, x_1; 0, 0) \cdot [\sum_{x_2'}(0, 0; d_2, x_2')] \cdot [\sum_{x_1'}(d_1, x_1'; 0, 0)] \cdot (0, 0; d_2, x_2) \\
&= (d_1, x_1; 0, 0) \cdot (0, 0; d_2, x_2) \\
&= (d_1, x_1; d_2, x_2),
\end{aligned}$$

where in the second step we have used equations (I.3.6a) and (I.3.6b). So all strongly local theories are complete.

Now, assume locality and completeness. We find

$$(d_1, x_1; d_2, x_2) = \sum_{x_2'}(d_1, x_1; d_2, x_2') \cdot \sum_{x_1'}(d_1, x_1'; d_2, x_2)$$

from equation (I.3.8) and

$$(d_1, x_1; d_2, x_2) = (d_1, x_1; 0, 0) \cdot (0, 0; d_2, x_2)$$

from equations (I.3.7). Hence, all local complete theories are strongly local, completing the proof.

Experimental evidence can be taken as counting against strong locality. We must then give up either Einstein locality or Jarrett completeness as principles operating in the microphysical domain. Einstein locality is a well established physical principle; a principle not easily given up without contravening much independent experimental evidence. We are, therefore, compelled to sacrifice Jarrett completeness in the microphysical domain. Quantum mechanics is, of course, in agreement with experiment and, hence, violates strong locality and, we deduce, Jarrett completeness.

Why does Jarrett call this condition of stochastic independence of joint measurement outcomes completeness? First, we note that any deterministic theory automatically satisfies completeness; crudely speaking, a deterministic theory, by definition, does not yield probabilistic state descriptions, so results of joint measurements must be stochastically independent.[11] Now, a theory which yields a stochastic state description of our two-electron system which allows for a measurement result on one electron to be conditionalized on the outcome of a measurement on the other electron, obviously does not contain information which is predictively relevant for the outcome of this measurement. We see, then, that Jarrett completeness is a generalization for stochastic theories of the EPR completeness criterion.

A theory which correctly describes the microphysical domain (such as quantum mechanics) must yield state descriptions which are Jarrett incomplete. Hence, to say a theory does not satisfy Jarrett completeness is not to say it can be "completed." In Jarrett's words, "Although the term 'incompleteness' may connote a defect (as if all incomplete theories

[11] Hence, assuming Einstein locality, deterministic hidden variable theories are automatically strongly local.

may be ‘completed’), incomplete theories (e.g., quantum mechanics) are by no means ipso facto defective.” Let us call a theory which does not satisfy Jarrett completeness and is the correct theory of its domain an “essentially incomplete” theory, meaning that there is no correct complete theory of that domain. Any theory of the microphysical domain must then be essentially incomplete.

I.4 Stochastic Interpretations

It was noticed quite early in the development of quantum mechanics that there existed strong similarities between Schrödinger’s equation and equations of stochastic theories, such as brownian motion. In particular, formal analogies were drawn between the diffusion equation and the one-dimensional Schrödinger equation, and between the Heisenberg relations and similarly derived uncertainty relations with the diffusion coefficient replacing $\hbar$.[12]

In the late sixties, L. de la Pena-Auerbach et al. (1967, 1968a, 1968b), in a series of papers, demonstrated that there exists a possible “isomorphism” between non-relativistic quantum theory and the stochastic theory of Markov processes. He also showed that from this markovian-quantum theory, he could extract quantum formalism that is added in a much less natural way to the standard quantum theory. We will try to explicate some of the fundamental ideas and approaches of de la Pena-Auerbach’s work, after first introducing some of the basic formalism of Markov processes that is needed for this approach.

[12] See Jammer (1974), Ch. 2, for a history of these theories.

Markov Processes

The theory of Markov Processes[13] concerns a stochastic description of a given arbitrary distribution of particles with the assumption that the probabilities of the dynamical variables (such as position or velocity) of any given particle do not depend upon the entire past history of the particle. It is assumed, rather, that these probabilities are determined if one knows particular values of the variables at some particular point in the past.

For a Markov process, then, we speak of the conditional probability, which we designate as ρ (also called the probability density), that given the value, say x_0, for the position of a particle at $t = 0$, one will find the value for the position between x and $x + dx$ at time t. Symbolically,

$$\rho = \rho(x_0|x, t). \tag{I.4.1}$$

We now find a condition that this probability must satisfy by looking at how it changes in time. First, we have

$$[(\partial/\partial t)\rho(x_0|x, t)]\Delta t = \rho(x_0|x, t + \Delta t) - \rho(x_0|x, t). \tag{I.4.2}$$

We rewrite the first term on the right by making use of an identity from elementary probability theory, namely,

[13] See, for instance, Jammer (1974), pp. 437-8, Pathria (1978), pp. 462-3, Reif (1965), pp. 577-80, Wang and Uhlenbeck (1945).

$$\rho(x_0|x,t+\Delta t) = \int dx_1 \rho(x_0|x_1,t)\rho(x_1|x,\Delta t), \tag{I.4.3}$$

which is known as the Smolchowski equation. We now have for equation (I.4.2):

$$(\partial\rho/\partial t)\Delta t = \int dx_1 \rho(x_0|x_1,t)\rho(x_1|x,\Delta t) - \rho(x_0|x,t). \tag{I.4.4}$$

Parenthetically, we remark that this equation can be interpreted as saying that the change in probability is due to two sources. The first term on the right represents the probability of particles moving into $x, x+dx$ from any x_1, x_1+dx_1 during a time Δt, times the probability that they were in x_1, x_1+dx_1 at the time t (thus increasing $\rho(x_0|x,t)$.) The second term represents a loss of particles from $\rho(x_0,|x,t)$ during Δt into, say, any x_1, x_1+dx_1, since $\rho(x_0|x,t)$ is the probability that a particle was at $x, x+dx$ at time t (and must, therefore, have moved out during time Δt.)

Next, we let $x_1 \equiv x-\xi$, so that equation (I.4.4) becomes

$$(\partial\rho/\partial t)\Delta t = \int d\xi \rho(x_0|x-\xi,t)\rho(x-\xi|x,\Delta t) - \rho(x_0|x,t). \tag{I.4.5}$$

We note that $\xi \equiv x - x_1$ represents the "distance" between states x and x_1. We now expand the integrand of equation (I.4.5) in a Taylor series, in powers of $x_1 - x = -\xi$ about the point $x_1 = x$:

$$\rho(x_0|x-\xi,t)\rho(x-\xi|x,\Delta t)=\sum_n(-\xi^n/n!)(\partial^n/\partial x_1^n)[\rho(x_0|x-\xi,t)\rho(x-\xi,t)\rho(x-\xi|x,\Delta t)]_x,$$

and, since

$$(\partial/\partial x_1)[f(x_1)]_x=(\partial/\partial x)[f(x)],$$

equation (I.4.5) becomes

$$(\partial\rho/\partial t)\Delta t=\sum_n(-1^n/n!)(\partial^n/\partial x^n)[\rho(x_0|x,t)\int d\xi\xi^n\rho(x|x+\xi,\Delta t)]-\rho(x_0|x,t). \tag{I.4.6}$$

First, we note that the $n=0$ term in the sum is equivalent to

$$\rho(x_0|x,t)\int dx_1\rho(x|x_1,\Delta t),$$

and, since $\rho(x|x_1,\Delta t)$ must be properly normalized, i.e.,

$$\int dx_1\rho(x|x_1,\Delta t)=1,$$

we see that this term cancels with our last term in equation (I.4.6). Next, to simplify equation (I.4.6), we define the "nth moment of coordinate change" during Δt as

$$a_n \equiv (1/\Delta t) \int d\xi \xi^n \rho(x|x+\xi, \Delta t). \tag{I.4.7}$$

Using this in equation (I.4.6), we have for our rate of change of ρ:

$$\partial\rho/\partial t = \sum_n (-1^n/n!)(\partial^n/\partial x^n)[a_n \rho(x_0|x,t)]. \tag{I.4.8}$$

Next, we will assume what may be called the "Brownian motion approximation" by considering the particle of interest to be on a different scale than the particles constituting its environment—to be "relatively macroscopic" to the other particles. This is the approximation normally made in analyses of Brownian motion, but is, of course, a critical one for our discussion. We will discuss this assumption in more detail later.

The point of this assumption is that it allows us to assume that ξ, the size of our jump between x, and x, must be small during the "small" time interval $\mathit{\Delta} t$. Δt is considered "macroscopically infinitesimal." With this approximation, we ignore terms in equation (I.4.8) higher than order 2. Finally we have

$$\partial\rho/\partial t = -(\partial/\partial x)(a\rho) + (1/2)(\partial^2/\partial x^2)(b\rho), \tag{I.4.9}$$

where we have set $a_1 = a$ and $a_2 = b$. Equation (I.4.9) is known as the Fokker-Planck equation.

We will now need to generalize this equation for an n-dimensional Markov process:

$$\partial\rho/\partial t = \sum_i (\partial/\partial x_i)[a_i\rho + \sum_k (\partial/\partial x_k)(b_{ik}\rho)], \tag{I.4.10}$$

where a_i, and b_{ik}, are defined similarly to a and b ($-1/2$ has been absorbed into b_{ik}.) It can be shown that $a_i = k_i/\beta$ is the ith component of the external force per unit mass, **k**, divided by β, the friction coefficient and also that b_{ij} is the diffusion tensor.

We next recall the equation of continuity representing conservation of total probability,

$$(\partial\rho/\partial t) + \boldsymbol{\nabla}\cdot\boldsymbol{j} = 0, \tag{I.4.11}$$

which can also be written as

$$(\partial\rho/\partial t) + \sum_i (\partial/\partial x_i) j_i = 0. \tag{I.4.12}$$

Substituting for the left side of equation (I.4.10) from equation (I.4.12), we find

$$-\sum_i (\partial/\partial x_i) j_i = -\sum_i (\partial/\partial x_i)[a_i\rho + \sum_k (\partial/\partial x_k)(b_{ik}\rho)]$$

or

$$j_i = a_i\rho + \sum_k (\partial/\partial x_k)(b_{ik}\rho). \tag{I.4.13}$$

Also, since,

$$\boldsymbol{j} \equiv \boldsymbol{v}\rho, \tag{I.4.14}$$

where $\boldsymbol{v}$ is the macroscopic, or flow velocity of the particle, we have from equation (I.4.13)

$$v_i = a_i + (1/\rho)\sum_k (\partial/\partial x_k)(b_{ik}\rho). \tag{I.4.15}$$

These equations, (I.4.10) through (I.4.15), are the stochastic equations we will need to examine de la Pena-Auerbach's work.

de la Pena-Auerbach's Derivation of the Schrödinger Equation

In the first stage of de la Pena-Auerbach's work, he derived a Schrödinger equation by starting with a description of the motion of a particle in the context of Markov theory. First, we can write $\rho(x,t)$, our probability density in configuration space for the stochastic variable $x(t)$, as

$$\rho = e^{2R}, \tag{I.4.16}$$

where $R = R(x,t)$ is real, since ρ is positive definite. Using equation (I.4.16) in our equation of continuity (I.4.11) along with equation (I.4.14), we find

$$2(\partial R/\partial t) + \boldsymbol{\nabla} \cdot \boldsymbol{v} + 2\boldsymbol{v} \cdot \boldsymbol{\nabla} R = 0$$

or,

$$\partial R/\partial t = -(1/2)\boldsymbol{\nabla} \cdot \boldsymbol{v} - \boldsymbol{v} \cdot \boldsymbol{\nabla} R = 0, \tag{I.4.17}$$

where v_i is given by equation (I.4.15). We now assume that $\mathbf{v}$ can be written as the gradient of a real function, $S(x,t)$,

$$\mathbf{v} = \alpha \boldsymbol{\nabla} S, \tag{I.4.18}$$

which is equivalent to assuming that the external force is conservative. α is a real undetermined constant, that in some way characterizes the system. To obtain a Schrödinger-type equation, we introduce the field variable ψ,

$$\psi = e^{R+iS}, \tag{I.4.19}$$

so that $\rho = |\psi|^2$; i.e., ψ acts as the probability amplitude. We notice that equation (I.4.17), which is a differential relation between R and

$\nabla S = \alpha^{-1}\mathbf{v}$, can provide us with a differential equation for ψ. To this end, we multiply equation (I.4.17) by ψ to get

$$\psi(\partial R/\partial t) = -(1/2)\alpha\psi\nabla^2 S - \alpha\psi\boldsymbol{\nabla} S \cdot \boldsymbol{\nabla} R, \tag{I.4.20}$$

where we have used equation (I.4.18). To eliminate derivatives of R and S in favor of ψ, we compute

$$\partial\psi/\partial t = \psi[(\partial R/\partial t) + i(\partial S/\partial t)]$$

or,

$$\psi(\partial R/\partial t) = (\partial\psi/\partial t) - i\psi(\partial S/\partial t)$$

and

$$\begin{aligned}\nabla^2\psi &= \psi(\boldsymbol{\nabla} R + i\boldsymbol{\nabla} S)\cdot(\boldsymbol{\nabla} R + i\boldsymbol{\nabla} S) + \psi(\nabla^2 R + i\nabla^2 S)\\ &= \psi[(\boldsymbol{\nabla} R)^2 - (\boldsymbol{\nabla} R + i\boldsymbol{\nabla} S) + \psi(\nabla^2 R + i\nabla^2 S)\end{aligned}$$

or

$$\psi\nabla^2 S = -i\nabla^2\psi + i\psi[\nabla^2 R + (\boldsymbol{\nabla} R)^2 - (\boldsymbol{\nabla} S)^2] + 2\psi\boldsymbol{\nabla} S \cdot\boldsymbol{\nabla}\, R.$$

Using these relations in equation (I.4.20), we get

$$(\partial\psi/\partial t) - i\psi(\partial S/\partial t) = (1/2)i\alpha\nabla^2\psi - (1/2)i\alpha[\nabla^2 R + (\boldsymbol{\nabla} R)^2 - (\boldsymbol{\nabla} S)^2]\psi. \tag{I.4.21}$$

By defining a new function $V(x,t)$, such that

$$V = -(\partial S/\partial t) + (1/2)\alpha[\nabla^2 R + (\boldsymbol{\nabla} R)^2 - (\boldsymbol{\nabla} S)^2], \tag{I.4.22}$$

we finally get

$$i(\partial\psi/\partial t) = -(1/2)\alpha\nabla^2\psi + V\psi. \tag{I.4.23}$$

If we want this equation to describe the motion of a real particle with mass, m, then our constant $\alpha = \gamma/m$, where γ is now an undetermined constant. If we finally set $\gamma = \hbar$, then we have the Schrödinger equation. Setting the value of this constant is an independent postulate, but that is so in standard quantum theory also. For now, we will leave this constant arbitrary.

What has been shown, therefore, is that a particle that obeys stochastic laws (namely equations (I.4.11) and (I.4.18)), and whose flow velocity is given in the Brownian motion approximation by equation (I.4.15), can be described by a Schrödinger-like equation, with a complex probability amplitude whose norm is the stochastic probability density.

Derivation of Brownian Motion Equation from Schrödinger's Equation

To demonstrate the full relationship between Schrödinger's equation and the Brownian motion equation (I.4.10), de la Pena-Auerbach took the reverse course and derived an equation of type (I.4.10) starting from Schrödinger's equation. Then, starting with

$$i\hbar(\partial\psi/\partial t) = -(\hbar^2/2m)\nabla^2\psi + V\psi \tag{I.4.24}$$

and writing ψ as

$$\psi = e^{R+iS},$$

where R and S are real functions of the coordinates and time, we find

$$i\hbar[(\partial R/\partial t)+i(\partial S/\partial t)] = -(\hbar^2/2m)[(\boldsymbol{\nabla} R)^2-(\boldsymbol{\nabla} S)^2+2i\boldsymbol{\nabla} S\cdot\boldsymbol{\nabla} R+\nabla^2 R+i\nabla^2 S]+V$$

or,

$$\partial R/\partial t = -(1/2)\alpha\nabla^2 S - \alpha\boldsymbol{\nabla} R\cdot\boldsymbol{\nabla}\, S \tag{I.4.25}$$

and

$$\partial S/\partial t = -(1/2)\alpha\nabla^2 R - (1/2)\alpha[(\boldsymbol{\nabla} R)^2 - (\boldsymbol{\nabla} S)^2] + (V/\hbar), \quad \text{(I.4.26)}$$

where we have set $\alpha = \hbar/m$. We will discuss the significance of equation (I.4.26) later.

Taking equation (25) and multiplying through by the integrating factor e^{2R} we get

$$(\partial/\partial t)e^{2R} = -\alpha\ \boldsymbol{\nabla}\boldsymbol{\cdot}\ [e^{2R}\boldsymbol{\nabla} S].$$

Letting $\rho = e^{2R} = |\psi|^2$, we finally have

$$(\partial\rho/\partial t) + \boldsymbol{\nabla}\boldsymbol{\cdot}[\alpha\rho\boldsymbol{\nabla} S] = 0, \quad \text{(I.4.27)}$$

which is of the form of a continuity equation. It is this equation that can be written in the form of equation (I.4.10). To show this, we introduce a new function Q,

$$Q = R + S. \quad \text{(I.4.28)}$$

Equation (I.4.27) then becomes

$$(\partial\rho/\partial t) + \boldsymbol{\nabla}\boldsymbol{\cdot}[\alpha\rho(\boldsymbol{\nabla} Q - \boldsymbol{\nabla} R)] = 0$$

or

$$(\partial\rho/\partial t) + \boldsymbol{\nabla}\boldsymbol{\cdot}[\alpha\rho\boldsymbol{\nabla}Q - (1/2)\alpha\boldsymbol{\nabla}\rho] = 0, \tag{I.4.29}$$

where we have made use of $\rho = e^{2R}$. To identify this with equation (I.4.10), we see we must have

$$\alpha\boldsymbol{\nabla}Q = \boldsymbol{a} = \boldsymbol{k}/\beta \tag{I.4.30}$$

and

$$\alpha/2 = -b \equiv D, \tag{I.4.31}$$

where, to make the last identification, we have assumed that the diffusion tensor $b_{ij} = -\delta_{ij}D$; i.e., it is isotropic. To come to the conclusion, however, that equation (I.4.29) is a Brownian motion equation (that is, that Schrödinger's equation implies a stochastic process), we must accept the approximation made in deriving equation (I.4.10) earlier, the one which we stressed. This approximation is equivalent to assuming that the time interval of interaction, Δt, is large compared to the relaxation time of the medium. This relaxation time in Brownian motion theory is proportional to β^{-1}, the inverse of the friction coefficient already introduced in the above work. This restriction, though, is not contained in Schrödinger's equation, since it is valid for all time intervals. We postulate, then, following de la Pena-Auerbach, that the condition $\varDelta t >> \beta^{-1}$or $\varDelta t\beta >> 1$ is an oversimplified version of the time-energy uncertainty relation.

To see this connection, we use equation (I.4.30) as

$$m\boldsymbol{k} = m\beta\alpha\boldsymbol{\nabla}Q.$$

Since $m\boldsymbol{k}$ is the external force, we may write, roughly,

$$\varDelta E \sim \alpha\beta m\varDelta Q,$$

since $m\boldsymbol{k}$ can be written as the gradient of a potential. Using $\beta\varDelta t >> 1$ in the above equation, we get

$$|\varDelta E|\varDelta t \geq \hbar\varDelta Q \sim \hbar. \tag{I.4.32}$$

What is being claimed, then, is that the restriction $\varDelta t >> \beta^{-1}$ must be also placed on Schrödinger's equation, and it is, in fact, valid only over time intervals satisfying this restriction.

This "two-way" derivation has shown that there exists an intimate connection between quantum mechanics and stochastic theory. It has also enabled us to make the identifications (I.4.30) and (I.4.31), thus determining the constant α and giving us a relation between the applied forces $\boldsymbol{k}$ and the parameters of our stochastic processes. We will proceed from this point, then, and show how further formalism analogous to that of quantum mechanics is developed quite naturally, and we will also examine the physical content of this formalism.

Further Analysis

If $\hat{f}$ indicates any operator, then we define

$$< \hat{f} >_{AV} \equiv \int \hat{f} \rho d\boldsymbol{r} \tag{I.4.33}$$

and

$$< \hat{f} > \equiv \int \psi^* \hat{f} \psi d\boldsymbol{r} \tag{I.4.34}$$

as its mean and expectation values, respectively. Equation (I.4.15) can be written, using equations (I.4.31) and (I.4.18), as

$$\boldsymbol{v} = \boldsymbol{a} - (D/\rho)\boldsymbol{\nabla}\rho = \alpha\boldsymbol{\nabla}S, \tag{I.4.35}$$

or, substituting for ρ,

$$\boldsymbol{v} = \boldsymbol{a} - 2D\boldsymbol{\nabla}R. \tag{I.4.35a}$$

If we now define the operator $\hat{\boldsymbol{v}}$ by

$$\hat{\boldsymbol{v}} \equiv \boldsymbol{a} - D\boldsymbol{\nabla}, \tag{I.4.36}$$

we will have

$$\hat{\boldsymbol{v}}\rho = \boldsymbol{v}\rho. \tag{I.4.37}$$

Next we define $\hat{\boldsymbol{p}}$ as

$$\hat{\boldsymbol{p}} \equiv -im\alpha\boldsymbol{\nabla} = -2imD\boldsymbol{\nabla}. \tag{I.4.38}$$

The mean value of $\hat{\boldsymbol{p}}$ is

$$< \hat{\boldsymbol{p}} >_{AV} = -2imD\int \boldsymbol{\nabla}\rho \cdot d\boldsymbol{r},$$

so

$$< \hat{\boldsymbol{p}} >_{AV} = 0, \tag{I.4.39}$$

since the integral can be written as a surface integral and ρ must vanish at infinity. In the same way, from equation (I.4.36) we get

$$< \hat{\boldsymbol{v}} >_{AV} = \bar{\boldsymbol{a}} = \bar{\boldsymbol{k}}/\beta, \tag{I.4.40}$$

which says that the mean value of the flow velocity of the particle is proportional to the mean value of the force per unit mass acting on it.

To calculate the expectation value of p, we substitute in equation (I.4.34)

$$\begin{aligned}<\hat{\boldsymbol{p}}> &= -2imD\int \psi^*\boldsymbol{\nabla}\psi\cdot d\boldsymbol{r}\\ &= -2imD\int \psi^*(\boldsymbol{\nabla}R + i\boldsymbol{\nabla}S)\psi\cdot d\boldsymbol{r}\\ &= -2imD <\boldsymbol{\nabla}R> +2mD <\boldsymbol{\nabla}S>,\end{aligned}$$

but since $\hat{\boldsymbol{p}}$ must be Hermitian,

$$<\boldsymbol{\nabla}R>=0, \tag{I.4.41}$$

and, since from equation (I.4.35)

$$<\boldsymbol{\nabla}S>=<\boldsymbol{v}>/2D,$$

we get

$$<\hat{\boldsymbol{p}}>=<m\boldsymbol{v}>=<m\boldsymbol{v}>_{AV}\equiv m\bar{\boldsymbol{v}}. \tag{I.4.42}$$

This last equation allows us to interpret $\hat{\boldsymbol{p}}$ as the momentum operator, since its expectation value is equal to the mean flow of the momentum associated with the particle.

Next we introduce the operator $\hat{E}$,

$$\hat{E} \equiv 2imD(\partial/\partial t). \tag{I.4.43}$$

Computing $< \hat{E} >$, we get

$$< \hat{E} >= -2mD < \partial S/\partial t >, \tag{I.4.44}$$

where, as above, $< \partial R/\partial t >= 0$ because of the Hermicity of $\hat{E}$.

We now look at the extra equation that we obtained from our second derivation, equation (I.4.26). This equation can be written as

$$- < \partial S/\partial t >= D < (\boldsymbol{\nabla} S)^2 - (\boldsymbol{\nabla} R)^2 - \nabla^2 R > + < V > . \tag{I.4.45}$$

We also notice that

$$< \hat{p}^2 >= -(2mD)^2 < \nabla^2 R + (\boldsymbol{\nabla} R)^2 - (\boldsymbol{\nabla} S)^2 > . \tag{I.4.46}$$

Combining equations (I.4.45), (I.4.46) and (I.4.44), we find

$$< \hat{E} >=< \hat{\boldsymbol{p}}^2/2m + V >=< \hat{H} >, \tag{I.4.47}$$

where $\hat{H} \equiv \hat{\boldsymbol{p}}^2/2m+V$ is the Hamiltonian operator. In turn, we interpret $\hat{E}$ as the energy operator. However, we can also write equation (I.4.45), using equation (I.4.44), and (I.4.35) for ∇S as

$$< \hat{E} >=< (1/2)m\boldsymbol{v}^2 + V + \phi_B >, \qquad \text{(I.4.48)}$$

where

$$\phi_B = -2mD^2[(\boldsymbol{\nabla} R)^2 + \nabla^2 R] \qquad \text{(I.4.49)}$$

is called Bohm's potential. It might be thought that, in this context, the expectation value of the energy operator can be interpreted as the sum of the average kinetic energy of the flow, $< (1/2)m\boldsymbol{v}^2 >$, plus an effective potential, $V + \phi_B$, as has been done in earlier stochastic interpretations, when the term ϕ_B, has occurred. However, if we compute the expectation value for ϕ_B and use equation (I.4.35) to find

$$(m\boldsymbol{v})^2 = 4D^2(\boldsymbol{\nabla} S)^2,$$

along with our expression (I.4.46) for $< \hat{\boldsymbol{p}}^2 >$, we get

$$< \phi_B >= (1/2m) < \hat{\boldsymbol{p}}^2 - (m\boldsymbol{v})^2 > . \qquad \text{(I.4.50)}$$

We see that $< \phi_B >$ is actually just the difference between the total kinetic energy and the kinetic energy of the flow. In other words, it is the mean stochastic kinetic energy.

This questionable potential does point out, though, that we may look at the total energy in an alternative way—one that emphasizes the stochastic nature of our formalism. Since $\boldsymbol{a}$ has the dimensions of velocity and is proportional to the applied force ($\boldsymbol{a} = \beta^{-1}\boldsymbol{k}$), we may speak of it as an applied velocity and designate it $\boldsymbol{u}$. Doing this, we notice from equation (I.4.35a) that

$$\boldsymbol{u} = \beta^{-1}\boldsymbol{k} = 2D\boldsymbol{\nabla} R + \boldsymbol{v},$$

or

$$\boldsymbol{\nabla} R = -(1/2D)(\boldsymbol{v} - \boldsymbol{u}). \tag{I.4.51}$$

Using this result, ϕ_B can be written as

$$\begin{aligned}\phi_B &= -2mD^2[(4D^2)^{-1}(\boldsymbol{v} - \boldsymbol{u})^2 - (2D)^{-1}\,\boldsymbol{\nabla}\boldsymbol{\cdot}\,(\boldsymbol{v} - \boldsymbol{u})] \\ &= -(1/2)m(\boldsymbol{v} - \boldsymbol{u})^2 + mD\,\boldsymbol{\nabla}\boldsymbol{\cdot}\,(\boldsymbol{v} - \boldsymbol{u}). \end{aligned} \tag{I.4.52}$$

We can now write the energy as

$$< \hat{E} >=< (1/2)m\boldsymbol{v}^2 - (1/2)m(\boldsymbol{v} - \boldsymbol{u})^2 + mD\boldsymbol{\nabla}\boldsymbol{\cdot}(\boldsymbol{v} - \boldsymbol{u}) + V > . \tag{I.4.53}$$

This equation can easily be put in the form of Euler's equation for an ideal fluid, but as de la Pena-Auerbach pointed out, this does not lead to any deeper physical meaning. Instead, it just reflects that the nature of our starting basic equations was hydrodynamic.

To explain the nature of his formalism, de la Pena-Auerbach suggested that the stochastic behavior of a quantum particle is due to its interaction with the vacuum. In this case then β, which is normally a friction coefficient in stochastic theory, gives a measure of the interaction of the quantum particle with the vacuum. However, we should note that, from the Bell's theorem results of the last section, any such solution will require there to be a non-local aspect to this medium.

I.5 The Quantum Potential Approach

An interpretation that is in some ways similar to the stochastic approach discussed in the last section is the quantum potential approach. This approach was originated by de Broglie shortly after the formulation of quantum mechanics. It was taken up many years later by Bohm (1952a, 1952b) who continues to strongly advocate it today (Bohm and Hiley (1975, 1984, 1985)).

If we take Schrödinger's equation for a single particle

$$i\hbar\partial\psi/\partial t = -(\hbar^2/2m)\nabla^2\psi + V\psi$$

and let

$$\psi = Re^{iS/\hbar}, \tag{I.5.1a}$$

$$P = R^2, \tag{I.5.1b}$$

where R and S are real functions of the coordinates, then we have

$$i\hbar[\partial R/\partial t + (i/\hbar)R(\partial S/\partial t)] = -(\hbar^2/2m)[\nabla^2 R + 2(i/\hbar)\boldsymbol{\nabla} R \boldsymbol{\cdot} \boldsymbol{\nabla}\, S \\ + (i/\hbar)R\nabla^2 S - (1/\hbar^2)R(\boldsymbol{\nabla} S)^2] + VR. \tag{I.5.2}$$

Equating the imaginary parts of this equation, we find

$$\partial R/\partial t = -(1/m)\boldsymbol{\nabla} R \boldsymbol{\cdot} \boldsymbol{\nabla}\, S - (1/2m)R\nabla^2 S$$

$$\partial P/\partial t = -(1/m)\boldsymbol{\nabla} P \boldsymbol{\cdot} \boldsymbol{\nabla}\, S - (1/m)P\nabla^2 S$$

$$\partial P/\partial t + \boldsymbol{\nabla}\boldsymbol{\cdot}(P\boldsymbol{\nabla} S/m) = 0 \tag{I.5.3}$$

This last equation can be interpreted as a continuity equation. Equating the real parts of equation (I.5.2), we find

$$\partial S/\partial t + (1/2m)(\boldsymbol{\nabla} S)^2 + V + Q = 0, \tag{I.5.4}$$

where

$$Q = -\frac{\hbar^2}{2m}\frac{\nabla^2 R}{R} \tag{I.5.5}$$

is called the quantum potential.

We can show that this potential is identical to Bohm's potential found in the last section. There

$$\phi_B = -2mD^2[(\boldsymbol{\nabla} R')^2 + \nabla^2 R'].$$

Comparing the different definition of R and R', we find

$$R = e^{R'},$$

so that

$$\begin{aligned}[(\boldsymbol{\nabla} R')^2 + \nabla^2 R'] &= (\boldsymbol{\nabla} \ln R)^2 + \nabla^2 \ln R \\ &= (\boldsymbol{\nabla} R/R)^2 + \boldsymbol{\nabla}\boldsymbol{\cdot}(\boldsymbol{\nabla} R/R) \\ &= \nabla^2 R/R.\end{aligned}$$

We will discuss the significance of this equivalence later.

Equation (I.3.4) is equivalent to a Hamilton-Jacobi equation containing an extra potential Q. Consequently, the quantum particle is considered to have a definite well-defined trajectory with velocity

$$\boldsymbol{v} = \boldsymbol{\nabla} S/m. \tag{I.5.6}$$

The particle's behavior is determined by the quantum potential Q (in addition to V) which, in turn, is determined by $R = |\psi|^2$. ψ is then interpreted as a physically real field which accompanies and "directs" the particle.

This interpretation is a manifestly realistic one, since particles, their paths, and their influences are considered objectively real. The quantum potential is responsible for the non-classical features of quantum phenomena. This potential is explicitly non-local; since it it independent of the magnitude of ψ, it can be large where the wave function is small. It is dependent on the form of ψ and not on its magnitude; so, where ψ is rapidly changing is where the quantum potential is large. Consequently, this potential does not yield an ordinary mechanical force. Instead, it is an "informational potential," informing the particle in a non-local manner of its surroundings. Bohm says that this potential represents "active" information. The collapse of a wave function due to a measurement is explained by the fact that the irreversible nature of the measurement, that determines which state the particle is actually in, causes the "information" present in the other states to become inactive.

The above analysis easily generalizes for N-body systems. For instance, for a two body system which obeys the Schrödinger equation

$$i\hbar(\partial\psi/\partial t) = -(\hbar^2/2m)(\nabla_1^2 + \nabla_2^2)\psi + V\psi,$$

where $\boldsymbol{\nabla}_i$ refers to the i^{th} particle, the quantum potential is

$$Q = -\frac{\hbar^2}{2m}\frac{(\nabla_1^2 + \nabla_2^2)R}{R}. \tag{I.5.7}$$

The quantum potential now provides a means of non-local interactions between particles. This potential, which "informs" both particles, depends, through ψ, on the state of the entire system, and, hence, each particle has the potential to "inform" the other of its condition.

Whether there are such non-local correlations between the two particles, or whether two particles behave independently, such as classical particles do, depends upon whether the quantum state describing them can be written as a mixture (e.g., $\psi = \phi_1\phi_2$) or a pure state (e.g., the Bohm-EPR state). In the former case, Q reduces to a sum of independent terms, $Q = Q_1 + Q_2$, and the particles behave independently. In the latter case, there are non-local correlations.

The foremost problem with the quantum potential approach is the ontological status of Q. What exactly is this potential due to, how does it "inform" an electron, and why is it a potential if its effects are so much different than an ordinary potential? Bohm has suggested, like the proponents of stochastic interpretations, that there may be some underlying "ether" in which the non-local potential may act. Lorentz invariance could still be maintained and Einstein locality not violated if the structure of this medium was not directly revealable except at very high energies.

I.6 Statistical Interpretations

In this section we consider those interpretations of quantum mechanics known as statistical interpretations. These interpretations claim that quantum mechanics is not a theory which describes individual systems, but is a theory which only applies to ensembles of similarly prepared systems. The probabilities that the state vector yields are not probabilities to be associated with the result of a single measurement, but are instead to be interpreted as the relative frequencies for results of measurements made on such ensembles of systems.

The statistical interpretation has a long history, generally agreed to have been first proposed (and continuously adhered to) by Einstein. There have been many attempts to demonstrate a logical priority of this interpretation over the Copenhagen interpretation, but none have been generally accepted even by the proponents of the statistical interpretation.

The statistical interpretation is usually claimed to be compatible with hidden-variable theories, so that such a theory would be the complete theory describing individual systems, and quantum mechanics would be its statistical approximation. We will have more to say on this point later.

The other major proponents of the statistical interpretation have been Alfred Landé and, most recently, Leslie E. Ballentine. We will consider Ballentine's (1970) arguments in what follows.

The pure quantum state, because of its assumed statistical status, is assumed explicitly not to provide a complete state description of individual systems. For example, it "considers a particle to always be at some position in space, each position being realized with relative frequency $|\psi(r)|^2$ in an ensemble of similarly prepared experiments." The role of observation in quantum theory no longer plays a special role, since there is no wave function collapse.

Similarly, there is no need for an explicit concept of wave-particle duality. The apparent need for this concept, along with the phenomenon of interference, can be explained by the quantizing effects of the object or device with which the particle interacts. For example, in scattering of electrons from a crystal, the discrete set of scattering angles observed is due to the fact that momentum transfer to and from a periodic object must be quantized. Individual electrons (which, of course, are not being described) need not be assumed to be spread out in interaction with the crystal, rather "the electron interacts with the crystal as a whole through the laws of quantum mechanics."

At this point it is instructive to consider one "solution" to the EPR paradox which has been proven wrong, but whose refutation stresses the unusual and essential nature of the superposed state describing the two-particle system.[14] It is assumed in this treatment that, after the two particles have separated, they are correctly described not by the superposed wave function (I.1.1), but instead by a mixture of simple product states, each of the form

$$\psi_{\hat{n}} = \hat{n}^{\pm}(1) \otimes \hat{n}^{\mp}(2).$$

In each of these terms (i.e., for every $\hat{n}$), each particle is in a definite state, with a definite value of spin: hence, the EPR paradox is avoided. This description yields the same results as the pure state description when the spin of the two particles is measured along the same direction, or when only one particle of a pair has its spin measured. When measurements are made along different directions, however, the mixture analysis yields different results than the pure state. Simply put, this is because the mixture result is a simple sum of products of probabilities, whereas the pure state result arises from a sum of probability amplitudes and thereby al-

[14] This attempt was first suggested by Schrödinger (1935) and by Furry (1936) who demonstrated its consequences.

lows for probability amplitude interference effects. Experiment confirms the pure state treatment.

The reason for considering this “mixture attempt” at this point is that this would be the resolution offered by a straightforward naive application of the statistical interpretation. Of course, this is not the resolution given by the statistical interpretation, but it makes clear that additional assumptions are needed.

Of course, as mentioned earlier, the statistical interpretation explicitly assumes that quantum mechanics does not provide a complete description of individual systems. So, in this sense, they are in agreement with EPR. To show agreement with “orthodox” quantum mechanical results when treating the wave function statistically, attention is focused on the measurement process. In particular, the difference between “state preparation” and “measurement” is stressed. The process which is usually called a measurement (such as deflecting a particle with a Stern-Gerlach apparatus) is actually a preparation of a sub-ensemble of particles which are now additionally specified by their value of spin. Measurement is said to take place when particles are detected with a suitable detector placed behind the Stern-Gerlach apparatus. Quantum uncertainty associated with a “measurement” translates into a statistical dispersion associated with state preparation. However, just as in the “orthodox” treatment (Copenhagen interpretation) the state of the “state preparation” apparatus must be included in the selection of a sub-ensemble of particles with a particular value of spin.

In order to reproduce the results of quantum mechanics, we see from the above discussion that the statistical interpretation must make one of two choices. It can deny the possibility of a description of individual systems in the microphysical domain by claiming that quantum mechanics, as the correct theory of this domain, yields a statistical description of phenomena. In this case, it is logically equivalent, albeit conceptually different, to the Copenhagen interpretation, since it denies the possibility of a physical model of microphysical processes.

The other choice the statistical interpretation has is to claim that a more complete theory describing individual systems is available; but then it is subject to Bell's theorem just as any other hidden-variable theory is, and must then reduce to a non-local hidden variable theory.

I.7 The Many-Worlds Interpretation

Here we discuss an approach that has become known as the many-worlds, or many-universes, interpretation of quantum mechanics. On this interpretation, when a measurement is made, the "world" splits into many-worlds, each real and each associated with a possible result of this measurement. This interpretation was first formulated in the late 1950's by Hugh Everett III (1957) (along with John Wheeler), by whom it was originally called the "relative state" formulation. It was later supported and extended by Bryce Dewitt (1973) and others.

The intent of this interpretation is to avoid the discontinuous change required by a "collapse" of the wave function during a measurement. Motivation was also found for this approach by the desire of general relativists to define a wave function for the whole universe. The unique role of the observer in other available interpretations made such a wave function meaningless. In fact, Everett (1973) later called his theory the "theory of the universal wave function."

We begin by assuming that quantum mechanics yields a complete description by describing all processes by the continuous change of Schrödinger's equation. Hence, we accept all the formalism of quantum mechanics, except we reject any notion of wave function collapse. With this formalism we are able to describe all isolated as well as interacting systems. Observation is described as two systems interacting; hence, the observer has no special role.

Consider, then, a system consisting of two interacting subsystems, S_1, and S_2. This system is in some state ψ^S. Now, for every state of S_2, call it η, we can associate a "relative state" in S_1

$$\psi^{\eta}_{rel} \equiv N \sum_i (\phi_i \eta, \psi^S)\phi_i,$$

where N is a normalization constant and $\{\phi_i\}$ is an orthonormal basis in S_1. It is easy to show that the relative state is independent of the choice of this basis. For any linear operator in S_1, its expectation value calculated for this state yields a probability conditioned on the state η in S_2.

The standard quantum formalism tells us that, after the interaction of two systems, the state of the joint system is a superposition of correlated states of the two subsystems. In fact, one such representation of this superposition consists of a state of one system and its relative state.

Now we consider these two subsystems to be an observer and an object-system. A "measurement" now forms a superposition of the observer in the relative state ψ^{η}_{rel} and the object-system in the state η. The correlation between these states corresponds to the fact that the observing device is now in a state indicating some definite value associated with the state η (if there was a "good" measurement.)

The orthodox approach, in order to explain that we, as observers, observe a particular measurement result, postulates an instantaneous collapse of the wave function due to some, as yet, unexplained mechanism. Everett assumes no such collapse ever takes place. Instead the observation-interaction has split the system of observer plus object-system (or the world) into many real distinct worlds. An observer finds a particular result because he must find himself in one particular world.

Similarly, one can show that if a second observer is allowed to observe this composite observer-object system, there will again be a linear superposition. Each correlated element of the superposition will include the two observers recording the same observation, along with the state of the object-system correlated with this observation. So, by virtue of the quantum-mechanical formalism itself, there is no way to observe the splitting into many universes.

Strictly speaking, the many-worlds interpretation is a theory of measurement since it does not attempt to interpret the quantum formalism any further than to suggest that measurement or observation is to be understood just as any other interaction. In fact, Everett, and later others, claimed to have proved that the Born probability interpretation (in which the eigenstate expansion coefficients squared are interpreted to yield the probabilities for the outcomes of the associated eigenvalues under a measurement) can be derived from the formalism itself. However, this claim did not hold up under scrutiny and it is generally agreed today that some additional assumption, for instance, explaining why some particular observer enters a particular branch of the splitting universes, is needed.

The advantage of the many-worlds interpretation is the simplicity in description one gains in only requiring a single continuous kind of time development. It also has no need for postulating any mechanism, such as a conscious observer, to obtain a wave function collapse, nor does it need to alter or extend the formalism of quantum mechanics. At the same time, it yields a complete and "realistic" description of all phenomena. In addition, it is the only known interpretation to allow for the conception of a wave function for the entire universe.

The reason for putting "realistic" in quotation above is due to the extraordinary assumption the many-worlds theory makes about the nature of reality. Reality is conceived as to be undergoing a continual splitting into an indefinite number of separate realities, each distinct and "unobservable" from all others. In some sense, it could be claimed that this is the ultimate of metaphysical assumptions: assuming an infinity of worlds that can never be detected or known.

II Epistemological Considerations

II.1 The Need

The need to address epistemological considerations in a discussion of the foundations of quantum mechanics is evident from two aspects of quantum mechanics: its lack of an acceptable model and the novelty of quantum phenomena. Since such difficulty of providing a model for a theory has not previously been met with, the quantum situation raises the question of the need of this heretofore epistemological cornerstone. The "novelty" of quantum phenomena directly concerns the long-standing epistemological problems of the nature of the scientific concept of object and the problem of causality.

This need is further exemplified by previous interpretational attempts when one considers the serious epistemological issues they raise (usually only implicitly) and their failure to address these issues. We have already discussed some of these interpretational attempts; here we will mention the epistemological problems each raises and how each fails to address them; later we will criticize the implicit epistemology they require using the epistemological framework which we will develope.

The Copenhagen interpretation, by its doctrine of complementarity, implicitly denies the possibility of constructing a model for quantum mechanics: since quantum phenomena are to be strictly defined only relative to a given experimental arrangement no independent picture of the quantum realm is ever possible. The epistemological consequences of this fact are not considered by this school. The wave-particle duality, non-local correlations found in EPR setups and other quantum phenomena are accepted prima facie as new insights into the true nature of physical reality, though these concepts are at odds with our present epistemological framework.

Stochastic, local and non-local hidden-variable and many-worlds interpretations make particular metaphysical assumptions; that is, they postulate scientific objects which are outside the realm of experience, and can in principle remain so. Hidden-variable theories postulate additional physical variables, stochastic interpretations and the quantum potential approach postulate a new vacuum medium, and many-worlds interpretations postulate an infinity of additional realities. The possibility of allowing such objects in a theory is a serious epistemological problem, and a non-ad-hoc, sound epistemological framework needs to be developed to justify and support any such postulate. All fail in this respect.

Finally, the statistical interpretation, by suggesting the absence of a one-particle theory for the microphysical domain, proclaims an "epistemological void" in our physics. Once again the consequences for our theory of knowledge are dramatic and are not dealt with in this interpretation.

II.2 A Framework for Investigation

We will now attempt to construct an epistemological framework appropriate for the discussion of a physical problem. The early 20th century Neo-Kantian, Ernst Cassirer (1956), has put forth such a program. It is a program that, although begun from first principles, is both transparent and immediately applicable to the quantum-mechanical problem. Following this program will allow us to avoid the thorny metaphysical discussion that other approaches become involved with and which more properly belongs strictly to our philosophy. We will, consequently, summarize some of Cassirer's ideas here; although not entirely agreeing with his specific conclusion about quantum mechanics, we will then propose an epistemological framework. We will use this framework to criticize existing interpretations and later use it to construct our own interpretation.

Quantum phenomena do exist, and they occur consistently according to the laws of quantum mechanics. As Cassirer makes clear, our physical knowledge of the world consists of precisely these sorts of laws. We would like to know, then, if this new knowledge of the world has been given to us in a radically different way and if it has changed the way in which we consider knowledge of the world.

Cassirer presents a program for a consistent, pragmatic epistemology. As such it considers questions concerning our theory of knowledge only in so far as they concern the methodology of physics; hence, it is not a metaphysical but a "critical epistemology" to be used to analyze existing and possible concepts in physics. For Cassirer, experience is of "first" importance. Scientific knowledge is to be considered a rational ordering of experience; therefore, epistemological statements are statements concerning experience. There are, according to Cassirer, different types of such physical statements, which form a sort of hierarchy. These are, first, statements of the results of measurements, then statements of laws, and then statements of principles. We will not discuss here the epistemological significance of each of these types as explicated in detail by Cassirer. What is important for our purpose is that the "general principle of causality" occupies the last level of hierarchy in this categorization of physical statements.

The causal relation is a special concept, because with it we can form definite empirical concepts. In fact, by applying it to our experience, it is the means by which we construct laws of nature; i.e., it is a statement concerning method. Accepting this approach, therefore, we immediately realize that it would be wrong to consider the causal relation as a special law of nature and we realize that it cannot be tested as such. It is, then, totally inappropriate to speak of the causal law as being found to fail in some instance. Rather, such "indeterminism" would be properly defined as nature arbitrarily applying whatever law it wished from case to case. But even this concept is wrong, since now we would be improperly anthropomorphizing nature. In Cassirer's words, "In strict physical terminology 'nature' is nothing but an aggregate of relations, of laws; and to such an aggregate, to such a pure form, the category of active or passive is not applicable."[15]

Let us now consider the causal principle itself, expressed in its logical form: i.e., if x, then y. If, after establishing a physical law via the causal principle, we find we can apply this law to cases where there is doubt or uncertainty in the premise, x, what do we say about this causal relation? We can see from what has been said above, and we also know directly from traditional logic, that the validity of the premises does not affect the form of the causal law; i.e., it is obvious, once again, that it would be wrong to attack the causal principle in such a case. What we must realize, however, is that for an actual case to properly describe natural phenomena a cause must be a "true" cause; otherwise, the meaning of the causal relation is ambiguous. Here, a true cause is one that can somehow be directly or indirectly experimentally proven; or, in the language of the

[15] Cassirer (1956), p. 119

present framework, an alleged cause must be directly empirically given; that is, an observable fact.

Cassirer also considers the relationship between our scientific concepts of objects and physical reality. Such a concept, he claims, is the result, the consequence, of our experience; it is not some thing before us waiting to be discovered. It constitutes (at any given time) a limit of our experience, but not a permanent limit to knowledge. Hence, we can understand how our concepts and definitions of objects change through time. But, Cassirer points out, without need for any recourse to metaphysics we can realize that there must be a "fundamental demand" which is consistent and unchanging, which these concepts satisfy. As Cassirer says, "Just as the geometrician selects for investigation those relations of a definite figure, which remain unchanged by certain transformations, so here the attempt is made to discover those universal elements of form, that persist through all change in the particular material content of experience."[16]We can say, then, that these invariant "universal elements of form" underlie our epistemological consistency. If these invariants were lacking our scientific concepts of objects would be inconsistent or confused.

Our Framework

How does Cassirer's epistemological framework help us deal with the quantum-mechanical situation? We see first of all there is no question of an indeterminism. The causal relation has been applied to discover the laws of quantum mechanics, which are consistent and well-confirmed. What these laws are concerned with (i.e., the wave function), however, is not an observable—not a "true cause." For the eigenstate case, however,

[16] Cassirer (1953), pp. 268–9, reprinted in Cassirer (1956), p.l38.

or in certain situations in case the quantum mechanical state can be described as a mixture of eigenstates, the quantum-mechanical treatment can deal directly with observables.[17] So, although it is improper to speak of an "indeterminism" in quantum mechanics, application of the concept of causality in quantum mechanics is restricted in some sense. In other words, since the causal principle is the means by which we order our experience, in the quantum-mechanical situation we are epistemologically inhibited.

To more fully understand what this last statement means, let us consider, via the above analysis, the concept of object offered in quantum mechanics. We are already familiar with the fact that there is an epistemological novelty in our scientific concepts of objects in quantum mechanics; e.g., the wave-particle duality. This is a novelty in our epistemology because we have not discovered new objects, but rather have changed the way in which we define objects; and, as we argued above, the concept of object is not a thing to be discovered, but a conceptual device we apply to experience. This "novelty," then, represents a breakdown in the normal process of ordering experience. This is consistent with the above analysis of determinism in quantum mechanics, where we also found that we are to expect some breakdown in our epistemology of the quantum realm. What this means for the nature of physical reality is also clear from our above analysis: since quantum mechanics is well-confirmed as the correct theory of its ontological domain, we expect some invariant, some universal element of form, to be lacking or faulted in this domain.

[17] In this context Cf. Howard (n.d.), where Bohr's complementarity interpretation is taken as stating a correspondence between quantum mixtures and a classical description. Also, recall the "mixtures resolution" of the EPR paradox, Section I.6.

II.3 Explicit and Implicit Philosophical Positions in Interpretations of Quantum Mechanics

We can divide the implicit or explicit epistemological positions taken by interpreters of quantum mechanics into two groups: those who assume a realist epistemology and those who assume a pragmatic one. It is the pragmatists' aim not to interpret the quantum formalism any further than is practically needed to correlate experiment with theory. It is generally their view that the role of physics is to relate observations to one another and to be able to predict with increasing accuracy outcomes of such observations. Physics as "explanation," they contend, has never been more that this, because it never really explains the why behind physical phenomena. The "aim" of physics is to merely order our experience, not to construct a model of physical reality. Such a pragmatic epistemology was supposedly favored by Bohr, and has been explicitly adopted by Copenhagen interpretation supporters today.[18] The fundamental idea behind the doctrine of complementarity, that any description of physical phenomena must be made with reference to a particular experimental arrangement, exemplifies these pragmatic ideas. A picture or model of the microphysical domain is clearly renunciated and prohibited. Once again, the "aim" is to order experience and subsequently to communicate it, and this requires descriptions to be in classical terms, those in conformance with experimental arrangements.

The need to communicate in classical terms was, for Bohr, the motivating factor behind the adoption of this philosophy. This is also probably the weakest point in his approach. This suggests that our concepts and descriptions can never develope and evolve. This is consistent with

[18] See, for instance, Stapp (1972).

the obvious criticism of the realists—that without a picture or a model progress becomes stagnated because it becomes difficult to discover new phenomena. In other words, there is a certain sterility inherent in this pragmatic approach. We do not wish so much to criticize this particular philosophy, here, as to point out that it entails an in depth philosophical structure. The obvious criticism of a pragmatic epistemology is that this is a subjective philosophy. To defend against this charge requires an extensive ontology of ideas and their relationship to experience (which, we note, is quite at odds with a Kantian epistemology.) To rely so heavily on such an (controversial) ontology to describe a particular branch of physics is not very practical.

Our epistemological framework is a truly practical one. It is a framework to be used for all domains of experience. It makes a minimum of assumptions, and requires no ad-hoc ontological framework for any particular domain. It gives no special status to any ontological domain nor a special place to any particular set of concepts.

The statistical interpretation claims to avoid the epistemological problems encountered by the Copenhagen interpretation by avoiding a one-particle theory. The consequences are the same, however, since no picture or model of the microphysical domain is possible—the statistical interpretation obtains this by fiat. In addition, however, we saw that, in order to give a consistent account of measurement, defining physical phenomena with reference to an experimental arrangement was still required. So, in fact, the statistical interpretation assumes implicitly the same pragmatic epistemology, and is, therefore, subject to the same criticisms as the Copenhagen interpretation.

II EPISTEMOLOGICAL CONSIDERATIONS

The "realist" school of interpretations of quantum mechanics wishes to fit the quantum formalism into a more complete formalism which will be expressly realistic; that is, they wish to construct a theory which describes a physical reality existing independently of observation, but which reduces to the quantum mechanical formalism under certain conditions. The realists believe that there must always be a description available of a given ontological domain such that all elements of physical reality are simultaneously describable. Furthermore, they find it necessary, in order to achieve this in the microphysical domain, to make metaphysical assumptions about additional elements of reality, which are even in principle unobservable.

Hidden-variable theories, stochastic interpretations, the quantum potential approach, and the many-worlds interpretations all adopt a realist ontology in the microphysical domain. Non-local hidden variable theories (in which the quantum potential approach as well as, possibly, a non-local version of a stochastic theory can be included) and the many-worlds interpretation have not been excluded by Bell's theorem and its corresponding experimental evidence. As recent work, such as Jarrett's, has shown, however, any non-local theory of the microphysical domain (of which quantum mechanics is an example) needs to be incomplete, as Jarrett defined it. It is difficult to imagine how any theory can be Jarrett incomplete and still provide a realist picture.

The many-worlds interpretations is alone in escaping completely the results of Bell's theorem. Its privileged position comes at an enormous price, however. It does not simply postulate an additional element of physical reality, but, rather, an infinite number of almost identical realities. It is testament to the power of such metaphysical assumptions, and this may be the ultimate, that such unavoidable agreement with quantum mechanics can be obtained.

II.4 Motivation for the Present Prescription

We have constructed a critical epistemological framework from first principles appropriate for an analysis of the quantum realm. Whereas in the first chapter we showed how interpretational attempts have failed, here we have shown that they are even epistemologically unprepared to deal with the quantum realm. We have subjected the theory of quantum mechanics to our framework and found that the complete consistent knowledge we have of other ontological domains is impossible to obtain for the quantum domain. Through our analysis we came to realize that the consistency in our knowledge that we experience in most ontological domains is due to some underlying invariances of physical reality. Some invariant element(s) must then be lacking or faulted in the quantum domain. This is as much as we can say from a philosophical analysis; however, it provides a foundation and motivation for our physical analysis.

In modern physics the concept of invariance has played a major role. These invariances are found to correspond to symmetries, usually either geometrical or mathematical (viz. equations.) In one area, though, symmetry is found to be involved in a more fundamental ontological way; this is in the study of the elementary forces. Here symmetry is an abstract concept which takes the form of a high-level principle, in the sense that it directs our formation of certain physical laws. Epistemologically, it is very much a superior concept and fits very well with the idea of an underlying "element of form."

To understand the quantum-mechanical situation, then, we are encouraged to investigate the concept of symmetry and in particular the role it plays in the theory of elementary forces; i.e., gauge theory. This task will occupy the next two chapters.

III Symmetry

III.1 Ontology and Significance

Definition of Symmetry

When one speaks of symmetry, geometrical symmetry is usually thought of. However, the concept of symmetry is found to play a role in other mathematical realms also, from analysis of electrical circuits to differential equations to quantum field theory. Here we will try to define symmetry and develope a scheme for its application in the most general terms, so as to not unnecessarily restrict its ontological scope. Geometrical language, though, will remain a convenient language to develope these ideas. Some heady praise has been heaped upon the concept of symmetry. Hermann Weyl in his well-known book, Symmetry, went as far as to say, “all a priori statements in physics have their origins in symmetry.”[19] We will, in fact, utilize symmetry as the epistemological basis of our assertions, and, in conjunction with our previously defined philosophical framework, this will allow us to develope a consistent ontological foundation for microphysics.

In early Greek philosophy, the concept of symmetry can be found in ideas on proportions, vis a vis harmony, in the world. Today a concise, but rough definition of symmetry involves an invariance of a mathematical or geometrical object under a set of automorphic transformations. This “definition” fails at being a true definition for two reasons. Firstly, notice the word “involves;” that is, it does not even have the form of a definition. Secondly, our terms and their relation to one another are not defined. The latter failure can be fairly straightforwardly dealt with, and

[19] Weyl (1952), p. 126.

is what will consume the next section. The first, apparently trivial, problem of writing down a literal definition will be seen to be not so trivial and will be dealt with in stages, as insights into it are discovered during the next section.

First, let us choose the general term "system" to be substituted for a "mathematical or geometrical object".[20] For now we say no more in general about "the system" except that it is something that has properties, is the thing we are investigating, and that it needs to be defined for a particular case. As we proceed, this concept of system will become clearer. We use the term "state," A, to designate a possible "condition" of the system: this condition designating some or all of those attributes of the system that are accidental; that is, those attributes that can change without changing the definition of the particular system at hand. All other attributes are "permanent" and essentially make up the definition of the system. The set of those states all referring to the same set of accidental attributes form a "state space." (Notice from these definitions of state and state space that one can define many different state spaces for the same system.) We also define a subsystem as a system wholly contained within a system, in the sense that its permanent attributes form a subset of those of the system.

From here on we concern ourselves with states and state spaces. If an accidental attribute (or attributes) can be formulated as an equivalence relation ($\equiv$) between states—that is, satisfies the conditions of Reflexivity, $A_i \equiv A_i$ Symmetry, $A_i \equiv A_j \Leftrightarrow A_j \equiv A_i$, and Transitivity, $A_i \equiv A_j, A_j \equiv A_k \Rightarrow A_i \equiv A_k$—then this relation can be used to decompose the state spaces into equivalent subspaces (i.e., subsets of equivalent

[20] The following discussion roughly follows the treatment given by Rosen (1983), Ch. 3.

states.) An automorphic transformation, T, can now be considered as a one-to-one mapping of a state space onto itself; i.e., $T(A_i) = A_j$, for all A_i. If this transformation preserves equivalent subspaces for some equivalence relation (does not map any state A, into an inequivalent one), then we call this transformation a "symmetry transformation." S. We write

$$S(A_i) = A_j \equiv A_i$$

for all A_i.

We have now clearly defined our terms and, in the end, finally remet with the word symmetry. Extraction of a clearcut definition from this use, however, is still not easily done. Let us review what we have done. We started by reducing the concept of system to a space of states, a construct easily used to investigate the accidental properties of the system. We next allowed this space to be given a substructure via the definition of an equivalence relation. Then, under the effect of an automorphic transformation, which essentially exchanges states within a state space, we identified those that don't "violate the boundaries" of the equivalence relation as symmetry transformations.

Now, in general, there is a nontrivial set of symmetry transformations which preserve the subspace structure for a given equivalence relation. This set, in fact, turns out to form a group, called the symmetry group. This can be seen as follows. Firstly, these transformations are clearly associative. Secondly, since

$$S_1(A_i) = A_j \equiv A_i$$

and

$$(S_2S_1)(A_i) = S_2(S_1(A_i)) = S_2(A_j) = A_k \equiv A_j \equiv A_i,$$

we have closure. For the identity transformation, I, we have $I(A_i) = A_i \equiv A_i$, so the identity transformation is a symmetry transformation. Finally, if $S(A_i) = A_j \equiv A_i$, then $S^{-1}(A_j) = A_i$, but if $A_j \equiv A_i$, then $A_i \equiv A_j$, from the symmetry of the equivalence relation, so S^{-1} is a symmetry transformation.

We find, then, that given a system with a defined state space and specified equivalence relation, we uniquely determine a symmetry group (which is, in fact, a subgroup of the group of all the possible automorphic transformations on this state space.) However, from this statement, we see that we cannot ascribe a symmetry group directly to a system for two reasons. First, for a given system it is possible to construct many different state spaces. Second, the symmetry group we find is determined by the equivalence relation we choose, of which there are many. So, even before we try to determine how to apply the concept symmetry directly to a system, we find its use so far is a bit ambiguous in that we cannot ascribe a unique symmetry group to a system.

However, this does not mean we must be slave to this generality. If we consider similar systems, choose state spaces which concern the same attributes, and select the same equivalence relation, we can compare the symmetry groups of these systems in some way. To compare things we need an "ordering" of the things. Quantifying the symmetry groups (i.e., assigning quantities to symmetry groups, which we can call symmetry) would constitute an ordering of the symmetry groups. Of course, to do such a thing we need to consider the abstract groups related to the symmetry groups; then, if one group can be considered a proper subgroup of another, we can assign the latter a higher symmetry. Furthermore, if we are dealing with finite groups, then we can consider the order of the groups to quantify and order them. In this way, we can quantify and order a set of systems according to their symmetry. It is not clear, however, that this procedure will always unambiguously order the systems at hand. For each case it is necessary to choose the method of ordering carefully.

Let us now regress and, with the aid of these new definitions, give a clearer explication of the concept of system. We define the "structure" of a system as that aspect of the system that is invariant under any symmetry transformation. In other words, when we perform a symmetry operation on a state space of a system, that aspect of the system that does not change we call its structure. This definition is made with direct reference to the symmetry transformations of a state space of the system, so that, if we could find no equivalence relations on a state space of the system and, hence, no non-trivial symmetry transformations (an asymmetric system), we could say the structure of the system is trivial or, rather, that it has no structure at all.

Most systems are composite, that is, they have a "substructure." This leads us to consider in detail the nature of subsystems and to discover how the concept of symmetry can play a role in such an analysis. What do we mean when we say a system has a substructure? We can take any system, defined by its permanent attributes, and arbitrarily form a subsystem by taking a subset of these permanent attributes; however, if we are to be consistent with our earlier definition of structure, such a subsystem would not constitute a substructure unless it has a state space with non-trivial equivalence relations. Put more simply, a substructure is a symmetrical part (subsystem) of a system. Now, a system may have symmetrical subsystems (substructures) and yet be asymmetrical; that is, symmetrical parts may be put together to form a whole which is not as symmetrical as the parts. Conversely, a system may have substructures of lower symmetry than itself (or may have no substructures.) For a system with substructures, the distinguishing factor between these two general cases is the existence of equivalence relations connecting the substructures. If there exists no such equivalence relations, then the only symmetry transformations for the system will be those in common among the subsystems; the symmetry group of the system will be, therefore, just the intersection of the symmetry groups of the subsystems. We can write this as $G = \bigcap_i G_i$. We see that, in this case, the symmetry of the whole system must be equal to or smaller than that of any of its subsystems. We call this a heterogeneous system.

If the substructures are equivalent, the situation is more complicated. The new state space will, as before, support the symmetry transformations preserving common equivalence relations of the substructures (i.e., $\bigcap_i G_i$), but will now also allow symmetry transformations based upon equivalence relations between the substructures. We call this a non-heterogeneous (or homogeneous) system.[21] Let us list these new transformations (group elements): $g_1 = e, g_2, \ldots, g_n$. Now, $\bigcap_i G_i$ is a group and also a subset of the group elements of the new symmetry group, G, so it is a subgroup of G. We call it $H = \{h_i\}$. Now, list all right cosets of H with the new group elements:

[21] See Shubnikov and Koptsik (1974), pp. 328-350, for a discussion of such composite systems. Rosen (1983) does not consider non-heterogeneous systems.

$$\begin{array}{r}
h_1, h_2, \ldots, h_n \\
h_1g_2, h_1g_2, \ldots, h_1g_n \\
. \\
. \\
. \\
h_1g_n, h_2g_n, \ldots, h_1g_n
\end{array}$$

Examining this list, we see that no two elements in the same line can be equal, since then $h_ig_j = h_kg_j$, or $h_i = h_k$, a contradiction. Also no two elements in the first two lines are the same, since then $h_i = h_kg_2$,, or $g_2 = h_k^{-1}h_i \in H$, a contradiction, since g_2 is not a member of H. So all the elements listed in the second line must be "new" elements. This may exhaust all the new elements, but, if it does not, then by similar arguments another line constructed with a new element not found in the second line contains elements not found in the first two lines. At some point in this continuing process, we will exhaust all the new elements and the rest of the new lines will contain elements already listed. We see from this analysis that the number of new elements must be a multiple of the number of elements in H. This is, in fact, a statement of a famous theorem in group theory due to Lagrange,[22] which can also be stated as r (the order of H) is a divisor of r+n-1 (the order of G.) The trivial case n=1 (only the identity element) corresponds to the heterogeneous case.

The entire above listing may duplicate, but certainly exhausts all the elements of G; hence, we may write G as the union of these right cosets:

[22] See, for instance, Herstein (1975), pp.41-2.

$$\begin{aligned} G &= Hg_1 \bigcup Hg_2 \cdots \bigcup Hg_n \\ &= \bigcup_i Hg_i \\ &= \bigcup_i (\bigcap_j G_j) g_i \end{aligned}$$

Symmetry Conditions

The preceding discussion of the concept of symmetry was kept general; essentially, general in the sense of the concept of system used, in that this concept was mostly unrestricted. This was done so that—our problem being of a foundational nature—we remain as close to first principles as possible. In fact, we will maintain, in accordance with the remarks at the beginning of this chapter, that, although there may be equivalent parallel epistemological pathways, the concept of symmetry provides a direct and nitid pathway from first principles to foundational problems in physics.

As with all high-level concepts in science, symmetry is used to understand the constraints on the nature of the physical world. In particular, the concept of symmetry as applied to physical systems will allow us to discover an important principle, the symmetry principle. This principle will inform us of a specific constraint on the nature of physical reality, and, in doing so, will provide us with an important insight into our specific problem. We will investigate this principle next.

The Symmetry Principle

Consider a heterogeneous system containing two subsystems, A and B, which together constitute the entire system.[23] Next, define an appropriate state space for the whole system and, hence, state spaces for the two subsystems. We define a "causal relation" as follows. First of all, we note that choosing a state of the whole system determines the states of the subsystems. Now, we look for a correlation between the states of subsystem A and subsystem B. If we find that for every state of the whole system, the same state of subsystem A occurs with the same state of subsystem B (but that different states of A can appear with the same state of B), we say that A is a "cause subsystem" and B is the "effect subsystem", or that A causes B.

We now note that if this causal condition is met, there is a one-to-one correspondence between states of the whole system and states of the subsystem A. We can see this immediately since, first, as noted above, a state of the whole system determines the state of A, and, second, that since a given state of A determines the state of B and since subsystems A and B constitute the entire system, then the state of the entire system is determined by a state of A. The same cannot be said of B, of course, since different states of A can appear with the same state of B. So, at the logical level of state spaces, the state space of subsystem A is equivalent to the state space of the whole system. But, recall that for heterogeneous systems the symmetry of a subsystem is greater than or equal to the symmetry of the whole system; therefore, the symmetry of subsystem B must be greater than or equal to that of the system, and, hence, of subsystem A. In other words—substituting for the causal identities of the

[23] We again follow Rosen (1983), Ch. 4.

subsystems—the symmetry of the effect subsystem (or simply, the effect) must be greater than or equal to the symmetry of the cause subsystem (the cause.) This is a statement of the "symmetry principle."[24]

We can construct a more general derivation of the symmetry principle. The previous derivation, although rigorous, ignored much of the formalism we had just developed. Assume then that, naturally, our state spaces and equivalence relations for our system and subsystems have been chosen by virtue of the causal relation of interest; that is, the causal relation determines the equivalence relation of the cause—"cause equivalence"—and the equivalence relation of the effect—"effect equivalence." From this assumption and from our prescription for identifying our subsystems follows immediately an "equivalence principle": states of the cause subsystem which are cause equivalent must "yield" (i.e., appear with) states of the effect subsystem which are effect equivalent. In other words, states which are cause equivalent correspond to the "same cause" and a causal relation implies a unique effect: i.e., states which are effect equivalent. Also, we see that cause inequivalent states may yield effect equivalent states. If we now let the symmetry group of the cause be the symmetry group for cause equivalence and likewise for the effect, we find ourselves back at the symmetry principle, this time stated as: every member of the symmetry group of the cause must be a member of the symmetry group of the effect (but the effect symmetry group may contain elements not in the cause symmetry group.)

There are two ways in which the symmetry principle can be used: to set a lower bound on the symmetry of an effect; that is, if we know a cause—and its corresponding symmetry—we know the effect must have at least this much symmetry, or to set an upper bound on the symmetry

[24] This is our derivation. Rosen's derivation follows.

of a cause; that is, given an effect—and its symmetry—we must look for a cause that has no more symmetry than this (simplicity would like the cause to have the maximum permitted symmetry.) We can refer to the first as a minimalistic use and the second as a maximalistic use of the symmetry principle.

It is interesting at this point to take note of a parallel approach to this subject. George Birkhoff reformulated Leibnitz's "principle of sufficient reason" ("nothing happens without a sufficient reason") as follows:

> If there appears to be in a theory T a set of ambiguously determined (i.e., symmetrically entering) variables, then these variables can themselves be determined only to the extent allowed by the corresponding group G. Consequently any problem concerning these variables which has a uniquely determined solution, must itself be formulated so as to be unchanged by the operations of the group G (i.e., must involve the variables symmetrically).[25]

Although at first blush appearing not to have any connection with Leibnitz's principle, a little reflection shows it, in fact, to be a symmetry oriented extension of Leibnitz's simple but important principle. In addition, we see a minimalistic use of the symmetry principle here: the effect (i.e., the solution) has set upon it a lower bound on its symmetry (it must be "unchanged by the operation of the group G," the symmetry group of the cause.) We will discuss this connection further later.

Let us consider now the procedure by which symmetry and the symmetry principle can be used, still keeping our discussion general. Usually we are faced with looking for a cause given a certain effect. In this case,

[25] Birkhoff (1950), p.45.

we would use the symmetry principle in the maximal way. Very often, though, the effect does not exhibit a perfect symmetry but rather an approximate one. Our first task, then, is to identify the symmetry that is being approximated. We then assume that the cause has a similar structure; that is, it also "almost" has this symmetry, or rather, its "dominant" part has this symmetry, although, of course, it can never possess a greater symmetry than the effect.

We can make the discussion of this procedure more precise.[26] First we define an "approximate symmetry transformation," $T(u)$, such that

$$d(u, T(u)) \leq \epsilon,$$

where d is a metric on the state space with the properties

$$d(u,u) = 0$$
$$d(u,v) = d(v,u)$$
$$d(u,w) \leq d(u,v) + d(v,w).$$

In other words, d is an equivalence relation. With such an approximate symmetry transformation, then, we can obtain an approximate symmetry group, with ϵ providing a measure of the "goodness of approximation" of this group to the symmetry being approximated. Approximate symmetry is also sometimes called ”broken symmetry.” where ϵ would now be a measure of the symmetry breaking due to some symmetry breaking factor.

[26] See Rosen (1983), Ch. 5.

Situations involving approximate causal symmetry can be classified into three cases: stability, lability, and instability. In the case where there is stability, causal symmetry deviations (from some perfect symmetry) are "damped out": so that the approximate symmetry group of the cause is the minimal (exact) symmetry group of the effect. In the case of lability, the deviations from perfect symmetry in the cause are transmitted consistently: hence, the approximate symmetry group of the cause is the minimal approximate symmetry group of the effect. Finally, in the case of instability (also known as spontaneous symmetry breaking,) these deviations are amplified when transmitted, in such a way that the exact symmetry of the cause is the minimal (exact) symmetry of the effect. In this last case, the cause can quite often appear to have more symmetry than the effect, because the approximate symmetry group of the cause is larger than the symmetry group of the effect; however, the symmetry principle is not violated because the exact, or actual, symmetry group of the cause is not larger than the symmetry group of the effect.

Epistemological Considerations

The principle of symmetry as derived previously has the appearance of a mere tautology: it is a statement about the nature of cause and effect which uses the nature of the causal relation to define its specific constituents, namely, the cause and effect equivalence relations and the connection between them.[27] But, of course, we do not wish to add anything to the causal principle, but to better understand it and use it. In fact, as we have seen, the principle of symmetry is a useful analytical

[27] We leave Rosen's treatment here.

tool, and, given that it follows directly from the causal principle, should prove useful in an epistemological analysis.

The principle of sufficient reason, discussed above and identified as an application of the symmetry principle, can also be seen to be an explication of the causal principle. Basically, an establishing of the principle starts by assuming one has a problem which has a unique solution; i.e, one assumes a cause-effect relation. If the problem does not distinguish between two situations (causes), then the solution (effect) should be the same for these two situations. This observation, then, immediately yields the principle of sufficient reason (as stated above) which all problems must obey.

The conclusion that we draw from these two observations, then, is that we have found a direct connection between a high level a priori principle (arguably the highest level) and another principle with significant epistemological import. In fact, we can see that the entire content of causality has been translated into the symmetry principle and its attendant concepts: we identified the presence of a causal relationship with the existence of subsystems whose states were related to each other in very definite ways, the symmetry principle being a quantitative restriction on this relationship.

What we wish to take note of here is not the particular form of this restriction, but rather its nature as relating to symmetries. Imagine that it was claimed that these symmetries (of cause and effect) were somehow ill-defined in some instance (later we will give a suggestion as to how they may be ill defined.) This would be tantamount to saying that the cause and effect subsystems were ill-defined, since inherent in their identification is a realization of some definable symmetry. Consequently the existence of a causal relation would be brought into question. Of course given the a priori status of the causal relation (as stressed in the last chapter,) it is not possible to talk about the correctness of the causal relation, but rather it is traditionally only possible to say there either is or is not a causal relationship.

Let us now make connection with our discussion in the last chapter. There we claimed that some necessary invariances of the microphysical domain were lacking in such a way as to restrict our use of the causal principle in understanding this domain. The above connection between symmetry and the causal principle puts this claim on a stronger footing and also allows us to leave the awkward business of talking about the status of an a priori principle. Rather, we can concentrate our investigation on the concept of symmetry.

It is instructive, at this point, to compare the two derivations of the symmetry principle given earlier. Crucial for the first derivation were the assumptions that the system at hand was heterogeneous and that the cause and effect subsystems constituted the entire system. We can see, though, that the latter assumption is not so critical. Imagine a system satisfying the above two assumptions and a causal relation being demonstrated between the two subsystems. We can now easily imagine enlarging the whole system without destroying its heterogeneous nature or changing the two subsystems; certainly, then, the causal relationship between the two subsystems will not be destroyed and, hence, the validity of the equivalence principle will not be affected. Next, imagine a heterogeneous system which is larger than the union of two subsystems which have been established as cause and effect subsystems. If the system is truly heterogeneous, we can reduce the system to just the size of the union of the two subsystems, without effecting the logical relations between states, and, once again, the first derivation of the symmetry principle will be valid. So, it is really only the requirement that the system be heterogeneous that it is critical.

Our second deviation is stated in more precise language, but makes no explicit reference to the heterogeneous nature of the system. However, we recall that the requirement for a system to be heterogeneous is that it not contain equivalence relations connecting its substructures. By connecting the causal relation with the symmetry principle as the means by which the orders of the cause and effect subgroups were related, it is implicit that there should be no such direct connection between these structures; that is, in choosing the substructures of a system—as cause and effect—and the appropriate equivalence relations, allowing the presence of such connecting equivalence relations would belie the original intent of dividing the system in an analyzable manner (i.e., in terms of cause and effect.)

This last argument is valid only so long as this implied freedom of choosing subsystems, state spaces, and equivalence relations makes sense. The possibility exists that, when faced with certain systems, choices of state spaces and equivalence relations and even subsystems may not be as rich as need be. In this case, desiring identification in a system of cause and effect subsystems may force us to analyze our system in a non-heterogeneous way; i.e., in a way that might lead us to say the associated symmetries are ill-defined. In this case, we might very well expect confused results in terms of a symmetry principle and, once again, results difficult to interpret in terms of cause and effect.

III.2 Symmetry and Physical Systems

We will now study how the concept of symmetry and the ideas we developed in the first section of this chapter are applied to isolated physical systems.[28] We first need to identify those aspects of physical systems which correspond to the abstract notions developed in the first section. Still, within this narrower context, we will try to keep our discussion as general as possible, until we find it necessary to further narrow our scope.

Our general abstract concept of system, then, we identify with physical dynamical process. The causal relations we find in physical processes are, of course, the laws of nature, and the cause and effect subsystems are, correspondingly, the initial and final conditions, respectively.

The familiar concept of an "isolated physical system" (i.e., a group of bodies with properties) will not be the symmetry-oriented construct of system that we have been discussing. Rather, it is the dynamic process that we call our "system" and whose structure, substructures, and their interrelations that we will be investigating; however, we will find these two concepts of system to be closely related.

The relation between physical system and our logical system (process) is as follows. First, we note that a given physical system can support different processes, but also a given process can take place in different physical systems, so neither concept logically includes the other. We are familiar with what is a state of a physical system. The dynamical process state space is more subtle—it is the space of all possible specific processes for a given dynamical process. There turns out to be a one-to-one correspondence between a given process state space and the physical state space of a physical system: a state of the cause subsystem

[28] See Rosen (1983), Ch. 6.

is a physical state (a specific set of initial conditions) and, correspondingly, every physical state can serve as a cause subsystem state and, further (as we noted in the first section of this chapter), there is a one-to-one correspondence between states of the whole system and states of the cause subsystems (for heterogeneous systems.) Consequently, all the transformation properties—i.e., equivalence relations, etc., of our system (process) space—will be identical for physical state space.

The obvious next step is to carry over our results of the first section of this chapter (namely, the symmetry principle) to processes and, hence, physical systems. Before we do this, however, we need to discuss the nature of our causal relations here; i.e., the laws of nature. A law of nature can be thought of as a mapping—a (temporal) mapping of states from the cause subsystem to the effect subsystem (any general causal relation could be interpreted this way, however, such an interpretation is more easily motivated in the physical case.) We can represent such a mapping as $u \xrightarrow{N} Nu$ where N represents our law of nature mapping, u is a state of the cause subsystem, and Nu is a state of the effect subsystem. Now, these states are also physical states (initial and final conditions), so this mapping is also a mapping of physical state space into itself.

Consider, then, a process and some physical (non-temporal) transformation, T, on the state space of this process; i.e., an invertible mapping $u \xrightarrow{T} Tu$, for all states, u. We can apply this transformation to final states of this process; that is, we can write

$$Nu \xrightarrow{T} TNu$$

for all u. Further, let us consider the temporal development of states Tu; i.e.,

$$Tu \xrightarrow{N} NTu$$

for all u.

Now we assume T is a symmetry of our law of nature, by which we mean that it will map cause subsystem states to cause equivalent states and effect subsystem states to effect equivalent states (i.e., it acts as a symmetry transformation in these spaces.) It then follows that the states on the right-hand sides of the above two relations must be equivalent; i.e.,

$$TNu \equiv NTu$$

for all u—or, equivalently,

$$TN = NT.$$

In words: the transformed result of a process is the same as the result of a transformed process. If the above is true of a transformation, T, and our system is heterogeneous,[29] it will also, via our correspondence between physical and logical state spaces, be a symmetry transformation

[29] These results will not necessarily hold for non-heterogeneous systems, as we will discuss later.

on physical state space (in the general sense of a symmetry transformation.)

The preceding discussion and consequent definition of a symmetry transformation of a law of nature provides us with a connection between a physical symmetry and the abstract notions of symmetry associated with causal systems. We can extend this connection to the concept of symmetry groups. The symmetry group of a law of nature, N, is simply defined as the group of all invertible transformations which obey the above commutation relation. By our above analysis, we see that all members of this group will map any state of the cause subsystem into a cause-equivalent state. But, also, we see that they will map all cause-equivalent states to one-another, otherwise two cause-equivalent states could be considered physically inequivalent. The symmetry group of N is then isomorphic to the symmetry group of the cause (we could have seen this immediately from the one-to-one correspondence between the state space of the whole system, the cause subsystem, and the physical system.)

Similarly, all members of the symmetry group of N will map any state of the effect subsystem into an effect-equivalent state; however, in this case there is no reason why there may not be states which are effect-equivalent and which are not mapped into one-another by a symmetry transformation of N. In fact, we know that cause-inequivalent states can evolve into effect-equivalent states. So, the symmetry group of N is a subgroup of the symmetry group of the effect.

The above two results yield two principles for physical state spaces which are parallel to those for logical state spaces. First is an equivalence principle: equivalent physical states must evolve into equivalent states while inequivalent physical states may evolve into equivalent ones. Second is a symmetry principle: for an isolated physical system, the degree of symmetry cannot decrease, but either remains constant or increases. This is called the "general symmetry evolution principle." It should be noted here that one symmetry of the laws of nature is special and does not follow the above analysis; that is time-reversal symmetry. This is because this symmetry acts on the evolution transformation, N. As before, we can consider the transformed (time-reversed) initial state

$$u \xrightarrow{t} tu,$$

where t stands for the time reversal transformation. And once again we can apply this transformation to a final state

$$Nu \xrightarrow{t} tNu.$$

However, to consider the evolved resultant of the transformed initial state, we must take into account that the process is time-reversed; in other words, we must map with the inverse of N, N^{-1}:

$$tu \xrightarrow{N^{-1}} Ntu.$$

If t is a symmetry transformation of N, then we have

$$tNu \equiv N^{-1}tu.$$

for all u with the obvious necessary condition that N^{-1}exists (i.e., N is one-to-one and onto), or we may write

$$tN = N^{-1}t.$$

We also note that it is possible that a symmetry transformation may only hold for a subspace of a physical state space, so that the relation

$$TNu \equiv NTu \; (or \; tNu \equiv N^{-1}tu)$$

may be true only for all u belonging to this subspace.

The general symmetry evolution principle is “general” because it follows directly from general considerations; it does not include any further specific assumptions. This principle, however, has a drawback from a practical point of view: it concerns the entire state space of a system whereas in analysis of particular physical systems, we usually follow the evolution of a particular initial state. In terms of processes, this principle concerns the space of many possible processes, not a particular one of interest.

We define the symmetry group of a particular state as the group of those transformations on physical state space which take this state into an equivalent one. This group is isomorphic to the group of permutations of all states belonging to the same equivalence subspace as this state; hence, the degree of symmetry of a state is simply measured by the population of the equivalence subspace of this state. The usefulness of this definition will be demonstrated if we can derive a symmetry principle for this symmetry. This will require, however, a special assumption which will restrict the scope of applicability of the final result. What we must assume is that different states evolve (are mapped by N) into different states. This is called the assumption of "non-convergent evolution." With this assumption, we see that the number of states equivalent to a state cannot decrease since all states which are initially equivalent (cause-equivalent) must evolve into final-equivalent (effect-equivalent) states (by the equivalence principle) and no two states can evolve into the same state. Of course, other entire equivalence subspaces may evolve into the final equivalence subspace of this state (again, by the equivalence principle.) This leads us to "the special symmetry evolution principle:" "the degree of symmetry of the state of an isolated system cannot decrease during evolution but either remains constant or increases." The assumption of non-convergent evolution is generally true for microscopic processes. However, when such processes are considered macroscopically (i.e., by statistical methods), this assumption is generally not true, since, for such macro-states, many different initial states can lead to the same final state. In fact, such macro-states can be considered as equivalent subspaces of micro-states and, of course, such subspaces can converge.

If, for a given process, N is time-reversal symmetric, we can say something special about the symmetry of evolving states, namely, that states must evolve with a constant degree of symmetry. In other words, equivalence subspaces cannot converge (otherwise the inverse process would involve a decrease in symmetry.) In fact, we can see that, more generally, the existence of an inverse for N guarantees evolution with constant symmetry.

Conservation Laws

We now introduce a concept associated with physical systems that is intimately connected with the concept of symmetry; this is the concept of a conservation of the laws of nature, which occurs when some property of the states of a physical system does not change in time as the system evolves according to the laws of nature; i.e, we can write

$$Q(Nu) = Q(u),$$

where Q denotes the conserved property.

Any well-defined property of states, call it Q, naturally leads to a decomposition of state space into equivalence subspaces, each having a certain value of the quantity associated with this property. The associated equivalence relation, of course, then defines a symmetry group on state space. Conversely, a symmetry of a physical system is associated with a symmetry group on state space, which defines an equivalence relation, which decomposes state space into equivalence subspaces, which yields a labeling of states with a property; i.e., according to their membership in a subspace.

Now, if the physical symmetry associated with a symmetry group is a symmetry of a law of nature in some system, then the temporal evolution mapping, N, must preserve the associated equivalence subspaces; hence, N will be a member of this symmetry group. In this case, $Q(Nu) = Q(u)$, and this property will be conserved. We have found, very easily, a connection between symmetries of the laws of nature and conserved properties; namely, if a symmetry of state space is also a symmetry of nature, then there will be a conserved property of such states. Let us investigate this connection a little more carefully and see how and under what conditions a symmetry of a law of nature will lead to a conserved quantity.

We see that N must have the properties that we demand of the symmetry transformations on state space if it is to be a member of the symmetry group on state space. (We might more properly turn this restriction around: we could say that only those symmetry groups which contain the temporal development mapping represent symmetries of nature which lead to conserved quantities.) This of course restricts N to be invertible (for infinitesimal (continuous) transformations it need only be infinitesimally invertible.) We saw earlier that the existence of N^{-1} guaranteed evolution with constant symmetry. Without this condition (if we were considering a semigroup of symmetry transformations,) we could have what we might call "partial conservation": states belonging to a given equivalence subspace (characterized by a certain value of a quantity) could not evolve into inequivalent states (states with different values of the quantity), but could evolve into an equivalence subspace with a larger number of states (which all must now be characterized by one value) due to convergence of equivalence subspaces.

Let us now begin by assuming we have identified a conserved quantity; i.e., $Q(Nu) = Q(u)$. This means that the equivalence subspaces of state space determined by this quantity are preserved by the temporal development mapping; hence, N must be a member of the symmetry group of state space associated with the above equivalence relation; that is, it must be a symmetry transformation of state space. This, of course, makes the symmetry associated with this symmetry group also a symmetry of the laws of nature. So we find that to every conservation we can ascribe a symmetry of the laws of nature.[30] If we had instead begun with only partial conservation (as defined earlier), then N would not need to be invertible, but we would still be lead to a symmetry of the laws of nature.

We need to note that there is the possibility that the property we have identified as conserved is not assigned uniquely to all states; that is, some states may be "undetermined" with respect to this property. The symmetry group discussed above would not then, in general, be a symmetry group of the entire state space of our system and this symmetry could not be called a symmetry of the laws of nature. We add the qualifier "in general" above because it still may be possible that the symmetry group could be extended in some consistent manner to include mappings of these undetermined states, so that it would be a symmetry group of entire state space. We will see an example of this possibility later.

Let us summarize the above connections between symmetry and conservation. If a transformation group on state space is a symmetry group of this space and this symmetry is also a symmetry of the laws of nature for the associated system, then there will be a uniquely conserved property in this system. A necessary and sufficient condition for a symmetry on state space to be a symmetry of nature is that the temporal development mapping, N, be a member of the symmetry group.

We discovered in the last section, however, that for heterogeneous systems, the one-to-one correspondence between state and process space implies a similar correspondence between state and cause and effect subspaces, which, in turn, implies that, if there is a symmetry group defined on state space, there must be an isomorphic symmetry group on process

[30] Subject to the condition noted below.

space. The symmetry associated with this symmetry group is then a symmetry of nature. So, an equivalent sufficient condition for a symmetry on state space to be a symmetry of nature (or vice-versa) is that the system be analyzed in a heterogeneous manner.

We also found that if we discover a well-defined property of physical states (such that we can assign all states to unique equivalence subspaces in a consistent manner) which is conserved (or even only partially so) by the laws of nature operating in this physical system, then there will be a symmetry of the laws of nature uniquely associated with this property.[31]

Following Rosen and Freundlich (1978), we consider two general classes of physical systems. First, we consider systems describable by linear vector spaces. Transformations on such a space can be in the form

$$Mu_i, = \mu_i u_i,$$

where such states, u_i, can be considered eigenstates of M with eigenvalues μ_i, although there may be states which are not eigenstates of M. In this way the mapping M defines a property of states: they either have a unique value μ, or are not assigned a value.

We now assume the mapping M is a symmetry of the laws of nature of this system. We can then write $MN = NM$ and apply this equation to an eigenstate of M, u_i:

[31] The general symmetry-conservation formalism discussed above was suggested, but not carried through, by Rosen (1980).

$$\begin{aligned} MNu_i &= NMu_i \\ &= N\mu_i u_i \\ &= \mu_i N u_i. \end{aligned}$$

We have shown that if u_i is an eigenstate of M with eigenvalue μ_i, then so also is Nu_i. Consequently, equivalence subspaces of those states which are eigenstates of M are preserved by N. What about those states which are not eigenstates of M? Let us assume that the evolved state Nu is an eigenstate of M with eigenvalue μ, but that u is not an eigenstate of M. Then,

$$\begin{aligned} NMu &= MNu \\ &= \mu N u \\ &= N\mu u. \end{aligned}$$

Using the invertibility of N, we multiply both sides of the above equation by N^{-1}. We find

$$Mu = \mu u,$$

in contradiction with our assumption. So all states which are not eigenstates of M will remain so. Hence, this property which is assigned to states by the symmetry mapping M is conserved.

If we were considering non-invertible mappings (i.e., symmetry semigroups,) however, then the last step in the above demonstration could not go through. In this case, states which are not eigenstates of M can evolve into those which are. This is just the situation we described in our general formalism earlier where we called this partial conservation. Here, states which are eigenstates of M have the property labeled by μ, conserved, but those states which are not eigenstates of M we can say nothing about.

We next start by assuming we have at least partial conservation on our linear vector space for the quantity associated with a mapping M with eigenstates u_i:

$$MNu_i = \mu_i N u_i;$$

that is, Nu_i is also an eigenstate of M with the same eigenvalue as u_i. We then find

$$\begin{aligned} MNu_i &= N\mu_i u_i \\ &= NMu_i. \end{aligned}$$

We can deduce a symmetry of nature from this last relation provided that the u_i form a complete set, since this will assure us that the above relation will be true for all states on state space (a requirement that we noted above in our general discussion.)

The second class of physical systems we consider are describable by continuum state spaces; i.e., these are systems whose temporal development mapping is continuous: $N(t, t_1)u_1$, is a mapping of state u_1, at time t_1, to a new state at some later time t. We can say immediately, then, that only symmetries which are continuous can yield conservations for these systems, since only groups of continuous symmetry transformations can contain a continuous temporal development mapping. Of course such transformations, in contrast to discrete transformations, can be represented by infinitesimals. The existence of inverses for such transformations is guaranteed.

We again follow Rosen and Freundlich (1978), except we use a slightly simpler and less general treatment which will suffice for our purposes. We first assume a generalized form of Hamilton's principle: that the action,

$$I = \int_{t_1}^{t_2} dt L(N(t, t_1)u_1, t) \tag{III.2.1}$$

—where L is the Lagrangian considered as a function of state and time—under arbitrary variations of the mapping N, and for all u_1, t_1, t_2, depends only on the situation at the endpoints of the integral. Consider, then, an arbitrary infinitesimal variation of N, so that

$$\delta I = \int_{t_1}^{t_2} dt \delta L, \tag{III.2.2}$$

where

$$\begin{aligned}\delta L &\equiv L(N'(t,t_1)u_1,t) - L(N(t,t_1)u_1,t)\\ &= \epsilon[\frac{d}{d\epsilon}L(N'(t,t_1)u_1,t)|_{\epsilon=0}]\\ &= \epsilon(\frac{d}{dt}\Pi + F). \end{aligned} \tag{III.2.3}$$

In the last line we assume we can in general write δL as the sum of a term which can be written as a total time derivative and another which does not in general contain such additive terms. Hamilton's principle then implies that

$$F \overset{\circ}{=} 0. \tag{III.2.4}$$

This is just the equation of motion, and $\overset{\circ}{=}$ indicates "equality if and only if the equations of motion are satisfied."

Now, if Hamilton's principle can be satisfied without recourse to the equations of motion for a particular variation; i.e., if

$$\delta L = \epsilon\frac{d}{dt}R, \tag{III.2.5}$$

then it follows that

$$\frac{d}{dt}Q \overset{\circ}{=} 0, \tag{III.2.6}$$

where $Q = Q(N(t,t_1)u_1,t) \equiv R - \Pi$, and we have conservation of the function Q.[32] This can also be put into our definitional form: $Q(N(t,t_1)u,t) \mathrel{\mathring{=}} Q(u_1,t_1)$. So, those variations which satisfy the condition (III.2.5) lead directly to a conservation. If applied to classical field theory, this result would reduce to the Bessel-Hagen extension of Noether's theorem in classical field theory.

We next consider symmetry transformations. As usual, consider a mapping M of state space onto itself. We will, of course, consider infinitesimal mappings here. Now, for M to be a symmetry of a law of nature in these systems, it must, as before, commute with N:

$$MN(t,t_1) = N(t,t_1)M, \tag{III.2.7}$$

so that the temporal development of a symmetry transformed state is equivalent to a particular infinitesimal variation of the temporal development itself, which we can write as

$$N' \equiv MN(t,t_1). \tag{III.2.8}$$

Hamilton's principle must then hold for the transformed action,

$$I = \int\limits_{t_1}^{t_2} dtL(N'(t,t_1)u_1,t). \tag{III.2.9}$$

If we expand this integral to first order in the variation, we find

[32] In the following, Π and all capital letters following O are to be considered functions of state and time.

$$I = \int\limits_{t_1}^{t_2} dt[L(N(t,t_1)u_1,t) + \delta L]. \tag{III.2.10}$$

We again allow an arbitrary variation of N (which we will indicate by δ'):

$$\delta' I = \int\limits_{t_1}^{t_2} dt[\delta' L + \delta' \delta L]. \tag{III.2.11}$$

Hamilton's principle can be satisfied in one of two ways; either

$$\delta L = \epsilon \frac{d}{dt} R \tag{III.2.12}$$

or

$$\delta' \delta L \mathrel{\hat{=}} \epsilon \frac{d}{dt} U. \tag{III.2.13}$$

The first way yields a conservation as before of $Q = R - II$. The second possibility, however, does not yield a conservation. In fact, the situation is much the same as before we applied the symmetry transformation to our system: the existence of a symmetry transformation does not necessarily lead to a conservation, although for some symmetry transformations we do have a formula for connecting the transformation to a conserved quantity. This situation may seem, at first, to conflict with our general formalism and also with conventional Noether's theorem. In the familiar, conventional Noether's theorem all continuous symmetries are associated with conservations; however, conventional Noether's theorem simply defines a symmetry transformation to be one which leaves the action invariant. The first condition (III.2.12) is then automatically satisfied. What we realize here, in comparing with our general formalism, is that we have not demanded that the symmetry transformation of our laws of nature is also a symmetry of physical state space. It must be, then, that the transformations defined by condition (III.2.13) (which, we note, is also a condition involving the Lagrangian) are not symmetry transformations on state space and, hence, force a non-heterogeneous analysis of our system.[33]

These results have their most obvious application in classical field theories (and, in fact, similar results have been found for these theories,) but, of course, have a broader scope only limited by the fundamental physical assumptions made at the beginning of the discussion. As such, this result could be called a meta-Noether's theorem. We also note that these results can be further generalized to allow for symmetry variations which vary and depend on the time of the object state, but the basic connections between symmetry and conservation remain the same for such generalizations.[34]

We can also obtain an inverse meta-Noether's theorem. If we start by assuming conservation of Q (i.e., equation (III.2.6)), then we can write

$$\frac{d}{dt}Q = f(F), \qquad \text{(III.2.14)}$$

[33] This will be shown explicitly for field theories in the next chapter. Rosen and Freundlich (1978) and Rosen (1980) do not come to this conclusion since they do not consider non-heterogeneous systems.

[34] Rosen and Freundlich (1978) and Rosen (1980) obtain these generalizations.

where $f(0) = 0$. Now, if we can find a variation of N such that

$$f(F) = F, \tag{III.2.15}$$

then we immediately obtain equation (III.2.12) with the help of the definition (III.2.3). We can now reverse our reasoning used to obtain meta-Noether's theorem: equation (III.2.10) and, therefore (III.2.9), must be valid actions; therefore, we can write down equation (III.2.8) and identify this variation as a symmetry of the equation of motion.

Now the condition (III.2.15) is a relation between the functional form of our conserved function on state space, Q, and the equations of motion. We can interpret this condition in light of our general formalism as the condition that Q assigns values to states in a unique and consistent manner. Inability to satisfy this condition would suggest that this conserved property is in some sense an "ill-defined" one.

We finally note that we can always associate a variation of L with a conserved quantity, even though it may not be a symmetry transformation, as long as $f(F)$ contains a linear term in F. Write equation (III.2.14) as

$$\begin{aligned}\frac{d}{dt}(P - S) = f(F)\\ \frac{d}{dt}P = g(F) + aF + \frac{d}{dt}S\\ \epsilon\frac{d}{dt}P = \delta L + \epsilon g(F),\end{aligned} \tag{III.2.16}$$

where a is some constant and $g(F)$ is like $f(F)$ except containing no linear terms in F. Now δL becomes equal to a total time derivative of a function of state and time upon satisfaction of the equation of motion; therefore,

δL is in fact a valid variation of the Lagrangian. Such transformations are generally called "Noether transformations." We also note that δL could in general contain a term of the form $\epsilon g(F)$, so we may redefine δL as

$$\delta L \equiv e(\frac{d}{dt}\Pi + g(F) + F), \tag{III.2.17}$$

so that the class of Noether transformations becomes identical to those transformations leading to equation (III.2.5).[35] With this more general δL, satisfaction of condition (III.2.15) requires

$$g(F) + F = \frac{d}{dt}T. \tag{III.2.18}$$

If $f(F)$ contains no linear term in F, then it cannot be related to a variation of L.

III.3 Symmetry as an Abstraction Tool

Although the concept of symmetry in physical systems is most commonly thought of as a useful tool in tackling specific problems, we have found here a much more fundamental role for this concept. We have, in the previous pages of this chapter, developed an intimate connection between the concept of causality and concepts of symmetry. Just as in mathematical fields such as geometry and group theory symmetry is used as a means of abstraction to gain a deeper and broader understanding of those fields, here symmetry has been used as an abstraction tool,

[35] In other words, those transformations associated with a conserved quantity. The relevance of these transformations will become apparent in the next chapter when we discuss symmetry-conservation formalism in classical field theory.

translating the epistemological content of the causal relation into a more abstract and tractable form. We have also found a direct connection between these abstract ideas and the symmetries of physical systems. Furthermore, we demonstrated a simple and direct connection between these two concepts and the fundamental concept of a conserved property. This concept of conservation is so fundamental because it is those properties which are conserved (or sometimes only partially so) that are identified as fundamental properties. Our general approach at the beginning of this chapter lead us to consider a class of systems that are usually not considered as candidates for physical systems; namely, non-heterogeneous systems. In fact, we found that our two main results, the symmetry principle and meta-Noether's theorem—both intuitively straightforward results—breakdown for such systems. We already noted how in the case of the symmetry principle such a breakdown indicates a flawed application of the causal relation. In a similar way, breakdown of a symmetry-conservation formalism indicates a flawed defined property, which could lead to a flawed conception of object. We have again, then, made connection with our remarks at the end of Chapter 2, where we predicted that inconsistency in some underlying element of symmetry could lead to a confused concept of object.

IV Gauge Theory

IV.1 The Action in Field Theory

The Action functional is a useful tool for both classical and quantum physics. In quantum field theory it is used either in the S-matrix approach or in the Feynman Path Integral approach to yield physical quantities. Its significance lies in its direct connection to measurable quantities, its constructive and exact (not approximate) nature, and the connection it provides with classical concepts and results. Classically it generates the canonical transformation which takes the canonical variables from one time to another. The Action can be written as

$$I = \int_{\tau_1}^{\tau_2} d^4xL[x] \tag{IV.1.1}$$

where L is the Lagrangian density. We can construct the possible Lagrangians for field theory by placing certain demands or restrictions on the Action.[36] First, in order that our physics obeys the postulates of special relativity, we require that the Action be Poincaré invariant. In particular, a necessary and sufficient condition for translation invariance is that the Lagrangian be a function of the fields and their derivatives only.

Second, we require that the Lagrangian depends on the fields only at one space-time point. This constrains us to local field theories.

Thirdly, the Action must be real. This demand carries over from classical physics where imaginary terms lead to non-conservation of probability.

[36] See, for instance, Ramond (1981), Section I.5.

Fourth, we require that the Action leads to (classical) equations of motion containing derivatives no higher than second-order. Higher order derivatives can lead (classically) to non-causal solutions. We can satisfy this condition by only allowing one or two ∂_μ operators in a term in the Lagrangian.

In addition to these restrictions on the Action, we will later introduce ad hoc symmetry requirements which will fix the dynamical properties of the fields.

Placing these four demands on the possible forms of the Lagrangian, we can write down the most general Lagrangians.

The requirement of Lorentz invariance divides up our problem into distinct possibilities. Our canonical variables are fields, which are continuous functions of the coordinates. These fields then must have definite transformation properties under the action of the Lorentz group: that is, they must transform under a definite representation of the Lorentz group. They can transform either as scalars, spinors, vectors, tensors, etc. These different transformation properties distinguish particles according to their spin. Fields which transform as scalars under Lorentz transformations describe particles with spin-0. Those which transform as spinors describe particles with spin-1/2. Those which transform as four-vectors have spin-1, and those which transform as higher rank tensors and/or spinors have higher spins.

In our analysis of possible field theories, we will be considering the fields as classical; i.e., as canonical coordinates which are functions of the space-time coordinate. We can apply the methods and analysis of classical mechanics here. Final justification for any results and their interpretation, of course, comes from quantizing the theory (as, for instance, using the Feynman path integral approach) and its renormalization. We will note where necessary what modifications need to be made to these results for a valid quantized theory We will be, naturally, stressing those aspects concerned with symmetry.

Symmetry and Conservation

Before proceeding to consider the possible field theoretic Actions, we consider one more general property of the Action, namely Noether's theorem. Instead of proceeding from our general result of the last chapter, it is more instructive to proceed from prior principles; although we will, of course, find that our results will parallel those that we found for general continuous temporal development. We will again find that there are symmetry transformations that do not lead to a conservation and that there are Noether transformations (transformations associated with conservations) which are not symmetry transformations. Two questions will arise due to the more specific nature of our problem: how do we explicitly affect a transformation for a given system, and how do we define a symmetry transformation?

Our systems are defined by the Action. Since it is written as an integral of a function of the fields and their derivatives (the Lagrangian), it can be made to vary either through a change in the integration measure (due to an implicit space-time coordinate variation) or through an explicit variation of L (usually called a form variation) or a variation in L induced by a variation of the fields. The variation of the fields, in turn, can either be explicit or induced by a variation of the space-time coordinates.

We also note that since we are using a manifest relativistic treatment, a "conservation condition" will not involve only one of the coordinates specially (i.e., the time); rather, what is conserved for field theories is a four-current and the conservation condition is a continuity equation. However, if certain space boundary conditions are imposed (namely, that the fields go to zero at infinity), then time constants can be identified (these are the charges.)

In the following $\phi(x)$ is any local field or collection of fields (all field indices are suppressed) and $x = (x^\mu), u = 0, 1, 2, 3$ are our space-time coordinates. The summation convention is assumed.

Let us begin by putting together the above possible means of varying the Action and consider its most general variation. Consider the following infinitesimal transformations:

$$x \to \bar{x} = x + \delta x \tag{IV.1.2}$$

$$\phi(x) \to \bar{\phi}(x) = \phi(x) + \delta_0\phi(x, \phi(x), \partial_\mu\phi(x)), \tag{IV.1.3}$$

where δ_0, is the explicit, or functional, change; i.e., the change at one space-time point (hence ∂_μ commutes with δ_0),[37] and

$$L(\phi(x), \partial_u\phi(x)) \to \bar{L}(\phi(x), \partial_u\phi(x)) = L(\phi(x), \partial_u\phi(x)) + \delta_0 L(\phi(x), \partial_\mu\phi(x)). \tag{IV.1.4}$$

Let us write this explicit change in L as

$$\delta_0 L = \partial_\mu \delta_0 L_1^\mu + \delta_0 L_2, \tag{IV.1.5}$$

[37] See Boyer (1966) for a discussion of the importance of this variation.

thus separating the variation into a divergence and a non-divergence term. The total change in L is then

$$\delta L = \delta_0 L + \delta x^\mu \partial_\mu L + \delta_0 \phi [\partial_\mu \Pi^\mu + EL], \tag{IV.1.6}$$

where E are the Euler-Lagrange operators:

$$E = -\partial_\mu \frac{\partial}{\partial[\partial_\mu \phi]} + \frac{\partial}{\partial \phi},$$

and

$$\Pi^\mu \equiv \Pi^\mu(L) = \frac{\partial L}{\partial[\partial_\mu \phi]}.$$

(Both E and $\Pi_\mu(L)$ can easily be generalized for Lagrangians containing higher order derivatives of the fields.[38]) The change in the integration measure given by

$$d^4x \to d^4\bar{x} = d^4x + \delta(d^4x) = d^4x + d^4x\partial_\mu \delta x^\mu.$$

The infinitesimal variation in the Action is then

[38] See, for instance, Boyer (1967) and Rosen (1972). In addition, relating to the following discussion of Noether's theorem, see Rosen (1974a, 1974b), Palmieri and Vitale (1970), Candotti et al. (1970, 1972) and references therein.

$$\begin{aligned}\delta I &= \int\limits_{\bar{V}} \bar{L}(\bar{\phi}(\bar{x}), \partial_\mu \bar{\phi}(\bar{x}))d^4\bar{x} - \int\limits_{V} L(\phi(x), \partial_\mu \phi(x))d^4x \\ &= \int\limits_{V} [\delta L] d^4x\end{aligned} \qquad \text{(IV.1.7)}$$

where

$$[\delta L] = \partial_\mu[L\delta x^\mu + \Pi^\mu \delta_0 \phi + \delta_0 L_1^\mu] + \delta_0 \phi EL + \delta_0 L_2 \qquad \text{(IV.1.8)}$$

Equation (IV.1.8) is a formal expression for the most general variation in the Action due to the arbitrary variations (IV.1.2), (IV.1.3) and (IV.1.4). For particular variations and particular Actions, it may be possible to write $[\delta L]$ in an essentially different way than that appearing in (IV.1.8), and this of course is essential to Noether's theorem, as we will see later. We now apply our generalized Hamilton's principle to equation (IV.1.8). Hamilton's principle involves an arbitrary variation of the canonical coordinates only—the fields in this case. It is also the principle for deriving the field configurations for a given system —given L—and so a form variation of L would make no sense. Hamilton's principle then says that the physical field configurations are those for which the variation of the Action (under the above condition) depends only on the situation at the boundary surface of integration. We then find immediately from equation (IV.1.8) the equations of motion:

$$F \equiv EL \overset{\circ}{=} 0 \qquad \text{(IV.1.9)}$$

—where again $\stackrel{\circ}{=}$ indicates equality if and only if the equation of motion hold (i.e., when ϕ is set equal to the physical field configurations)—as a necessary and sufficient condition for Hamilton's principle to hold.

Noether Transformations

We now proceed to find the most general class of Noether transformations—those transformations which are variations of the Action leading to a continuity equation. These can be found, in fact, by demanding that the variation of the Action satisfy Hamilton's principle without recourse to the equation of motion, except that we do not restrict the variations to field variations, but we must make the restriction that

$$\delta_0 L_2 \stackrel{\circ}{=} 0; \tag{IV.1.10}$$

that is, the form variation of the Lagrangian is restricted in such a way that its non-divergence part must go to zero upon satisfaction of the equations of motion. We can see this as follows. Let[39]

$$[\delta L] = \partial_\mu k^\mu \tag{IV.1.11}$$

[39] Candotti et al. (1970) generalized the possible variation of $[\delta L]$ to include an additional term which goes to zero when the equations of motion are satisfied—i.e., $[\delta L] = \partial_\mu k^\mu + k_2$, where $\mathrm{k}_2 \stackrel{\circ}{=} 0$—and still obtained Noether transformations. Rosen (1972) adopted their generalization and further generalized the approach by including the term $\delta_0 L_2$ above. However, Rosen failed to realize his generalization made that of Candotti et al. redundant. We can see this by noticing that the term k_2 above can be absorbed into $\delta_0 L_2$ without any loss of generality, since both $\stackrel{\circ}{=} 0$. This disallowed Rosen and Candotti et al. from discovering the connection between the most general Noether transformations and Hamilton's principle, leading them to conclude that there is no general connection between Noether transformations and symmetry transformations.

so that

$$\partial_\mu Z^\mu = g, \tag{IV.1.12}$$

where

$$Z^\mu = L\delta x^\mu + \Pi^\mu \delta_0 \phi + \delta_0 L_1^\mu - k^\mu \tag{IV.1.13}$$

and

$$g = -\delta_0 \phi EL - \delta_0 L_2. \tag{IV.1.14}$$

Condition (IV.1.10) gives us

$$\partial_\mu Z^\mu \mathrel{\mathring{=}} 0. \tag{IV.1.15}$$

The situation is actually much simpler than it appears above. As we indicated earlier, the terms above are formal terms for arbitrary variations. In particular, the condition (IV.1.11) will force cancellations in equation (IV.1.13). Let us examine equation (IV.1.8) given the condition (IV.1.11). For the right hand side of equation (IV.1.8) to be a total divergence depends solely on the behavior of the penultimate term, since only it and the last term are not in general divergence terms and the last term is by definition not a divergence. There are two ways, then, to satisfy conditions (IV.1.10) and (IV.1.11).

First we can have

$$\delta_0 \phi EL = \partial_\mu f^\mu \tag{IV.1.16}$$

and

$$\delta_0 L_2 = 0, \tag{IV.1.17}$$

in which case we have

$$\partial_\mu f^\mu \mathrel{\overset{\circ}{=}} 0 \tag{IV.1.18}$$

if $\delta_0 \phi EL$ is not identically zero. In this case we call equation (IV.1.16) a "weak continuity equation" of the "first kind."[40] (If $\delta_0 \phi EL$ is identically zero, we obtain what is called a "strong continuity" equation. These do not properly lead to conservation laws.) These continuity equations are the most familiar and are those that follow from conventional Noether's theorem. In fact, condition (IV.1.17), which requires that the form variation of the Lagrangian be a total divergence, reduces the above treatment to the conventional approach.[41]

There is a second possibility for associating variations with a continuity equation. We can have

[40] We follow Rosen (1974a) in this nomenclature.

[41] We take the conventional approach to be the Bessel-Hagen extension (see Candotti et al. (1970)) of Noether's theorem, which is equivalent to this treatment.

$$\delta_0 \phi EL = \partial_\mu f_1^\mu + f_2 \tag{IV.1.19}$$

where f_2 is a non-divergence part and

$$f_2 \mathrel{\mathring{=}} 0 \tag{IV.1.20}$$

because of condition (IV.1.10). We also see that condition (IV.1.17) is not necessary but rather

$$f_2 + \delta_0 L_2 = 0 \tag{IV.1.21}$$

must be true. Again equation (IV.1.19) is a weak continuity equation—i.e.,

$$\partial_\mu f_1^\mu \mathrel{\mathring{=}} 0 \tag{IV.1.22}$$

—this time of the "second kind".[42]Equations (IV.1.19)–(IV.1.21) then define the most general Noether transformations for a given Lagrangian ($f_2 = 0$ reduces (IV.1.19)–(IV.1.21) to (IV.1.16) and (IV.1.17).)

This leaves coordinate transformations arbitrary, meaning they are not relevant to the existence or construction of a conserved current.[43] The same is true for form variations of the Lagrangian equal to a to-

[42] See footnote 40

[43] See Boyer (1967) for a discussion of this point. Rosen (1974a) also discusses this point, although he seems to add some mystery to it.

tal divergence. As for the non-divergence part of the Lagrangian form-variation, it obviously does not contribute to the conserved current, f_1^μ, and since it is fixed by condition (IV.1.21) it is not concerned with the existence of the conserved current either. So no form variation of the Lagrangian, in general, is concerned with Noether transformations. This leaves pure field variations (equation (IV.1.3)) as the only candidates for Noether transformations.

This result is satisfying since our general analysis of the last chapter lead us to expect that only the canonical coordinates should be involved in the construction of a physical conserved object. We also see the similarity between equations (IV.1.19) and (III.2.18) (the expression we derived for general continuous development systems.) This confirms that we have, in fact, discovered the most general variations leading to Noether transformations.

To summarize, a necessary and sufficient condition for a transformation to be a Noether transformation for a given system is that the consequent variation of the Action satisfy Hamilton's principle without recourse to the equations of motion (where only the fields need to be varied, as normally dictated by Hamilton's principle.) An equivalent necessary and sufficient condition involves the equations of motion and is given by equations (IV.1.19) and (IV.1.20).

Symmetry Transformations

The previous discussion of Noether transformations was unambiguous. Once we had decided we wanted to find the most general transformations of a given system leading to a continuity equation, the procedure was straight-forward. The varied approaches and results one finds concerned with symmetry-conservation investigations in field theory arise from differing definitions of symmetry transformations, many of which even involve the concept of the Noether transformation in their definition.[44]

In the last chapter we began from first principles to develope a definition of a symmetry transformation. We then, in stages, made this concept more specific and found a definition for general, continuous, temporal, development systems. We now again make this definition more specific, to apply to field theory; hence, our definitions will be unambiguous and traceable to first principles.

For symmetry transformations the question is not only what types of variations of the Action are candidates for the infinitesimal variations connected with symmetries, but what should be the condition on the Action for a variation to represent a symmetry. We found in the last chapter that a symmetry mapping, M, commutes with the temporal development mapping, N. In field theory N is given by the field solutions (physical fields), which we label by ϕ^0; i.e., we have the correspondence

$$N \Longleftrightarrow \phi^0. \tag{IV.1.23}$$

For M to act on N—as we allowed it to in the last chapter when we defined $N' = MN$ as an infinitesimal variation of N—it must vary these field solutions; i.e.,

$$MN \Longleftrightarrow \phi^0 \rightarrow \overline{\phi^0} = \phi^0 + \delta_0^m \phi|_{\phi=\phi^0} \tag{IV.1.24}$$

[44] See, for instance, Jackiw (1972), Section II.

or more generally a candidate symmetry variation must vary the fields

$$M \Longleftrightarrow \phi \to \bar{\phi} = \phi + \delta_0^m \phi. \tag{IV.1.25}$$

We see that symmetry transformations are properly associated only with variations of the fields, not with variations of the space-time coordinates nor with form variations of the Lagrangian.

At this point, it may seem that the proper way to define a symmetry transformation is by direct reference to, and as a condition on, the equations of motion—or more specifically their solutions—with no need to refer to the Action. In principle, for classical field theory this is the most straight-forward way to proceed. In quantum field theory, however, the equations of motion cannot be relied upon so heavily: so going so far as to formulate the condition for a symmetry transformation in terms of the equations of motion would obviously not be fruitful. Reference to the equations of motion in defining a symmetry transformation, as in the relation (IV.1.24), is allowed (as, in fact, such reference is made expressly so in deriving quantum electrodynamics through the process of second quantization.) The condition, however, should be one on the Action, which should therefore be achieved by letting the symmetry variation act on general ϕ, as through the relation (IV.1.25).

Let us write down this condition on the field solutions. First we note one last correspondence:

$$NM \Longleftrightarrow \bar{\phi} \to (\bar{\phi})^0 = (\phi + \delta_0^m \phi)^0. \tag{IV.1.26}$$

Our commutation condition for M then yields the condition: a symmetry transformation must map field solutions into field solutions in the transformed system; i.e,

$$NM = NM \Longleftrightarrow \overline{\phi^0} = (\bar{\phi})^0. \tag{IV.1.27}$$

Consider the variation of the equation of motion under some variation of the fields:

$$EL(\bar{\phi}, \partial_\mu \bar{\phi}) = EL(\phi, \partial_\mu \phi) + \delta_\phi(EL). \tag{IV.1.28}$$

Let this be a symmetry variation, so that

$$EL(\bar{\phi}, \partial_\mu \bar{\phi})|_{\bar{\phi}=(\bar{\phi})^0=\overline{\phi^0}} = EL(\phi, \partial_\mu \phi)|_{\phi=\phi^0} + [\delta_\phi^m(EL)]_{\phi=\phi^0}, \tag{IV.1.29}$$

yielding

$$[\delta_\phi^m(EL)]_{\phi=\phi^0} = 0 \tag{IV.1.30}$$

—the obvious condition that the equation of motion must be invariant under a symmetry variation at its solutions. It also becomes obvious that we cannot infer anything about the response of the structure of the equations of motion—and consequently nothing about the behavior or structure of the Lagrangian or Action—from their behavior at one (or several) values of the fields. It seems that the definition of a symmetry transformation needs to be further specified to yield any more restrictive conclusions concerning the Action.

Before we do this, however, let us take this narrow definition a little further by paralleling our analysis of the last chapter. Namely, since Hamilton's principle is a principle whose application yields the physical field solutions, it must hold for the transformed Action,

$$I = \int d^4xL(\bar{\phi}(x), \partial_\mu\bar{\phi}(x)), \tag{IV.1.31}$$

which we can write as

$$I = \int d^4x[L(\phi(x), \partial_\mu\phi(x)) + \delta_\phi^m L], \tag{IV.1.32}$$

where

$$\delta_\phi^m L = \partial_\mu \Pi^\mu \delta_0^m \phi + \delta_0^m \phi EL. \tag{IV.1.33}$$

Allowing an arbitrary variation of this Action we find

$$\delta I = \int d^4x[\delta_\phi L + \delta_\phi \delta_\phi^m L]. \tag{IV.1.34}$$

Again, as in the last chapter, we find we can satisfy Hamilton's principle for this Action either by having

$$\delta_\phi^m L = \partial_\mu k \tag{IV.1.35}$$

or by

$$\delta_\phi \delta_\phi^m L \doteq \partial_\mu k'. \tag{IV.1.36}$$

The first condition is equivalent to exactly the sort of wider definition of a symmetry transformation that we mentioned in the preceding paragraph; namely, it is equivalent to requiring that the symmetry transformation be such that the equations of motion be form invariant; or, in other words, equation (IV.1.30) should hold not only for the field solutions but for all ϕ. We can see this easily as follows. If $\delta_\phi^m L$ is a total divergence then EL is invariant, since the Euler derivative of a total divergence is identically zero, by a well-known theorem in variational calculus[45]—which also states that if EL is invariant under a variation of L, that variation must be a total divergence. We also see that this broader definition of a symmetry variation, by leading to the condition (IV.1.35), leads immediately to a continuity equation of the first kind.

However, there remains the possibility of condition (IV.1.36). Since condition (IV.1.35) and (IV.1.36) are mutually exclusive, it must be that condition (IV.1.36) is obtained for those transformations which do not leave the equations of motion form invariant, but only map field solutions to field solutions. In fact, condition (IV.1.36) can be shown to be the necessary and sufficient equivalent of equation (IV.1.30).

[45] See, for instance, Courant and Hilbert (1953), p. 193.

It now becomes evident, especially in light of our general formalism of the last chapter, that the above proposed "extension" of, or broader definition of a symmetry transformation (as a condition on all ϕ), is really a demand that this symmetry also be a physical symmetry: i.e., a symmetry on physical state space.

We can summarize the situation as follows. Condition (IV.1.30) is the condition on a transformation such that it be a symmetry of the laws of nature. Condition (IV.1.36) (which is equivalent to condition (IV.1.30) written instead as $\delta_\phi^m(EL) \mathrel{\mathring{=}} 0$) is a condition that the transformation be a symmetry of Nature but that it not be a symmetry on physical state space. Condition (IV.1.35) requires that the transformation be a symmetry of Nature and a symmetry on physical state space. Of course, this last requirement is exactly the requirement that a transformation lead to a conservation under our general symmetry-conservation formalism,[46] and condition (IV.1.35) is in fact identical with our condition for Noether transformations which lead to a continuity equation of the first kind.[47]

As for those field transformations given by equations (IV.1.19) and (IV.1.20), which lead to a continuity equation of the second kind, a simultaneous non-divergence form transformation of the Lagrangian is required (given by equation (IV.1.21)), hence these transformations are explicitly non-canonical and cannot be related to a symmetry of Nature. From our results of the last chapter, we see that these should be transformations, which when associated with a property of states, such property will be an "ill-defined" one.[48]

[46] Cf. Section III.2.

[47] Cf. equation (IV.1.16).

[48] Cf. Section III.2. Other investigators have missed these connections because they did not consider non-heterogeneous systems, for which there is a distinction between a symmetry of Nature and a physical symmetry.

We need to make a special note about space-time symmetries. Although we have made it clear that we are to properly only consider functional variations (i.e., δ_0) in the fields when considering symmetry transformations, when investigating physically meaningful space-time symmetries, conditions are placed on the fields, $\phi(x)$, taking into account their dependency on the space-time coordinates. In other words, the condition is given on $\delta\phi$, which is equal to

$$\delta\phi = \delta_0\phi + \delta x^\mu \partial_\mu \phi, \tag{IV.1.37}$$

so that we should consider

$$\delta_0\phi = \delta\phi - \delta x^\mu \partial_\mu \phi, \tag{IV.1.38}$$

as our symmetry variation. For instance, translation invariance requires of the fields that under a translation there should be no change in the fields so that

$$\delta\phi = 0, \tag{IV.1.39}$$

which means that (since $\delta x^\mu = \epsilon^\mu$ for a translation)

$$\delta_0\phi = -\epsilon^\mu \partial_\mu \phi \tag{IV.1.40}$$

is the proper variation. It is still only the variation $\delta_0\phi$, however, that is being considered and that is associated with the symmetry transformation and Noether transformation. Equation (IV.1.38) merely gives the correct expression for $\delta_0\phi$ given a physical requirement on $\delta\phi$; it does not indicate that we are separately considering variations of the space-time coordinates.

If we have found a continuity equation associated with a variation, then if we integrate it over all space and over a finite time interval we find

$$\int\limits_{t_1}^{t_2} dx^0 \int\limits_{-\infty}^{+\infty} d^3\mathbf{x}\partial_\mu j^\mu \mathrel{\overset{\circ}{=}} 0$$

or

$$\int\limits_{t_1}^{t_2} dx^0 \tfrac{\partial}{\partial x^0} \int\limits_{-\infty}^{+\infty} d^3\mathbf{x} j^0 + \int\limits_{t_1}^{t_2} dx^0 \int\limits_{-\infty}^{+\infty} d^3\mathbf{x}\partial_i j^i \mathrel{\overset{\circ}{=}} 0. \tag{IV.1.41}$$

The last term is a surface term which will vanish if the fields vanish at infinity. In this case

$$\int\limits_{-\infty}^{+\infty} d^3\mathbf{x} j^0(t_2,\mathbf{x}) - \int\limits_{-\infty}^{+\infty} d^3\mathbf{x} j^0(t_1,\mathbf{x}) \mathrel{\overset{\circ}{=}} 0. \tag{IV.1.42}$$

We can now identify

$$Q(t) \equiv \int_{-\infty}^{+\infty} d^3\mathbf{x} j(t, \mathbf{x}) \tag{IV.1.43}$$

as a conserved charge, since IV.1.42 is true for arbitrary time intervals; i.e., we can write

$$\frac{dQ}{dt} \overset{\circ}{=} 0. \tag{IV.1.44}$$

Finally, we note that we may wish to investigate certain approximate symmetries. The field variations associated with an approximate symmetry will cause the variation of the Action to contain a non-divergence term which does not go to zero upon imposition of the equations of motion. This can happen if and only if equation (IV.1.19) holds but equation (IV.1.20) does not, so that

$$\partial_\mu f_1^\mu \overset{\circ}{=} -f_2. \tag{IV.1.45}$$

Examples

The term $\delta_0 \phi EL$ can be written out in a convenient form. Since the total variation in L due to a variation $\delta_0\phi$ is

$$\begin{aligned} \delta_\phi L &= (\frac{\partial L}{\partial \phi})\delta_0\phi + \Pi_\mu \partial_\mu \delta_0 \phi \\ &= \partial_\mu(\Pi^\mu \delta_0 \phi) + \delta_0 \phi EL, \end{aligned} \tag{IV.1.46}$$

we can write

$$\delta_0\phi EL = \delta_\phi L - \partial_\mu(\Pi^\mu \delta_0\phi). \tag{IV.1.47}$$

We can also write $\delta_\phi L$ in general as

$$\begin{aligned}\delta_\phi L &= (\delta_{\phi_x} - \delta x^\mu \partial_\mu)L \\ &= \delta_{\phi_x} L - \partial_\mu(\delta x^\mu L) + L\partial_\mu \delta x^\mu,\end{aligned} \tag{IV.1.48}$$

where $\delta_{\phi_x} L$ is the variation in L due to variations of ϕ and x. For pure field transformations (internal symmetries) the last two terms in equation (IV.1.48) are zero and the condition for such a transformation to yield a continuity equation of the first kind is

$$\delta_\phi L = 0, \tag{IV.1.49}$$

and the conserved current is

$$j^\mu = \Pi^\mu \delta_0\phi. \tag{IV.1.50}$$

For space-time symmetries the condition for a continuity equation of the first kind is

$$\delta_{\phi_x} L = -L\partial_\mu \delta x^\mu, \tag{IV.1.51}$$

and the conserved current is

$$j^\mu = \Pi^\mu \delta_0 \phi + L \delta x^\mu. \tag{IV.1.52}$$

We note that the current can always be defined, even if equation (IV.1.51) or (IV.1.49) does not hold and j^μ is not conserved.

Transformations with which we associate symmetries form groups. In particular infinitesimal canonical transformations form Lie groups; if space-time transformations are being considered the fields must transform under a representation of the space-time transformation group, otherwise they will transform under a representation of some other Lie group. The fields, as canonical coordinates, in general obey the Lie bracket products

$$[\Pi(t, \mathbf{x}), \phi(t, \mathbf{y})] = \delta(\mathbf{x} - \mathbf{y}) \tag{IV.1.53}$$

$$[\Pi(t, \mathbf{x}), \Pi(t, \mathbf{y})] = [\phi(t, \mathbf{x}), \phi(t, \mathbf{y})] = 0, \tag{IV.1.54}$$

where Π is the appropriate canonical conjugate. In quantum field theory, where the fields are quantized operators, these brackets become (anti-)commutators;[49] in classical field theory they are Poisson brackets. The abstract Lie group generators obey the algebra

$$[T_a, T_b] = -c^c_{ab} T_c \tag{IV.1.55}$$

[49] In this case appropriate factors of i also need to be inserted into the commutators because of hermicity requirements on the operators.

where c^c_{ab} are the structure constants. Their representatives, defined through the field variation

$$\delta_0\phi = \epsilon^\mu D_\mu \phi(x), \tag{IV.1.56}$$

also obey this algebra.

The conserved charges are also representatives of the generators of the transformation:

$$\delta_0\phi = [Q, \phi]. \tag{IV.1.57}$$

We can prove this as follows. From equations (IV.1.43) and (IV.1.52) we have

$$\begin{aligned}
[Q(t,\mathbf{y}), \phi(t,\mathbf{x})] &= \int d^3\mathbf{y}[\Pi^0(t,\mathbf{y})\delta_0\phi(t,\mathbf{y}) + L(t,\mathbf{y})\delta t, \phi(t,\mathbf{x})] \\
&= \int d^3\mathbf{y}\{[\Pi^0(t,\mathbf{y}), \phi(t,\mathbf{x})]\delta_0\phi(t,\mathbf{y}) \\
&+ \Pi^0(t,\mathbf{y})[\delta_0\phi(t,\mathbf{y}), \phi(t,\mathbf{x})] + [L(t,\mathbf{y})\delta t, \phi(t,\mathbf{x})]\} \\
&= \delta_0\phi(t,\mathbf{x}) + \int d^3\mathbf{y}\{\Pi^0(t,\mathbf{y})[\delta_0\phi(t,\mathbf{y}), \phi(t,\mathbf{x})] \\
&+ [L(t,\mathbf{y})\delta t, \phi(t,\mathbf{x})]\},
\end{aligned}$$

where we have used equation (IV.1.53) to obtain the first term. In the last two terms, $\delta_0\phi$ and L, as functionals of ϕ and $\partial_\mu\phi$, only have non-vanishing Lie bracket products due to their dependence on $\partial_0\phi$. We can write generally

$$[F(\phi(t,\mathbf{y}),\partial_\mu\phi(t,\mathbf{y}),\phi(t,\mathbf{x})]=\frac{\partial F}{\partial[\partial_0\phi]}[\partial_0\phi(t,\mathbf{y}),\phi(t,\mathbf{x})].$$

Consequently we find

$$\begin{aligned}[Q(t,\mathbf{y}),\phi(t,\mathbf{x})]&=\delta_0\phi(t,\mathbf{x})+\int d^3\mathbf{y}\{\Pi^0(t,\mathbf{y})\frac{\partial\delta_0\phi}{\partial[\partial_0\phi]}[\partial_0\phi(t,\mathbf{y}),\phi(t,\mathbf{x})]\\&+\frac{\partial L}{\partial[\partial_0\phi]}\delta t[\partial_0\phi(t,\mathbf{y}),\phi(t,\mathbf{x})]\}\\&=\delta_0\phi+\int d^3\mathbf{y}\{\Pi^0(t,\mathbf{y})(-\delta t)\\&+\Pi^0(t,\mathbf{y})\delta t\}[\partial_0\phi(t,\mathbf{y}),\phi(t,\mathbf{x})],\end{aligned}$$

where in the last line we have used equation (IV.1.38) to evaluate the first derivative term, and have assumed that

$$\frac{\partial\delta\phi}{\partial[\partial_0\phi]}=0,$$

which is to say that $\delta\phi$ is not a function of $\partial_0\phi$, which is true for all symmetries which we will encounter. We also made the identification

$$\Pi^0\equiv\frac{\partial L}{\partial[\partial_0\phi]}$$

for the canonical conjugate. The term in brackets above vanishes, and we have completed our proof. We note that we have made no use of the time independence of Q, so this result holds for conserved as well as non-conserved charges.

We now apply the previous analysis to examples of well-known transformations, still keeping our Lagrangian unspecified. We have already mentioned space-time translations. We found that for these transformations $\delta x^\mu = \epsilon^\mu$, however, they can be represented more generally as

$$\delta x^\mu = \epsilon^\rho P_\rho x^\mu, \tag{IV.1.58}$$

where

$$P_\rho = -\partial_\rho \tag{IV.1.59}$$

is the most general representation of the four translation generators. In this case we can write (in agreement with equation (IV.1.39) and (IV.1.40))

$$\delta_0 \phi = -\epsilon^\mu \partial_\mu \phi$$

so that

$$\begin{aligned} \delta_0 \phi EL &= -\epsilon^\mu \partial_\mu \phi \left(\frac{\partial L}{\partial \phi} - \partial_\nu \Pi^\nu\right) \\ &= -\epsilon^\mu (\partial_\mu L - \partial_\mu \phi \partial_\nu \Pi^\nu) \\ &= \epsilon_\nu \partial_\mu \theta_c^{\mu\nu}, \end{aligned} \tag{IV.1.60}$$

where

$$\theta_c^{\mu\nu} = \Pi^\mu \partial^\nu \phi - g^{\mu\nu} L \tag{IV.1.61}$$

is the canonical energy momentum tensor. Apparently all Lagrangians possess translation symmetry and have associated a conserved quantity, $\theta_{\mu\nu}$. This is because we have chosen L not to depend explicitly on x_μ and this guarantees invariance under translations, as can be seen from equation (IV.1.51). Equation (IV.1.61) could have also been obtained directly from equation (IV.1.52).

The generators obviously obey

$$[P_\mu, P_\nu] = 0 \tag{IV.1.62}$$

so that the structure constants are zero. The conserved charge is

$$\begin{aligned} P_\mu &= \int d^3\mathbf{x}\theta_{0\mu} \\ &= \int d^3\mathbf{x}(P_0\partial_\mu\phi - g_{0\mu}L), \end{aligned} \tag{IV.1.63}$$

which obeys

$$\epsilon^\mu [P_\mu, \phi(x)] = -\epsilon^\mu \partial_\mu \phi. \tag{IV.1.64}$$

Next we consider Lorentz transformations. The space-time coordinates transform as

$$\delta x^\mu = \epsilon^{\mu\nu} x_\nu, \tag{IV.1.65}$$

where $\epsilon^{\mu\nu}$ is an infinitesimal antisymmetric tensor parameter which consequently represents six independent parameters. This can be written as

$$\delta x^\mu = -(1/2)\epsilon^{\rho\sigma} L_{\rho\sigma} x^\mu, \tag{IV.1.66}$$

where the six generators are defined as

$$L_{\mu\nu} \equiv (x_\mu \partial_\nu - x_\nu \partial_\mu). \tag{IV.1.67}$$

These obey the algebra

$$[L_{\mu\nu}, L_{\rho\sigma}] = g_{\nu\rho} L_{\mu\sigma} - g_{\mu\rho} L_{\nu\sigma} - g_{\nu\sigma} L_{\mu\rho} + g_{\mu\sigma} L_{\nu\rho}, \tag{IV.1.68}$$

which is the Lie algebra of $SO(3,1)$. $L_{\mu\nu}$, however, are not the most general representation of the generators of this group. Operators which obey the same algebra as the $L_{\mu\nu}$ but instead act directly on the spacetime indices can be added to the $L_{\mu\nu}$.[50] This most general representation is given by

[50] See, for example, Ramond (1981), Section I.2.

$$M_{\mu\nu} \equiv (x_\mu \partial_\nu - x_\nu \partial_\mu) + \Sigma_{\mu\nu}. \tag{IV.1.69}$$

The antisymmetric tensor matrix $\Sigma_{\mu\nu}$ is called the spin matrix. There are distinct possibilities for the $\Sigma_{\mu\nu}$ and it is these which determine the distinct representations of the Lorentz group. As we noted in the beginning of this chapter, the fields must transform under definite representations of the Lorentz group and are therefore intimately connected to $\Sigma_{\mu\nu}$.

For instance, scalar fields are defined to transform as Lorentz scalars so that

$$\phi'(x') = \phi(x) \tag{IV.1.70}$$

under a Lorenz transformation. This means

$$\delta\phi = 0 \tag{IV.1.71}$$

or

$$\delta_0\phi + \epsilon^{\mu\nu} x_\nu \partial_\mu \phi = \delta_0\phi + 1/2\epsilon^{\rho\sigma} L_{\rho\sigma}\phi = 0$$

which yields

$$\delta_0\phi = -(1/2)\epsilon^{\rho\sigma} L_{\rho\sigma}\phi. \quad (scalar\ fields) \tag{IV.1.72}$$

Comparing with the most general representation

$$\delta_0 = (1/2)\epsilon^{\rho\sigma} M_{\rho\sigma}\phi \tag{IV.1.73}$$

tells us that $\Sigma_{\mu\nu} = 0$ for scalar fields. Similarly it can be shown that

$$\Sigma^{\mu\nu}_{(ij)} = (1/2)\sigma^{\mu\nu}_{ij} \tag{IV.1.74}$$

for Dirac spinors[51] (ij are the spinor indices), and

$$\Sigma^{\mu\nu}_{(\alpha\beta)} = g^{\mu}_{\alpha} g^{\nu}_{\beta} - g^{\mu}_{\beta} g^{\nu}_{\alpha} \tag{IV.1.75}$$

for vector fields ($\alpha\beta$ are space-time indices).

We now proceed to write down the current associated with the transformation (IV.1.73), which we can write as

$$\delta_0 = (1/2)\epsilon_{\mu\nu}[\Sigma^{\mu\nu} + (x^{\mu}\partial^{\nu} - x^{\nu}\partial^{\mu})]\phi. \tag{IV.1.76}$$

We note this indicates that

$$\delta\phi = (1/2)\epsilon_{\mu\nu}\Sigma^{\mu\nu}\phi. \tag{IV.1.77}$$

[51] $\sigma_{\mu\nu} = (i/2)[\gamma_{\mu}, \gamma_{\nu}]$.

Using equation (IV.1.52), we find (extracting the infinitesimal parameter, as is customary) the canonical angular momentum current

$$M_c^{\mu\alpha\beta} = \Pi^\mu \Sigma^{\alpha\beta}\phi + \Pi^\mu (x^\alpha \partial^\beta - x^\beta \partial^\alpha)\phi + (x^\alpha g^{\beta\mu} - x^\beta g^{\alpha\mu})L,$$

which can be written more simply m terms of the canonical energy momentum tensor as

$$M_c^{\mu\alpha\beta} = \Pi^\mu \Sigma^{\alpha\beta}\phi + x^\alpha \theta_c^{\mu\beta} - x^\beta \theta_c^{\mu\alpha}. \tag{IV.1.78}$$

This last result represents a fact common to all space-time symmetries; that is, we can write all currents associated with space-time symmetries in terms of the canonical energy momentum tensor. We can see this as follows. We can generally write equation (IV.1.52) as

$$j^\mu = \Pi^\mu \delta\phi - \Pi^\mu \delta x_\nu \partial^\nu \phi + L\delta x^\mu,$$

where we have used equation (IV.1.38),and so

$$\begin{aligned} j^\mu &= \Pi^\mu \delta\phi - \delta x_\nu (\Pi^\mu \partial^\nu \phi - g^{\mu\nu} L) \\ &= \Pi^\mu \delta\phi - \delta x_\nu \theta_c^{\mu\nu}. \end{aligned} \tag{IV.1.79}$$

The current (IV.1.78) is conserved when condition (IV.1.51) holds:

$$\epsilon_{\mu\nu}\Sigma^{\mu\nu}L = -L\partial_\alpha(\epsilon^{\alpha\beta}x_\beta) = 0. \tag{IV.1.80}$$

which means that L must be invariant under the variations δ_{ϕ_x} (or under the action of $\Sigma^{\mu\nu}$;) i.e., it must be a Lorentz scalar.

The charges associated with the current (IV.1.78) are

$$M^{\alpha\beta} = \int d^3\mathbf{x} M_c^{0\alpha\beta} = \int d^3\mathbf{x}(\Pi^0\Sigma^{\alpha\beta}\phi + x^\alpha\theta_c^{0\beta} - x^\beta\theta_c^{0\alpha}). \tag{IV.1.81}$$

Of course, they obey

$$(1/2)\epsilon_{\alpha\beta}[M^{\alpha\beta},\phi] = (1/2)\epsilon_{\alpha\beta}[\Sigma^{\alpha\beta} + x^\alpha\partial^\beta - x^\beta\partial^\alpha]\phi \tag{IV.1.82}$$

and obey the algebra given by equation (IV.1.68). We can also construct the Lie bracket of these generators with the translation generators

$$[M_{\mu\nu}, P_\rho] = -g_{\mu\rho}P_\nu + g_{\nu\rho}P_\mu. \tag{IV.1.83}$$

Equations (IV.1.62), (IV.1.68), and (IV.1.83) define the Lie algebra of the ten parameter Poincaré group.

We consider another group of space-time transformations—the dilations.[52] These transformations are defined by

$$\delta x^\mu = \epsilon x^\mu. \tag{IV.1.84}$$

This can be written as

$$\delta x^\mu = \epsilon x^\nu \partial_\nu x^\mu. \tag{IV.1.85}$$

The most general representation of the dilation generator, however, can be written as

$$D = (d - x^\mu \partial_\mu), \tag{IV.1.86}$$

where d is a constant matrix called the scale dimension matrix, so that

$$\delta_0 \phi = \epsilon (d - x^\mu \partial_\mu)\phi \tag{IV.1.87}$$

and

$$\delta \phi = \epsilon d \phi. \tag{IV.1.88}$$

[52] The term dilatation is more often used. We feel dilation is the more natural noun.

Equation (IV.1.88) can be seen to be a simple generalization of equation (IV.1.84). Equations (IV.1.84) and (IV.1.88) are similar to a "rescaling" of variables (and in fact dilations are often referred to as scale transformations.) Since a dilation is a space-time transformation, however, it differs essentially from a simple rescaling. Since only functions of x$^\mu$ are affected by a dilation, terms in the Lagrangian containing dimensionful constant parameters will transform differently.

A consistent definition of the matrix d, therefore, is that it multiplies the fields by their physical dimensions (for $\hbar = c = 1$, in units of inverse length.) These dimensions are assigned to the fields so that their dynamical term in the Lagrangian is such that the Action is dimensionless. The form of this term, as we will see, is fixed by the demands of Poincaré symmetry and the other restrictions on the Action that we stated earlier.

The Lie algebra is, of course, trivial:

$$[D, D] = 0 \tag{IV.1.89}$$

Since Poincaré symmetry will be assumed in constructing field theories, the eleven parameter Lie group consisting of the Poincaré and dilation transformations is of more significance. It is defined by the algebra given by equations (IV.1.62), (IV.1.68), (IV.1.83), (IV.1.89) and by

$$[D, P_\mu] = P_\mu \tag{IV.1.90}$$

$$[D, M_{\mu\nu}] = 0 \tag{IV.1.91}$$

and is called the Weyl group.

Using the expression (IV.1.79) to construct the canonical dilation current we find

$$D_c^{\mu} = \Pi^{\mu} d\phi - x^{\mu}\theta_c^{\mu\nu}. \tag{IV.1.92}$$

The condition (IV.1.51) for this current to be conserved becomes

$$\epsilon dL = -L\partial_{\mu}(\epsilon x^{\mu}),$$

or

$$dL = -4L. \tag{IV.1.93}$$

This states that the Lagrangian must have a scale dimension of 4, which is its physical dimension. For a Lagrangian to be dilation symmetric, then, it must contain only terms which contain no dimensionful constants (i.e., only terms whose scale dimensions is equal to their physical dimension.)

The dilation charge is

$$D = \int d^3\mathbf{x} D_c^0 = \int d^3\mathbf{x}(\Pi^0 d\phi - x^0\theta_c^{0\nu}). \tag{IV.1.94}$$

It obeys

$$\epsilon[D, \phi] = \epsilon(d - x^\mu \partial_\mu)\phi. \tag{IV.1.95}$$

The last space-time transformations we consider are the conformal (angle preserving) transformations:

$$\delta x^\mu = \epsilon_\nu(2x^\mu x^\nu - g^{\mu\nu} x^2). \tag{IV.1.96}$$

Once again we can rewrite these as

$$\delta x^\mu = \epsilon_\beta(2x^\alpha x^\beta - g^{\alpha\beta} x^2)\partial_\alpha x^\mu. \tag{IV.1.97}$$

To understand how the fields transform, we rewrite equation (IV.1.96) in a more suggestive form:

$$\begin{aligned}\delta x^\mu &= \epsilon_\alpha(x^\mu x^\alpha + x^\mu x^\alpha - g^{\mu\alpha} x_\beta x^\beta) \\ &= \epsilon_\alpha x_\beta[g^{\mu\beta}(x^\alpha) - (x^\beta g^{\mu\alpha} - x^\alpha g^{\mu\beta})].\end{aligned} \tag{IV.1.98}$$

The terms in parenthesis are just the infinitesimal dilation and Lorentz transformations, respectively, sans their infinitesimal parameters. Substituting for their field representations we find

$$\delta\phi = 2\epsilon_\alpha x_\beta[g^{\beta\alpha} d - \Sigma^{\beta\alpha}]\phi \tag{IV.1.99}$$

so the we have

$$\delta\phi = \epsilon_\alpha [2x_\beta(g^{\beta\alpha} d - \Sigma^{\beta\alpha}) - (2x^\beta x^\alpha - g^{\beta\alpha} x^2)\partial_\beta]\phi, \tag{IV.1.100}$$

and the most general representation of the conformal generator

$$K^\alpha = 2x_\beta(g^{\beta\alpha} d - \Sigma^{\beta\alpha}) - (2x^\beta x^\alpha - g^{\beta\alpha} x^2)\partial_\beta. \tag{IV.1.101}$$

It obeys the following Lie brackets:

$$[K^\alpha, K^\beta] = 0 \tag{IV.1.102}$$

$$[D, K^\alpha] = -K^\alpha \tag{IV.1.103}$$

$$[M^{\alpha\beta}, K^\gamma] = g^{\alpha\gamma} K^\beta - g^{\beta\gamma} K^\alpha \tag{IV.1.104}$$

$$[P^\alpha, K^\beta] = -g^{\alpha\beta} D + 2M^{\alpha\beta}. \tag{IV.1.105}$$

These equations together with equations (IV.1.62), (IV.1.68), (IV.1.83), (IV.1.89), (IV.1.90), and (IV.1.91) define the algebra of the fifteen parameter Lie group called the conformal group. As we can see from equation (IV.1.105), there is no closed Lie group containing only the Poincaré and the conformal transformations. Interestingly, the conformal transformations can be obtained by applying the space-time inversion to the translations, whereas no new transformations are obtained by applying the inversion to the Lorentz or dilation transformations. So the conformal group is the smallest group of space-time transformations containing the Poincare group and the inversion.

To construct the canonical conformal current, we use equation (IV.1.79). We find

$$K_c^{\alpha\mu} = 2x_\beta \pi^\mu (g^{\beta\alpha} d - \Sigma^{\beta\alpha})\phi - (2x^\alpha x_\beta x^2)\theta_c^{\mu\beta}. \tag{IV.1.106}$$

The condition for this current to be conserved is, from equation (IV.1.51),

$$2x_\beta [g^{\beta\alpha} d - \Sigma^{\beta\alpha})L = -L\partial_\mu(2x^\mu x^\alpha - g^{\mu\alpha}x^2),$$

or

$$2x^\alpha dL - 2x_\beta \Sigma^{\beta\alpha} L = -8Lx^\alpha. \tag{IV.1.107}$$

This equation is satisfied if the Lagrangian is both dilation and Lorentz symmetric, as represented by equation (IV.1.93) and (IV.1.80), respectively. ~~So all Poincaré and dilation symmetric field theories are conformally symmetric.~~[53]

Correction: So all conformally symmetric field theories are Poincaré and dilation symmetric.

The charge associated with the conformal current is

$$K^\alpha = \int d^3x K_c^{\alpha 0} = \int d^3x [2x_\beta \Pi^0 (g^{\beta\alpha} d - \Sigma^{\beta\alpha})\phi - (2x^\alpha x_\beta - g_\beta^\alpha x^2)\theta_c^{0\beta}]. \tag{IV.1.108}$$

[53] An additional condition is often found for field theories to be conformal symmetric. See, for instance, Jackiw (1972), pp. 209–10. This condition, which also adds a superfluous term to the conformal current, apparently arises from a confusion of the role of coordinate variations in Noether's theorem.

We also have

$$\epsilon_\alpha[K^\alpha, \phi] = \epsilon_\alpha[2x_\beta(g^{\beta\alpha}d - \Sigma^{\beta\alpha}) - (2x^\beta x^\alpha - g^{\beta\alpha}x^2)\partial_\beta]\phi. \quad \text{(IV.1.109)}$$

Finally we consider internal field symmetries. We can write these transformations generally as

$$\delta_0 = -i\epsilon_a L_a \phi, \quad \text{(IV.1.110)}$$

where once again the ϵ_a are all real infinitesimal parameters and the L_a are our matrix representations of the Lie group generators. These obey some Lie algebra

$$[L_a, L_b] = iC_{abc}L_c. \quad \text{(IV.1.111)}$$

We already stated the general result for the current, equation (IV.1.50), which we can write now as

$$J^\mu_a = i\Pi^\mu L_a \phi. \quad \text{(IV.1.112)}$$

These currents are conserved when equation (IV.1.49) holds:

$$L_a L = 0, \quad \text{(IV.1.113)}$$

which states that the Lagrangian must be invariant under the action of the generators.

Action Constructs and their Symmetries

We are now ready to write down the most general Lagrange densities subject to Poincaré invariance and the other restrictions we stated at the beginning of this section. For a single scalar field, we can write

$$L = (1/2)\partial_\mu \phi(x) \partial^\mu \phi(x) - V(\phi(x)), \tag{IV.1.114}$$

where the 1/2 is put in by convention and V is a scalar function of $\phi(x)$. The physical interpretation of this and other Lagrangians we shall write down comes when they are used to find the canonical equations of motion or in the path integral treatment.

The first term, the "kinetic term," in equation (IV.1.114), is invariant under the full conformal group. Terms appearing in V, the "potential term", containing dimensionful parameters will, however, ruin dilation and conformal invariance. The kinetic term also tells us that the scalar field has a dimension (physical and scale) of L^{-1}.

The kinetic term is also invariant under a field shift

$$\phi \rightarrow \phi + a, \tag{IV.1.115}$$

—whereas the potential term is not—and under the transformation

$$\phi \to -\phi, \tag{IV.1.116}$$

under which the potential term may or may not be invariant. In particular if $V(\phi) = V(\phi^2)$, the Lagrangian will posses the symmetry (IV.1.116).

If a system consists of more than one scalar field, then other symmetries may exist. For instance, the kinetic term for N real scalar fields,

$$(1/2)\sum_{a=1}^{N} \partial_\mu \phi_a \partial^\mu \phi_a \tag{IV.1.117}$$

is invariant under a rotation of the fields into one another,

$$\delta\phi_a = \epsilon_{ab}\phi_b, \tag{IV.1.118}$$

where ϵ_{ab} is, naturally, an antisymmetric infinitesimal parameter. If $V(\phi) = V(\phi_a\phi_a)$, this theory possesses the internal symmetry (IV.1.118). In this case there are $(N^2 - N)/2$ conserved currents

$$J^\mu_{ab} = \phi_a \partial^\mu \phi_b - \phi_b \partial^\mu \phi_a. \tag{IV.1.119}$$

In particular, if we have the Lagrangian

$$L = (1/2)\partial_\mu\phi_1\partial^\mu\phi_1 + (1/2)\partial_\mu\phi_2\partial^\mu\phi_2 - V(\phi_1\phi_1 + \phi_2\phi_2) \tag{IV.1.120}$$

and we define

$$\phi \equiv (\phi_1 + i\phi_2)/\sqrt{2}, \tag{IV.1.121}$$

then we have

$$L = \partial_\mu \phi^* \partial^\mu \phi - V(\phi^* \phi). \tag{IV.1.122}$$

This Lagrangian is invariant under the infinitesimal "phase" transformation

$$\delta\phi = -i\epsilon q \phi, \tag{IV.1.123}$$

—where q is the generator of this transformation (in this case just a number)—and has associated the conserved current

$$J^\mu = iq(\phi^* \partial^\mu \phi - \phi \partial^\mu \phi^*) \equiv iq\phi^* \overleftrightarrow{\partial^\mu} \phi. \tag{IV.1.124}$$

For spin-1/2 fields the kinetic term is

$$(1/2)\bar{\psi} i\gamma^\mu \overleftrightarrow{\partial_\mu} \psi = (1/2)\psi_L^\dagger \sigma^\mu \overleftrightarrow{\partial_\mu} \psi_L + (1/2)\psi_R^\dagger \bar{\sigma}^\mu \overleftrightarrow{\partial_\mu} \psi_R, \tag{IV.1.125}$$

where ψ and $\overline{\psi}$ are four component complex objects called Dirac spinors and $\overline{\psi} \equiv \psi^\dagger \gamma_0$; they are constructed from ψ_L and ψ_R, two component objects called left and right Weyl spinors respectively which are the actual Lorentz group representations, such that

$$\psi \equiv \begin{pmatrix} \psi_L \\ \psi_R \end{pmatrix} \tag{IV.1.126}$$

so that parity is well defined for ψ:

$$P : \psi \to \psi^P = \begin{pmatrix} \psi_R \\ \psi_L \end{pmatrix} \equiv \gamma_0 \psi. \tag{IV.1.127}$$

We see from the terms (IV.1.125) that spinor fields have dimension -3/2. These spinor fields are, classically, Grassman (anti-commuting) functions.[54]

The kinetic term, under a constant field shift, gains a total divergence so that the Action is invariant under this transformation if there is no potential term. It is also invariant under conformal transformations. The two terms for ψ_L and ψ_R also separately posses these two invariances.

The Dirac kinetic term is invariant under two different "phase" transformations:

$$\delta\psi = -i\epsilon q \psi \tag{IV.1.128}$$

and

[54] The terms (IV.1.125) can also be written without the factor of 1/2 and with the derivative operator only acting to the right. The difference between these two forms is just a total divergence.

$$\delta\psi = -i\epsilon p\gamma_5\psi. \qquad \text{(IV.1.129)}$$

The second transformation is called a chiral transformation. (These two transformations reshuffle into two ordinary phase transformations on the two Weyl terms.) The two currents associated with the above transformations are

$$J^\mu = q\bar{\psi}\gamma^\mu\psi = iq\psi_L^\dagger\sigma^\mu\psi_L + iq\psi_R^\dagger\bar{\sigma}^\mu\psi_R^\mu \qquad \text{(IV.1.130)}$$

and

$$J_5^\mu = p\bar{\psi}\gamma^\mu\gamma_5\psi = ip\psi_L^\dagger\sigma^\mu\psi_L + ip\psi_R^\dagger\bar{\sigma}^\mu\psi_R^\mu. \qquad \text{(IV.1.131)}$$

We do not discuss spin-1 fields here as they will naturally evolve from our discussions in the next section.

IV.2 Gauge Invariance

The Abelian Case

In this section we investigate in more detail those internal symmetries which we found to be possessed naturally by the kinetic terms of scalar and spinor fields, whose form was dictated by the physical assumptions we made at the beginning of the preceding section. These phase symmetries fall into a larger class of "global gauge symmetries" which apply to multiplets of an arbitrary number of fields. Global gauge symmetries form a subset of the set of internal symmetries found in Nature. What distinguishes these symmetries is that when one allows the infinitesimal parameters which characterize them to depend on x, they are still found to represent symmetries in Nature. These symmetries then become "local gauge symmetries" and are said thereby to be "gauged." They are local because, by allowing the characterizing parameter to be a function of x, different variations of the fields may take place at different space-time locations under the imposition of a given symmetry transformation. From a constructive viewpoint the motivation for investigating such symmetries is clear. Imposition of (global) space-time symmetries allowed us to construct a restricted number of dynamical functions of the space-time coordinates, namely the kinetic terms for the fields. A symmetry which involves transformations which are prescribed by their location in space-time will add an additional order to our dynamics; namely, the imposition of forces or interactions. Making the symmetry parameter x-dependent also adds another dynamical function to the theory.

General relativity was the first theory to realize a symmetry made local. Einstein made Lorentz symmetry a local symmetry. For general relativity the added dynamical function is the gravitational field, which defines the connection between symmetry transformations made at different space-time points.

Soon after the success of general relativity, Herman Weyl attempted to apply this same technique to the phenomenon of electromagnetism.[55] He chose another space-time symmetry, dilation symmetry, to make into a local symmetry (hence, the Weyl group.) The electromagnetic field would then define the connection between dilations made at different space-time points. Weyl introduced the term gauge theory because of

[55] See, for instance, Quigg (1983), pp. 37–8, or Moriyasu (1982), for a discussion.

the gauge blocks which were used as reference lengths. Let us consider how this might work.

Let ϕ be a scalar or spinor field. Under an ordinary infinitesimal dilation its transformation is given by equation (IV.1.87). $\partial_\mu \phi$ transforms like

$$\delta_\phi[\partial_\mu \phi] = \partial_\mu \delta_0 \phi = \epsilon(d_\phi - 1 - x^\nu \partial_\nu)\partial_\mu \phi = \epsilon D \partial_\mu \phi, \tag{IV.2.1}$$

where d_ϕ is the dimension of ϕ and D is given by equation (IV.1.86). This result can be seen either by calculating $\partial_\mu \delta_0 \phi$ directly, or by noting that δ_0 and ∂_μ commute and $\partial_\mu \phi$ transforms like ϕ, except it has a dimension of one inverse length extra.

If our infinitesimal parameter is now allowed to depend on x, the transformation of ϕ has the same form; i.e.,

$$\delta_0 \phi = \epsilon(x)(d - x^\nu \partial_\nu)\phi, \tag{IV.2.2}$$

but this is not so for $\partial_\mu \phi$, since δ_0 no longer commutes with ∂_μ. Instead we find

$$\begin{aligned}\delta_\phi[\partial_\mu \phi] &= \partial_\mu \delta_0 \phi \\ &= \delta_0 \partial_\mu \phi + (\partial_\mu \epsilon(x))(d - x^\nu \partial_\nu)\phi \\ &= \epsilon(x) D[\partial_\mu + \partial_\mu \ln \epsilon(x)]\phi \\ &\simeq \epsilon(x) D[\partial_\mu + \partial_\mu \epsilon(x)]\phi, \end{aligned} \tag{IV.2.3}$$

where in the last line we made a first order approximation. The extra term we get because of the x-dependence of ϵ we have been able to put in a suggestive form in equation (IV.2.3). It suggests that, if we redefine $\partial_\mu \phi$ to

$$D_\mu \phi \equiv (\partial_\mu + eA_\mu)\phi, \tag{IV.2.4}$$

(where e is a constant of proportionality between A_μ and ϕ) and require A_μ to transform simultaneously under the infinitesimal transformation in such a way that it cancels the extra term in equation (IV.2.3),

$$\delta A_\mu = (-1/e)\partial_\mu \epsilon(x), \tag{IV.2.5}$$

then this new derivative, the "covariant derivative," will transform as $\partial_\mu \phi$ in equation (IV.2.1),

$$\delta_\phi[D_\mu \phi] = \epsilon(x) D D_\mu \phi. \tag{IV.2.6}$$

So, if $L(\phi, \partial_\mu \phi)$ is globally dilation invariant, $L(\phi, D_\mu \phi)$ will be locally dilation invariant. Weyl interpreted this needed function A_μ as the electromagnetic potential, which in fact possesses the freedom given by equation (IV.2.5).

Unfortunately, dilation invariance is not realized in Nature. At the quantum level, a particle of definite mass has associated a de-Broglie wavelength which sets a preferred scale. In field theory, since finite mass terms contain dimensionful parameters, dilation invariance requires either that all particles be massless or that the mass spectrum be continuous. This is, of course, not observed. We will have more to say on this subject later.

Phase transformations turn out to follow an analysis identical to the one we described above for dilations. This "coincidence" is due to the fact that both transformations are parametrized by a scalar parameter. It is the scalarity of the characterizing parameter that allows the identification appearing in equation (IV.2.5), and it is that transformation law which is the key to introducing the vector field A_μ. (Recall, none of the other fundamental space-time symmetries we discussed depend upon a scalar parameter.)

Let us reproduce the above results for a phase transformation, except that we consider, as is often done, a finite phase transformation.[56] We can write

$$U = e^{-iq\omega} \tag{IV.2.7}$$

for our finite global phase transformation, where we use ω to represent our parameter, a real number. The transformation (IV.2.7) is, of course, an element of a Lie group. It is called $U(1)$ because its algebra contains one generator and its representations are unitary. So

$$\phi(x) \rightarrow U\phi(x) \tag{IV.2.8a}$$

$$\partial_\mu\phi(x) \rightarrow U\partial_\mu\phi(x). \tag{IV.2.8b}$$

[56] See, for instance, Huang (1982), Ch. 3, or Quigg (1983), Ch. 3.

A local phase transformation

$$U(x) = e^{-iq\omega(x)} \tag{IV.2.9}$$

gives

$$\phi(x) \rightarrow U(x)\phi(x) \tag{IV.2.10}$$

$$\partial_\mu \phi(x) \rightarrow U(x)[\partial_\mu - iq\partial_\mu \omega(x)]\phi(x). \tag{IV.2.11}$$

Once again, if we let

$$D_\mu \equiv \partial_\mu + iqA_\mu \tag{IV.2.12}$$

and

$$A_\mu \rightarrow A_\mu + \partial_\mu \omega(x) \tag{IV.2.13}$$

we find

$$D_\mu \phi(x) \rightarrow U(x) D_\mu \phi(x). \tag{IV.2.14}$$

$L(\phi, D_\mu \phi)$ will be locally gauge invariant if $L(\phi, \partial_\mu \phi)$ is globally gauge invariant. Global gauge invariance is a symmetry found in Nature, so these are meaningful results.

In order for $L(\phi, D_\mu\phi)$ to be a closed dynamical system, it must contain a dynamical term for A_μ; that is, a term containing $\partial_\mu A_\nu$ quadratically. This term, of course, must also be a Lorentz scalar and must not spoil the new found symmetry which motivated its introduction: local gauge invariance.

The covariant derivative was constructed to be gauge covariant. Consider the following commutator:

$$\begin{aligned}(1/iq)[D_\mu, D_\nu] &= (\partial_\mu A_\nu - \partial_\nu A_\mu) + iq[A_\mu, A_\nu] \\ &= \partial_\mu A_\nu - \partial_\nu A_\mu \\ &\equiv F_{\mu\nu}, \end{aligned} \tag{IV.2.15}$$

since A_ν commutes with itself. $F_{\mu\nu}$ is called the field strength tensor, and by construction is gauge-invariant. We can then form the Lorentz scalar

$$L_\gamma \equiv -(1/4)F_{\mu\nu}F^{\mu\nu}, \tag{IV.2.16}$$

where the -1/4 is by convention. There is, in fact, no other term not proportional to this one that meets our requirements.

Consider now a system containing one complex scalar field and a spinor field, such that

$$L = (D_\mu\phi)^* D^\mu\phi + \bar\psi i\gamma^\mu D_\mu\psi - m^2\phi^*\phi - M\bar\psi\psi - (1/4)F_{\mu\nu}F^{\mu\nu}. \tag{IV.2.17}$$

This system is locally gauge invariant under the following simultaneous transformations:

$$\begin{aligned}
\phi &\to e^{-iq\omega(x)}\phi \\
\psi &\to e^{-iq\omega(x)}\psi \\
A_\mu &\to A_\mu + \partial_\mu\omega(x) \qquad \text{(IV.2.18)}
\end{aligned}$$

and physically represents a scalar and a spinor field interacting with an electromagnetic field. We can exhibit the interaction more explicitly by writing out the kinetic terms:

$$\begin{aligned}
(D_\mu\phi)^* D^\mu\phi &= (\partial_\mu - iqA_\mu)\phi^*(\partial^\mu + iqA^\mu)\phi \\
&= \partial_\mu\phi^*\partial^\mu\phi + q^2 A_\mu A^\mu\phi^*\phi - iq\phi^*\overleftrightarrow{\partial_\mu}\phi A^\mu \\
&= \partial_\mu\phi^*\partial^\mu\phi - iq(\phi^*\overleftrightarrow{D_\mu}\phi)A^\mu - q^2 A_\mu A^\mu\phi^*\phi \\
&= \partial_\mu\phi^*\partial^\mu\phi - J_\mu A^\mu - q^2A^2\phi^2 \qquad \text{(IV.2.19)} \\
\bar\psi\gamma^\mu D_\mu\psi &= \bar\psi i\gamma^\mu(\partial_\mu + iqA_\mu)\psi \\
&= \bar\psi i\gamma^\mu\partial_\mu\psi - qA_\mu\bar\psi\gamma^\mu\psi \\
&= \bar\psi i\gamma^\mu\partial_\mu\psi - J^\mu A_\mu. \qquad \text{(IV.2.20)}
\end{aligned}$$

In equations (IV.2.19) and (IV.2.20) the J_μ are the conserved currents associated with local gauge invariance:

$$J_\mu^{SCALAR} = iq(\phi^*\overleftrightarrow{D_\mu}\phi) = iq\phi^*\overleftrightarrow{\partial_\mu}\phi - 2q^2\phi^*\phi A_\mu \qquad \text{(IV.2.21)}$$

$$J_\mu^{SPINOR} = q\bar\psi\gamma_\mu\psi. \qquad \text{(IV.2.22)}$$

The scalar current gains an extra term because its canonical momentum changes,

$$\Pi^\mu \to \Pi^\mu - iqA^\mu\phi^*, \tag{IV.2.23}$$

whereas the spinor Π^μ does not change.

Applying Hamilton's principle to the Lagrangian (IV.2.17) alternately for each of the canonical fields, we will obtain an equation of motion for a massive scalar field with mass m (Klein-Gordon equation), an equation of motion for a massive spinor field with mass M (Dirac's equation), and Maxwell's equations with the currents J_μ as the sources. Adding a term of the form $m_\gamma^2 A^2$ to the Lagrangian (IV.2.17) which would yield a massive electrodynamics would spoil gauge invariance; hence, the masslessness of the photon is required by gauge invariance.

The Non-Abelian Case

We now consider the general gauge theory for multiplets of fields.[57] These results will be mostly a generalization of the preceding discussion except that complications arise from the fact that the elements of non-trivial Lie algebras do not commute, and since the gauge fields will be elements of the Lie algebra, they will not commute. These gauge fields are consequently referred to as non-abelian gauge fields. Alternatively, they are called Yang-Mills fields, Yang and Mills (1954) being the first to consider such theories.

In the last section we already indicated how a theory generalized for a set of similar fields gives rise to new symmetries. In fact, phase

[57] See, for example, Huang (1982), Ch. 4, or Quigg (1983), Ch. 4.

symmetry, or $U(1)$ gauge symmetry, for scalar fields resulted from the simplest such generalization, and spinor fields are by construction multicomponent objects. Consequently, we construct multiplets containing similar fields (i.e., either scalar or spinor fields) in such a way that they transform among themselves under an irreducible representation of a Lie group. We wrote down the general infinitesimal transformation for such a group at the end of the last section. We can write the general finite transformation (element of the Lie group) as

$$U = e^{-i\omega_a L_a} = e^{-i\omega} \tag{IV.2.24}$$

with

$$\omega = \omega_a L_a, \tag{IV.2.25}$$

which is also a generalization of equation (IV.2.7). For the representations, in addition to being unitary, we require that det $U = 1$ (which removes the complex phases) so that the representatives of the L_a, are traceless. Unitarity also insures that these representatives are Hermitian. These groups are called $SU\,(N)$, where N is the dimension of the smallest nontrivial irreducible representation (the fundamental representation) and $(N^2 - 1)$ is the dimension of the Lie algebra (the number of generators). (S, for “special,” indicates that the condition $\det U = 1$ holds.)

Practically, we can represent these multiplets as column vectors; such as

$$\phi = \begin{pmatrix} \phi_1 \\ \cdot \\ \cdot \\ \cdot \\ \phi_n \end{pmatrix} \tag{IV.2.26}$$

so that the fields ϕ_i can be considered components of these vectors which are rotated into one another by the n-dimensional unitary representation matrices. We represent a collection of these multiplets by Ψ. Then if $L(\Psi, \partial_\mu \Psi) = L(U\Psi, \partial_\mu U\Psi)$, this Lagrangian is globally gauge invariant.

Similarly under a local gauge transformation

$$\Psi \to U(x)\Psi(x) \tag{IV.2.27a}$$

$$\partial^\mu \Psi \to U(x)\partial^\mu \Psi(x) + [\partial^\mu U(x)]\Psi(x), \tag{IV.2.27b}$$

or, in terms of an infinitesimal transformation,

$$\delta\Psi(x) = -i\omega(x)\Psi(x) \tag{IV.2.28}$$

$$\delta_\Psi[\partial_\mu \Psi(x)] = -i\omega(x)\partial^\mu \Psi(x) - i[\partial^\mu \omega(x)]\Psi(x). \tag{IV.2.29}$$

Again, we would like $\partial^\mu \Psi$ to transform like Ψ, which means cancelling the second term on the right side in equation (IV.2.29). To this end we introduce the covariant derivative

$$D^\mu \Psi(x) \equiv [\partial^\mu + igA^\mu(x)]\Psi(x). \tag{IV.2.30}$$

We define

$$A^\mu(x) \equiv A^\mu_a(x) L_a \tag{IV.2.31}$$

as an element of the Lie algebra since $\omega(x)$ is such an element. Consequently, there will be N gauge fields, A^μ_a. We desire the transformation property of $A^\mu_a(x)$ to lead to the cancellation of the unwanted term in equation (IV.2.29). Since

$$D^\mu \Psi \to [\partial^\mu + ig(A^\mu + \delta A^\mu)](\Psi + \delta\Psi), \tag{IV.2.32}$$

the infinitesimal change in $D^\mu\Psi$ (to first order) is

$$\delta_\Psi(D^\mu\Psi) = i\omega D^\mu\Psi + ig\{\delta A^\mu - (1/g)\partial^\mu\omega + i[\omega, A^\mu]\}\Psi. \tag{IV.2.33}$$

$D^\mu\Psi$ transforms as desired if the last term above vanishes; i.e.,

$$\delta A^\mu_a(x) = (1/g)\partial^\mu\omega(x) - i[\omega(x), A^\mu(x)], \tag{IV.2.34}$$

which is the desired transformation law for A^μ. This can also be written in terms of the individual gauge fields as

$$\delta A_a^\mu(x) = (1/g)\partial^\mu \omega_a(x) + C_{abc}\omega(x) A_c^\mu(x), \tag{IV.2.35}$$

where the C_{abc} are the Lie algebra structure constants defined in the last section and are completely antisymmetric.

To find the kinetic term for the A_a^μ, we once again consider

$$\begin{aligned}(1/ig)[D_\mu, D_\nu] &= \partial_\mu A_\nu - \partial_\nu A_\mu + ig[A_\mu, A_\nu] \\ &\equiv F_{\mu\nu},\end{aligned} \tag{IV.2.36}$$

where the commutator does not vanish this time. If we define $F_a^{\mu\nu}$ by

$$F^{\mu\nu} \equiv F_a^{\mu\nu} L_a, \tag{IV.2.37}$$

then it can be shown that

$$F_a^{\mu\nu} = \partial^\mu A_a^\nu - \partial^\nu A_a^\mu - g C_{abc} A_b^\mu A_c^\nu \tag{IV.2.38}$$

which is our field tensor. This, of course, is not gauge invariant, but is gauge-covariant; however,

$$L_\gamma = -(1/4) F_a^{\mu\nu} F_{a\mu\nu} \tag{IV.2.39}$$

is gauge-invariant as well as Lorentz invariant.

Our scalar and spinor matter current, which couples to A_a^μ is now

$$\begin{aligned} J_a^\nu &= -ig[(D^\nu\phi)^* L_a\phi - \phi^* L_a(D^\nu\phi)] + (\bar\psi\gamma^\nu L_a\psi) \\ &= -ig(\phi^*\overleftrightarrow{\partial^\nu} L_a\phi) - g^2 A_b^\nu \phi^*\{L_a, L_b\}\phi + (\bar\psi\gamma^\nu L_a\psi). \end{aligned} \qquad \text{(IV.2.40)}$$

where {,} is an anticommutator.

The constant g, above, is called the gauge coupling constant. It was included to account for a difference of scale between the gauge fields and the matter fields. In quantum theory it determines the relative strength of the minimal coupling. If the gauge group is simple (cannot be written as a product of two or more Lie sub-groups), then the coupling constant can be absorbed into a redefinition of the gauge fields. If it is not simple, then there will be a different g for each independent Lie algebra and such a redefinition is not possible. The g's are the only arbitrary parameters of the theory.

From a constructive viewpoint global gauge invariance, while not forced upon us as a necessary symmetry, was at least strongly motivated by our requirements of Poincaré invariance and other physical requirements mainly concerned with causality; however, at first sight, there seems to be little such motivation for making this global symmetry into a local one. We note, however, that local gauge invariance is not a new symmetry, and in fact no new currents are introduced when a symmetry is gauged. So gauging a symmetry is not akin to imposing a new one. Rather, we hinted at the motivation for gauging a symmetry earlier—by gauging global gauge invariance, which is naturally a symmetry of the kinetic terms, we add a second order to our dynamics. We were motivated to consider terms with ∂_μ operators in order to give a dynamical aspect to our fields, not by any motivation from first principles, and we were restricted to two ∂_μ operators in a term by considerations of causality. Taking a symmetry which is closely connected to considerations of Poincaré invariance and of causality and generalizing it to obtain interactions of forces is a logical path to remain true to these considerations.

Once a gauge invariance is connected with a symmetry actually found in Nature (as was done when the connection $A_\mu(x)$ was identified as the electromagnetic potential,) a "conflict" with gauge invariance would imply a conflict with those basic principles which motivated it. In fact, gauge invariance is essential to proving the renormalizability of a quantum field theory.

Such conflicts are called gauge symmetry breaking. One possible way to break a gauge symmetry was provided by the example of a mass term for A_μ. Such a symmetry breaking is an explicit symmetry breaking. This sort of term straight-forwardly violates the gauge symmetry and eliminates it as a true symmetry of the system and of the equations of motion; although, if this sort of term is small, one might still have an approximate symmetry. Of more interest is a mechanism which breaks the symmetry of the system, but allows gauge symmetry to remain a symmetry of the solutions to the equations of motion. Such a mechanism would maintain gauge symmetry as a symmetry of Nature, but not as a physical symmetry—a situation we should be most interested in. Such a mechanism does in fact exist and it plays an important role in present day field theory. It is known as "spontaneous symmetry breaking." We shall be discussing this phenomenon in the next section.

IV.3 Spontaneous Symmetry Breaking

The Global Abelian Case

We now discuss the important phenomenon of spontaneous symmetry breaking. As before, our discussion will be in terms of classical fields, though we will often use quantum field theoretical terms in this discussion. The justification, in this case, rests on the fact that the classical structure so presented is equivalent to a quantum treatment in the tree approximation (no closed loops in Feynman graphs.)[58]

We start off with a general and qualitative description of this phenomenon. Spontaneous symmetry breaking is not a phenomenon limited to field theory and is in fact a common occurrence in Nature. It usually occurs when two or more symmetrical "forces" come into competition due to a change in some external parameter (such as temperature), and the system is forced to choose a relatively asymmetrical final state. This final state then does not exhibit the symmetry of the theory. For this reason the symmetry is sometimes said at this point to be hidden.[59]

As an idealized example, consider a three dimensional infinite ferromagnet (Heisenberg ferromagnet). At high temperatures the rotational invariance of the coulomb interaction is manifest, as the atomic spins are randomized. As it is cooled below the Curie point, however, the ferromagnet reaches its ground state in which all the spins are aligned. The direction of alignment is arbitrary until the direction is chosen spontaneously. The symmetry of the theory is now expressed by the fact that this ground state is infinitely degenerate. An arbitrary, non-implementable, rota-

[58] See Coleman (1973). This is true for gauge theories, but may break down for non-gauged field theories. See Frampton (1987), pp. 47–75, for a discussion of this point.

[59] See, for example, O'Raifeartaigh (1979) for a discussion along these lines.

tion of all spins would take the ferromagnet into an equivalent ground state. The infinite extent of the ferromagnet makes such a rotation of spins impossible. A person living in a universe containing such a ferromagnet would see a constant magnetic field in a certain direction and would not be aware of rotational invariance as an exact symmetry of the laws of Nature. If his measuring device interacted only weakly with this field he might suspect rotational symmetry is an approximate symmetry; however, if he knew the truth—that is, that the field is a permanent all-pervading feature of his universe—then he would properly label rotational symmetry as a broken symmetry of Nature.[60]As one final note, if we had considered a one-dimensional infinite ferromagnet instead, then the symmetry spontaneously broken would be a discrete one.

If, in quantum field theory there existed a field whose vacuum state was not invariant under certain transformations of the Hamiltonian of the theory, then this symmetry would be spontaneously broken and would not be manifest to us. Once again this vacuum state would be degenerate, but these vacua could not overlap, being states of an (infinite) quantum field theory; that is, they must lie in distinct Hilbert spaces.

These vacua, in order to be degenerate, must be non-empty. This immediately tells us that this field must be a scalar field; otherwise, the vacuum would have spin and Lorentz symmetry would be spontaneously broken, contrary to observation.

We consider first, then, the simplest such field, a single scalar field.[61] Let its potential term be

[60] So a broken symmetry results from global conditions. An approximate symmetry could, generally, be realized as a perfect symmetry under appropriate local conditions.

[61] See Huang (1982), Chs. 3 and 4, Quigg (1983), Ch. 5, and Coleman (1973) for the following discussion of spontaneous symmetry breaking in quantum field theories.

$$V(\phi) = (\mu^2/2)\phi^2 + (\lambda/4!)\phi^4, \tag{IV.3.1}$$

which yields a renormalizable field theory, and can be described as a self-interacting massive scalar field. λ must be positive, otherwise the Hamiltonian has no lower bound; μ^2 can be positive or negative and, hence, is not strictly identifiable as a mass term. The classical Hamiltonian is

$$H = \int d^3x[(1/2)(\partial_0\phi)^2 + (1/2)(\nabla\phi)^2 + V(\phi^2)]. \tag{IV.3.2}$$

It can be seen that a solution to the equations of motion which is a minimum of H is a constant $\phi(x) = \phi_0$, and is a minimum of $V(\phi_0)$. We recall that this theory is invariant under the discrete symmetry

$$\phi \to -\phi. \tag{IV.3.3}$$

If μ^2 is positive, the minimum of V is given by $\phi_0 = 0$ and this symmetry is manifest in the vacuum state; that is, it is invariant under the variation (IV.3.3). If μ^2 is negative, we find that the minima of the potential are

$$\phi_0 = \pm(-6\mu^2/\lambda)^{1/2}, \tag{IV.3.4}$$

so that there are two vacua (analogous to the one-dimensional ferromagnet.) One of these two solutions must be chosen by the system in its vacuum state. Let us assume it is the positive solution. To study the low-lying states of this system, it is then appropriate to expand about this solution. To this end we define the "shifted" field

$$\theta(x) = \phi(x) - a, \tag{IV.3.5}$$

where

$$a = (-6\mu^2/\lambda)^{1/2}. \tag{IV.3.6}$$

We can write the potential in terms of this field as

$$V(\theta) = (\lambda a^2/6)\theta^2 + (\lambda a/12)\theta^3 + (\lambda/4!)\theta^4. \tag{IV.3.7}$$

We still have a massive scalar field, now with $mass = (\lambda a^2/4!)^{1/2}$. Because of the θ^3 term, the symmetry is no longer manifest; although, equation (IV.3.7) is still invariant under the transformation

$$\theta \to -\theta - 2a. \tag{IV.3.8}$$

This transformation is, as stressed above, not implementable.

Consider next a complex scalar field with the potential function

$$V(\phi^*\phi) = \mu^2\phi^*\phi + \lambda(\phi^*\phi)^2. \tag{IV.3.9}$$

This theory possesses a global gauge invariance. Once again, if $\mu^2 > 0$, there will be a symmetric vacuum state at $\phi_0 = 0$. If $\mu^2 < 0$, however, we have spontaneous symmetry breaking with

$$\rho = (-\mu^2/2\lambda)^{1/2} e^{i\alpha_0} \equiv \phi_0 e^{i\alpha_0}/\sqrt{2} \tag{IV.3.10}$$

as a vacuum solution, where a_0, is an arbitrary real constant. There are now (as in the case of the three-dimensional ferromagnet) an infinite number of vacuum state solutions, which can be taken into one another by a phase transformation. As we know, however, once a system has spontaneously chosen a vacuum state, it knows nothing of the other distinct vacua (i.e., no transitions are possible between them.)

To determine the low-lying quantum states, we can again expand about a vacuum state in terms of classical small oscillations. We replace the complex fields by two real fields, such that

$$\phi(x) = [\phi_0 + \eta(x)] e^{i\alpha(x)}/\sqrt{2}, \tag{IV.3.11}$$

where $\eta(x)$ and $\alpha(x)$ have vacuum expectation values of zero, and for simplicity we have implicitly chosen the ground state such that $\alpha_0 = 0$. This yields an expansion in terms of $\eta(x)$ for the Lagrangian

$$L = 1/2\{\partial^\mu\eta\partial_\mu\eta - \lambda(2\phi_0 - \eta)^2\eta^2 + (\phi + \eta)^2\partial^\mu\alpha\partial_\mu\alpha\} \tag{IV.3.12}$$

or, to second order in the fields,

$$L \simeq 1/2(\partial^\mu\eta\partial_\mu\eta - 4\lambda\phi_0^2\eta^2) + 1/2(\phi_0^2\partial^\mu\alpha\partial_\mu\alpha). \tag{IV.3.13}$$

Here we see exhibited a massive scalar field $\eta(x)$, $mass = 2\phi_0\sqrt{\lambda}$, and a massless scalar field, $\alpha(x)$.

In classical language, examining equation (IV.3.11), we see that $\alpha(x)$ represents an angular mode of vibration, whereas $\eta(x)$ represents a radial vibration mode. The latter mode is opposed by the restoring force of the potential, the other mode is not. In quantum field theory, such modes translate into massive and massless particles respectively. Furthermore, this strange spin-less massless particle, called a "Goldstone boson," corresponds to zero-energy excitations which connect the possible vacua, but cannot, of course, cause transitions between them. The broken symmetry is sometimes said to be manifested in this mode (called the "Goldstone mode;") this language is appropriate if it is understood that this mode does not correspond to any canonical or unitary transformation on the system, which would otherwise violate the vacuum constraint.

The appearance of the Goldstone boson has been formulated into a theorem: for any local, manifestly Lorentz covariant field theory whose energy-momentum eigenvalues form a complete set and have positive definite norm, for every continuous[62] global symmetry of the system that is not a symmetry of the vacuum, there will occur in the theory a massless scalar particle whose properties are the same as the broken symmetry group generator. This is known as Goldstone's theorem. Unfortunately, no fundamental massless scalar particles are observed in Nature; so, in order for spontaneous symmetry breaking to operate in quantum field

[62] Spontaneously broken discrete symmetries do not yield Goldstone bosons.

theory, one or more of the assumptions that Goldstone's theorem relies on must be false. We shall see that, in fact, this happens for local gauge theories.

The Local Abelian Case

For systems with local gauge invariance, Goldstone's theorem breaks down and is replaced by the Higgs mechanism. The failure of Goldstone's theorem is directly attributable to the existence of the gauge fields and their associated gauge freedom. When quantizing gauge fields, the unphysical degrees of freedom associated with the gauge freedom must be eliminated by fixing the gauge. If one quantizes in a manifestly covariant gauge (i.e., the Lorentz gauge,) however, the theory contains states of negative norm (i.e., longitudinal photons;) whereas, if quantization is done in a gauge where only states of positive norm appear (such as the radiation gauge,) then manifest covariance is lost. Let us see how the Higgs mechanism works in a theory with a local $U(1)$ gauge symmetry.

We consider, then, the Lagrangian

$$L = (D^\mu \phi)^* (D_\mu \phi) - V(\phi^* \phi) - 1/4 F^{\mu\nu} F_{\mu\nu}, \tag{IV.3.14}$$

where

$$V(\phi^* \phi) = \lambda(\phi^* \phi - \rho_0^2)^2 \quad (\rho_0 \neq 0), \tag{IV.3.15}$$

which yields a spontaneously broken gauge symmetry. We call a scalar field such as this a Higgs field. The $U(1)$ symmetry of equation (IV.3.14) is manifested by the local gauge transformations

$$\delta A^{\mu}(x) = \partial_{\mu}\omega(x) \tag{IV.3.16}$$

and

$$\delta\phi(x) = -iq\omega(x)\phi(x) \tag{IV.3.17a}$$

$$\delta\phi^{*}(x) = iq\omega(x)\phi^{*}(x). \tag{IV.3.17b}$$

A lowest energy field solution for this system is

$$A_0^{\mu}(x) = 0$$

$$\rho_0(x) = \phi_0 e^{i\alpha_0}/\sqrt{2}.$$

The gauge field is empty in the vacuum state, as it must be. To examine the quantum states around the vacuum, it is useful once again to replace the complex scalar fields by a pair of real fields as in equation (IV.3.11). We also choose the vacuum state as before. From equations (IV.3.17a),(IV.3.17b) and (IV.3.11) we see that

$$\delta\eta = 0 \tag{IV.3.18a}$$

$$\delta\alpha = -q\omega(x) \tag{IV.3.18b}$$

under a local gauge transformation, so that

$$D_\mu \eta = \partial_\mu \eta \tag{IV.3.19a}$$

$$D_\mu \alpha = \partial_\mu \alpha + qA_\mu . \tag{IV.3.19b}$$

Let us substitute these covariant derivatives explicitly, directly into equation (IV.3.13). We obtain

$$L \simeq (1/2)(\partial^\mu \eta \partial_\mu \eta - 4\lambda \phi_0^2 \eta^2) + (1/2)\phi_0^2 \partial^\mu \alpha \partial_\mu \alpha + \phi_0^2 q A^\mu \partial_\mu \alpha \tag{IV.3.20}$$

$$+ (1/2)\phi_0^2 q^2 A^\mu A_\mu - (1/4)F^{\mu\nu} F_{\mu\nu} .$$

The radial mode-scalar field η has not been affected. Interpretation of the other two fields is difficult because of the quadratic cross term. We have not finished, however, since as mentioned earlier, our gauge must be fixed to eliminate the extra degree of freedom due to the gauge freedom of A^μ (in fact, if we were to interpret the penultimate term in equation (IV.3.20) as a mass term for A^μ, this system would posses five field degrees of freedom, whereas the Lagrangian (IV.3.14) we began with contains only four.)[63] Examining equation (IV.3.11), we can see that we can always choose a particular gauge such that $\alpha(x)$ is identically zero. Such a gauge is

[63] Scalar fields represent one field degree of freedom, massless vector fields two, and massive vector fields three.

$$\tilde{\phi}(x) = e^{-i\alpha(x)}\phi(x) = [\phi_0 + \eta(x)]/\sqrt{2} \qquad \text{(IV.3.21a)}$$

$$\tilde{A}_\mu(x) = A_\mu(x) + \partial_\mu \alpha(x). \qquad \text{(IV.3.21b)}$$

In this gauge our Lagrangian for small fields becomes

$$L \simeq 1/2(\partial^\mu \eta \partial_\mu \eta - 4\lambda \phi_0^2 \eta^2) + 1/2(\phi_0^2 q^2 \tilde{A}^\mu \tilde{A}_\mu) - (1/4)F^{\mu\nu}F_{\mu\nu}. \quad \text{(IV.3.22)}$$

The particle spectrum now consists of a massive scalar field, *mass* $= 2\phi_0\sqrt{\lambda}$, and a massive vector gauge field, *mass* $= \phi_0 q$. There is no α field as we explicitly intended by this choice of gauge.

In classical language, α was the (Goldstone) mode which manifested the symmetry broken by the degenerate ground state. By coupling this gauge symmetry to the dynamical field A_μ, however, this mode can no longer be permitted since this would require non-zero energy oscillations in the field A_μ which is zero (empty) in the ground state. Instead, the degree of freedom the Goldstone represented is now manifested in the mass of the gauge boson.

We see that the Higgs mechanism is also a mechanism for producing massive gauge fields without explicitly adding a gauge symmetry violating term. The appearance of the mass term in the Lagrangian (IV.3.22) differs from an ad hoc insertion of a mass term because this Higgs-generated term comes along with higher order interaction terms (which are not included in equation (IV.3.22)) which together preserve a non-implementable gauge invariance of the Lagrangian. This fact is most important in the full quantum theory where renormalization requires gauge invariance. Symmetry of the Lagrangian under gauge transformations even of this non-implementable form is sufficient for the cancellation of the ultraviolet divergences of the theory.

The gauge we chose above is called the physical gauge because only physical fields appear in the Lagrangian. It is also called the unitary gauge because only Feynman propagators for physical fields appear in the S-matrix. In other gauges, such as the Lorentz gauge, the Goldstone boson does not disappear, but remains as an unphysical constrained field. As one expects, however, since physical quantities, such as scattering amplitudes, are gauge-independent, the Goldstone boson must disappear from these when calculated in any gauge. To understand how this can happen in gauges other than the physical gauge, recall the Gupta-Bleuler mechanism in quantum electrodynamics. There the time-like component of the photon field, which by itself gives a non-unitary contribution to the S-matrix, is exactly cancelled by the longitudinal component of the photon field in on-shell calculations. The Higgs mechanism operates instead to cancel the time-like component of the gauge field with the Goldstone boson (thereby leaving the gauge field with three degrees of freedom instead of two, as required.)[64]

The Non-Abelian Case

We now examine spontaneous symmetry breakdown of theories with general gauge groups.[65] We postulate a set of Higgs fields which we denote by the vector ϕ. These fields are invariant under the global transformations specified by the gauge groups, G, of the theory:

[64] Cf. O'Raifeartaigh (1979) for the argument.

[65] See Huang (1982) for this discussion.

$$\delta\phi = -i\omega\phi = -i\omega_a L_a \phi. \tag{IV.3.23}$$

The potential function of ϕ has a minimum, taken to be zero, at a non-zero value $\phi = \rho$; ρ is independent of x. The degeneracy of ρ is expressed by transformations of it under the gauge group

$$\delta\rho = -\omega_a L_a \rho. \tag{IV.3.24}$$

As opposed to the $U(1)$ case, however, not all ρ's so obtained need be independent; i.e., $\delta\rho$ may be zero for some ω. In other words, there may be a subset of the elements of G—which will form a proper subgroup—which leave the vacuum invariant. We call this subgroup, H, the "little group." Its Lie generators, l_a, are a subset of the generators of G:

$$[l_\alpha, l_\beta] = iC_{\alpha\beta\gamma} l_\gamma. \tag{IV.3.25}$$

For these generators equation (IV.3.24) becomes

$$\delta\rho = -i\omega_\alpha l_\alpha \rho = 0, \tag{IV.3.26}$$

which yields

$$l_\alpha \rho = 0, \tag{IV.3.27}$$

since the ω_α are arbitrary. We say that the little group generators “annihilate the vacuum,” or that they are unbroken generators. H, then, is a remaining symmetry group after spontaneous symmetry breaking—the symmetry group G of the original full theory is broken down to H.

We can form distinct cosets of G with respect to H: $H, U_1H, U_2H\ldots$ The members of U_iH are of course not members of H; therefore, the number of generators of this coset space[66] G/H is equal to the number of generators of G minus the number of generators of H, a number we will label by K. G/H is not necessarily a group; in fact, it is only a group if H is a normal subgroup (i.e., every right coset is a left coset.) We can summarize by writing $G = H \cup G/H = H \cup (G - H)$.

The generators then fall into two sets. There are $(N - K)$ unbroken generators (N=number of generators of G) given by equation (IV.3.27). There are also K broken generators:

$$L_j\rho \neq 0. \tag{IV.3.28}$$

How the generators actually divide up between these two sets depends on the choice of the vacuum state, but, of course, the number in each set is independent of this choice. Since from Goldstone's theorem there must be a Goldstone boson for each broken generator, there must be K Goldstone bosons in this theory. The number of Goldstone bosons in a theory then depends strictly on group theoretic properties; namely, on the dimensions of the Lie algebras of G and H.

[66] The elements of the coset space are, of course, not transformations but cosets; however, there is a one-to-one correspondence between these cosets and the generators of G not in the set l_α. See, for instance, Herstein (1975), pp. 49-64.

The dimension of the representational space, naturally, depends upon what representation of the gauge group the Higgs field transforms under. Let us take this to be R-dimensional. This means that ρ has R components. Since equation (IV.3.28) represents K independent vectors, there is a mapping of the K generators to a K-dimensional subspace of the representation space. We call this the Goldstone space. The remaining $(R-K)$-dimensional subspace we call the Higgs space. There are, correspondingly, $(R-K)$ non-Goldstone fields. Of course if this number is zero, the little group is empty.

We straight-forwardly now generalize our approach for $U(1)$ spontaneously broken symmetry. We choose our vacuum state such that the Goldstone space is empty, and shift our non-Goldstone fields such that the newly defined fields also have zero vacuum expectation values. By next introducing the Yang-Mills fields, we institute the Higgs mechanism. Our vacuum state is then given by

$$\rho = \begin{pmatrix} 0 \\ \tilde{\rho} \end{pmatrix} \begin{matrix} K-dimensional\ Goldstone\ space \\ (R-K)-dimensional\ Higgs\ space \end{matrix}$$

$$(A_a^\mu)_0 = 0. \tag{IV.3.29}$$

We then fix the gauge to be the physical gauge so that the Goldstone space is identically empty. Low-lying state solutions are then given by

$$\tilde{\phi}(x) = \begin{pmatrix} 0 \\ \tilde{\rho} + \eta(x) \end{pmatrix}$$

$$\tilde{A}^\mu_a(x) \; small. \tag{IV.3.30}$$

$\eta(x)$ is an $(R–K)$-component field and is also small.

To demonstrate the particle spectrum we note that to second order in these small fields

$$\begin{aligned} V(\phi) &= \frac{\eta_n \eta_m}{2} \left[\frac{\partial^2 V}{\partial \eta_n \partial \eta_m} \right]_{\phi=\rho} \\ &= (1/2)\eta_n \left[\frac{\partial^2 V}{\partial \eta_n \partial \eta_m} \right]_{\phi=\rho} \eta_m \\ &\equiv (1/2)(\eta, V''(\rho)\eta), \end{aligned} \tag{IV.3.31}$$

where we have defined the inner product (,); also

$$\begin{aligned} J^\mu_a &= -g^2 A^\nu_b \phi^\dagger \{L_a, L_b\} \phi \\ &= -g^2 \rho^\dagger \{L_a, L_b\} \rho A^\mu_b \; + \; interaction\, terms \end{aligned}$$

from equation (IV.2.40), so that

$$A^a_\mu J^\mu_a = -g^2 A^a_\mu \rho^\dagger \{L_a, L_b\} \rho A^\mu_b \; + \; interaction\, terms. \tag{IV.3.32}$$

These are the terms in the Lagrangian from which we obtain the mass matrices:

$$\left(\mu^2\right)_{nm} \equiv \left[\begin{array}{ccc} 0 & | & 0 \\ -- & - & -- \\ 0 & | & V''(\rho) \end{array}\right] \begin{array}{l} \textit{Goldstone space} \\ \\ \textit{Higgs space} \end{array} \tag{IV.3.33}$$

$$\left(M^2\right)_{ab} \equiv (1/2)g^2\rho^\dagger \left\{L_a, L_b\right\}\rho = \left[\begin{array}{ccc} (M^2)_{ij} & | & 0 \\ -- & - & -- \\ 0 & | & 0 \end{array}\right] \begin{array}{l} \textit{Goldstone space} \\ \\ \textit{Higgs space} \end{array}$$

The first matrix gives the masses in the Higgs sector: $(R-K)$ massive scalar particles. The second matrix gives the masses in the gauge boson sector: K massive vector bosons and $(N–K)$ massless vector bosons. If we count the number of independent components these fields represent, we find $(2N+R)$—the number before spontaneous symmetry breaking: i.e., N massless vector gauge fields and R Higgs scalars.

The Weinberg-Salam Model

We now briefly present an example of a successful application of the ideas so far presented in this section and the last. This is the Weinberg-Salam model of the electroweak interactions.[67] As we "construct" this theory, we choose those characteristics that we are free to choose in conformance with our low-energy knowledge of the particle spectrum and their interactions. These characteristics are the multiplets our physical particles are to appear in, the gauge group whose representations these particles are to transform under, and the particular vacuum state structure that

[67] We follow, for the most part, Huang's (1982), Ch. 6, treatment.

the theory will contain. Of course, the first two of these characteristics are related—the multiplet structure of the Lagrangian yields its internal symmetries from which our gauged symmetries must be chosen, or given the gauge structure of the theory, our fundamental particle multiplets must transform under a representation of the gauge group (usually the fundamental one.)

Weak interactions are known to violate parity conservation maximally: only left-handed particles carry weak charge. Particle states with definite transformation properties under the action of the underlying gauge group must then be states of definite chirality. This suggests that we should use Weyl spinors as our representation of the Fermi fields. Another observed symmetry property of the weak interactions is that weak charged particles always appear in pairs. In the lepton sector, for instance, a left-handed electron always occurs with its neutrino. This symmetry is associated with lepton number. Hence our left-handed particles should appear as doublets; e.g.,

$$L \equiv \begin{pmatrix} \nu_L \\ e_L \end{pmatrix}. \tag{IV.3.34}$$

Because the neutrino is believed to be massless, there is no ν_R; consequently, the right-handed electron must transform alone, as a singlet:

$$R \equiv e_R \tag{IV.3.35}$$

We also include a Higgs doublet in our theory,

$$\phi = \begin{pmatrix} \phi_+ \\ \phi_0 \end{pmatrix} \tag{IV.3.36}$$

where ϕ_+, and ϕ_0 are positively charged and neutral scalar fields, respectively. The motivation for introducing this field is that the weak interaction is a short range force, which must be mediated by massive vector bosons. We can pick the vacuum state of this Higgs field so as to cause spontaneously symmetry breakdown and thereby give mass to these gauge bosons. We note also that an explicit mass term for the fermion fields must be of the form

$$m\bar{\psi}\psi = (\bar{L}R + \bar{R}L). \tag{IV.3.37}$$

But, since L and R transform so differently, this term cannot have the correct transformation properties to preserve weak gauge symmetry. In other words, a weak charge eigenstate is not a mass eigenstate. These masses can arise, however, also through an interaction with the Higgs field. Such a term of the form

$$L_{H-F} = \bar{L}\phi R + \bar{R}\phi^\dagger L \tag{IV.3.38}$$

has the correct symmetry properties (definite lepton number and chirality.) A nonzero vacuum expectation value for ϕ will very clearly yield a mass term from this equation of the form (IV.3.37) (this term will also yield the other requisite interaction terms between ϕ and L and R.)

We can now write down our Lagrangian. For simplicity, we will consider only one lepton family, the electron and its neutrino. The treatment is identical for additional families. Our ungauged Lagrangian is

$$L_0 = \bar{L} i \partial\!\!\!/ L + \bar{R} i \partial\!\!\!/ R + (\partial\phi)^\dagger(\partial\phi) - V(\phi^\dagger\phi) - (m/\tilde{\rho})(\bar{L}\phi R + \bar{R}\phi^\dagger L), \quad \text{(IV.3.39)}$$

where

$$V(\phi^\dagger\phi) = \lambda(\phi^\dagger\phi - \rho^2)^2. \quad \text{(IV.3.40)}$$

We have chosen the factor multiplying the last term in anticipation of its result after symmetry breaking. This Lagrangian is globally invariant under $SU(2)$; this symmetry is called weak isospin. This gauge group has three generators which we denote by the vector $\mathbf{t}$. The fields transform under the fundamental representation of these generators, $\boldsymbol{\tau}/2$, the familiar Pauli matrices:

$$\begin{aligned} &L \to e^{-i\boldsymbol{\omega}\cdot\boldsymbol{\tau}/2} L \\ &R \to R \\ &\phi \to e^{-i\boldsymbol{\omega}\cdot\boldsymbol{\tau}/2} \phi \end{aligned} \quad \text{(IV.3.41)}$$

Equation (IV.3.39) also contains two global $U(1)$ symmetries:

$$\begin{aligned} &L \to e^{-iI\theta} L \\ &R \to e^{-i\theta'} R \\ &\phi \to e^{-iI(\theta-\theta')} \phi, \end{aligned} \tag{IV.3.42}$$

where I is the unit two by two matrix.

These two transformations can be reshuffled (i.e., into two new linear combinations) so that we can identify one $U(1)$ symmetry with lepton number N, such that R and L have $N = 1$ and ϕ has $N = 0$. The other $U(1)$ symmetry cannot be the $U(1)$ of electric charge, since the doublets are not electric charge eigenstates. We call it weak hypercharge. Just as spontaneous symmetry breaking will yield physical states of definite mass, so too do we expect it to yield physical states that are charge eigenstates. Consequently, we assume a linear relationship between the eigenvalues of weak isospin, weak hypercharge and electric charge. We make the assignments of t_0, then, to obey the rule

$$q = t_3 + t_0, \tag{IV.3.43}$$

where q is the electric charge generator. These considerations fix our $U(1) \otimes U(1)$ transformations as follows

$$\begin{aligned} &L \to e^{-iI(\alpha t_0+\beta N)} L && (t_0 = {}^1\!/_2, N = 1) \\ &R \to e^{-i(\alpha t_0+\beta N)} R && (t_0 = -1, N = 1) \\ &\phi \to e^{-iI(\alpha t_0+\beta N)} \phi && (t_0 = {}^1\!/_2, N = 0) \end{aligned} \tag{IV.3.44}$$

There is no evidence that lepton number is gauged by Nature; therefore, we take the gauge group of the Lagrangian (IV.3.39) to be the product group $SU(2) \otimes U(1)$ of the two independent gauge symmetries of weak isospin and weak hypercharge.

It is this symmetry, then, that we next gauge by introducing the appropriate covariant derivatives and the dynamical terms for the gauge fields. We obtain

$$\begin{aligned} L = {} & \bar{L} i \not{D} L + \bar{R} i \not{D} R + (D^{\mu}\phi)^{\dagger} \cdot (D_{\mu}\phi) - V(\phi^{\dagger}\phi) - (m/\rho_0)(\bar{L}\phi R + \bar{R}\phi^{\dagger} L) \\ & - (1/4)(G^{\mu\nu}G_{\mu\nu} + H^{\mu\nu}H_{\mu\nu}) \end{aligned} \tag{IV.3.45}$$

with

$$D^{\mu} = \partial^{\mu} + ig\mathbf{W}^{\mu} \cdot \mathbf{t} + ig' W_0^{\mu} t_0, \tag{IV.3.46}$$

where $\mathbf{W}^{\mu}$ and W_0^{μ} are the weak isospin and hypercharge gauge fields, respectively, g and g' are their coupling constants, respectively, and $G^{\mu\nu}$ and $H^{\mu\nu}$ are the field tensors.

Since there is a 1-1 correspondence between generators and gauge fields, the linear relationship (IV.3.43) between generators implies a similar relationship between the associated gauge fields, which we write as

$$A^\mu = W_3^\mu \sin\theta + W_0^\mu \cos\theta_w. \tag{IV.3.47}$$

We write the combination in this way so that we can easily form the requisite orthogonal combination,

$$Z^\mu = W_3^\mu \cos\theta_w - W_0^\mu \sin\theta_w. \tag{IV.3.48}$$

We can invert these two equations to obtain

$$\begin{aligned} W_3^\mu &= A^\mu \sin\theta_w + Z^\mu \cos\theta_w \\ W_0^\mu &= A^\mu \cos\theta - Z^\mu \sin\theta_w. \end{aligned} \tag{IV.3.49}$$

θ_w is called the Weinberg angle and is a free parameter to be determined by experiment. Like A^μ, Z^μ must be an electrically neutral gauge field.

Equation (IV.3.47) also implies a relationship between the coupling constants. We can write

$$e = g \sin\theta_w = g' \cos\theta_w \tag{IV.3.50}$$

and

$$\tan\theta_w = g'/g$$
$$e = gg'/(g^2 + g'^2)^{1/2}, \tag{IV.3.51}$$

where $-e$ is the electron charge. We can now rewrite the covariant derivative (IV.3.46) in terms of these new fields:

$$D^\mu = \partial^\mu + ig(W_1^\mu t_1 + W_2^\mu t_2) + ieqA^\mu + ieq'Z^\mu, \tag{IV.3.52}$$

where q' is the generator to be associated with Z^μ,

$$q' = t_3 \cot\theta_w - t_0 \tan\theta_w. \tag{IV.3.53}$$

At this point, these new definitions do not represent physical gauge fields because they are not associated with any symmetry of the Lagrangian (IV.3.45). This will change, however, when we institute spontaneous symmetry breaking and choose the vacuum state so that the electric charge generator is the only generator to remain unbroken. Such a vacuum state solution is

$$\rho = \begin{pmatrix} 0 \\ \tilde{\rho} \end{pmatrix}, \tag{IV.3.54}$$

thereby placing the charged components of ρ in the Goldstone space and leaving the vacuum uncharged. For small ϕ, in the physical gauge, we have as usual

$$\phi = \begin{pmatrix} 0 \\ \tilde{\rho} + \eta(x) \end{pmatrix}. \tag{IV.3.55}$$

Let us now examine the term in the Lagrangian that gives rise to the gauge field masses. This is

$$\begin{aligned}(D^\mu \rho)^\dagger (D_\mu \rho) \equiv |D^\mu \rho|^2 = |\{\partial^\mu + ig[(1/2)(W_1^\mu - iW_2^\mu)\tau_+ \\ + (1/2)(W_1^\mu - iW_2^\mu)\tau_-] + ieqA^\mu + ieq'Z^\mu\} \begin{pmatrix} 0 \\ \tilde{\rho} \end{pmatrix} |^2,\end{aligned} \tag{IV.3.56}$$

where we have, naturally, put the generators in their fundamental representation. In this representation, using equation (IV.3.43),

$$q = \tau_3/2 + It_0 \tag{IV.3.57}$$

and, in particular,

$$q = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

for the Higgs field. This gives

$$q\rho = 0, \tag{IV.3.58}$$

or, in other words, the electric charge generator annihilates the vacuum, as planned. Correspondingly, the photon remains massless. The other gauge fields, however, do not escape the Higgs mechanism. Writing out equation (IV.3.56) yields the explicit mass terms for these fields

$$|D_\mu \rho|^2 = (1/2) g^2 \tilde{\rho}^2 \left[W_+^\mu W_{\mu -}^* + (1/2)(\cos^2 \theta_w) Z^\mu Z_\mu \right] , \qquad \text{(IV.3.59)}$$

where

$$W_\pm^\mu = (1/\sqrt{2})(W_1^\mu \pm i W_2^\mu) \qquad \text{(IV.3.60)}$$

is a complex (electrically charged) gauge field. These particles have been observed and their measured masses are in good agreement with these mass terms when the experimentally determined values of θ_w and the coupling constants are used to calculate them.

Writing out the other terms in the Lagrangian (IV.3.45) in terms of these fields straight-forwardly yields the interactions between the fields and, as discussed above, the mass terms for the electrons.

IV.4 Summary and Implications

In the first section of this chapter we discovered that for field theories a symmetry of Nature that is not a physical symmetry is equivalent to an invariance of the solutions to the equations of motion, or, in other words, an invariance of the physical field configurations. In a quantum field theory the physical field configurations correspond to the physical on-shell S-matrix elements; a physical symmetry, or a symmetry on physical state space, is in this case a symmetry on the Hilbert space constructed from the state vectors (called the Fock space.) This latter symmetry is expressible in terms of unitary transformations of the operators which act in this space; namely, the quantized fields.

We also found that when there existed an invariance of the Action, then there was correspondingly a symmetry of Nature and a symmetry on physical state space. This then leads to the existence of a continuity equation; i.e., a conserved current. In local gauge theories, it is the current associated with gauge symmetry which couples to the force carrying particle, the gauge boson. It is this coupling to a conserved current which insures that the gauge boson propagator can be made finite through a simple rescaling: that is, it insures charge renormalization. (It also insures that no radiative corrections to this propagator can shift its pole from $k^2 = 0$; i.e., it cannot gain a mass.) Let us consider spontaneous symmetry breaking in terms of this symmetry formalism. It is, of course, a symmetry of state space that is broken when the ground state of a theory becomes degenerate. The solutions to the equation of motion, or, equivalently, the physical on-shell S-matrix elements, remain invariant under the transformations of the symmetry group. (This is why, in certain gauges, unphysical particle states may appear, but do not contribute when physical quantities are calculated.)

We need to note here that, although the mechanism of spontaneous symmetry breaking has been used very successfully (such as in the Weinberg-Salam theory), it relies heavily on the ad hoc insertion into the theory of (unobserved) fundamental particles.

Superconductivity is the closest example of spontaneous symmetry breaking in another field. In fact, in the phenomenological Landau-Ginzberg model, a potential which is identical in form to the Higgs field potential is added to the Lagrangian. In this case, the role of “fundamental scalars," however, is taken by bound Cooper pairs of electrons ($|\phi|^2$ is the probability density for these Cooper pairs.)

Needless to say, there have been attempts to find such bound states of known fermions to take the role of the Higgs field, but no such attempts have been successful.[68] One attempt in another area, however, has been successful. In the partially conserved axial current (PCAC) hypothesis, one assumes a spontaneous breakdown of chiral symmetry $[SU(2)]_A$. Here, the (unusually light scalar) pions are taken as the resulting Goldstone bosons (since this is not a gauged symmetry) in an idealized massless limit. So here we have a global approximate spontaneously broken symmetry.

This example of PCAC highlights the difference between an approximate symmetry and a broken one. Chiral symmetry is a global symmetry of the fermion Lagrangian only in the massless limit. It is further assumed that this approximate symmetry (since fermions are not massless and neither are the pions) is spontaneously broken. An approximate symmetry can be considered in some reasonable limit to yield a perfect symmetry; hence, symmetry-conservation formalism can be invoked (i.e., one can define a partially conserved current.) A broken symmetry, however, has global effects (existence of Goldstone or Higgs bosons) which cannot be removed through small approximations.

In the Weinberg-Salam model a $SU(2) \otimes (1)$ symmetry is spontaneously broken, yielding a perfect $U(1)$ symmetry (of Q.E.D.) and a broken $SU(2)$ symmetry (to be associated with the weak interactions.) The "unified" theory, even after symmetry breaking, is still renormalizable, and this fact relies on the special feature of spontaneous symmetry breaking: that there still remains the (hidden) non-implementable gauge

[68] This is called dynamical symmetry breaking.

symmetry of the original product group. However, this spontaneous symmetry breaking allows one to consider the $U(1)$ symmetry as associated with a separate interaction; i.e., the theory of Q.E.D. This theory is renormalizable because it is based on an unbroken gauge symmetry. On the other hand, when one tries to consider the $SU(2)$ symmetry as being associated with a separate interaction, one finds a non-renormalizable theory—the theory of the weak interactions. If we accept the Weinberg-Salam model, then, there is no renormalizable theory describing separately the weak interaction.

Renormalizability is a requirement of consistency for quantum field theories. If a quantum field theory is non-renormalizable, it yields infinities for some physical results; results which are not merely wrong, but which are senseless. These physical results are interactions cross-sections; therefore, this indicates some inability to describe the weak interaction in causal terms. In fact, we find here an example of our general symmetry formalism where there exists a fundamental broken symmetry: we label the theory of the weak interactions as an essentially incomplete theory.

We find, then, a precedent for our description of quantum mechanics as an essentially incomplete theory. We have also seen in detail how in quantum field theory such an essentially incomplete theory results. We do not, however, expect the mechanism or methods of spontaneous symmetry breaking to be directly applicable to the solution in the non-relativistic quantum domain, since one-particle quantum mechanics is a much different theory than quantum field theories. In addition, spontaneous symmetry breaking is not possible in non-infinite systems: in quantum mechanics degenerate ground states overlap (they do not lie in distinct Hilbert spaces.)

V Quantum Mechanics as a Broken Symmetry

V.1 A Broken Symmetry Ontology

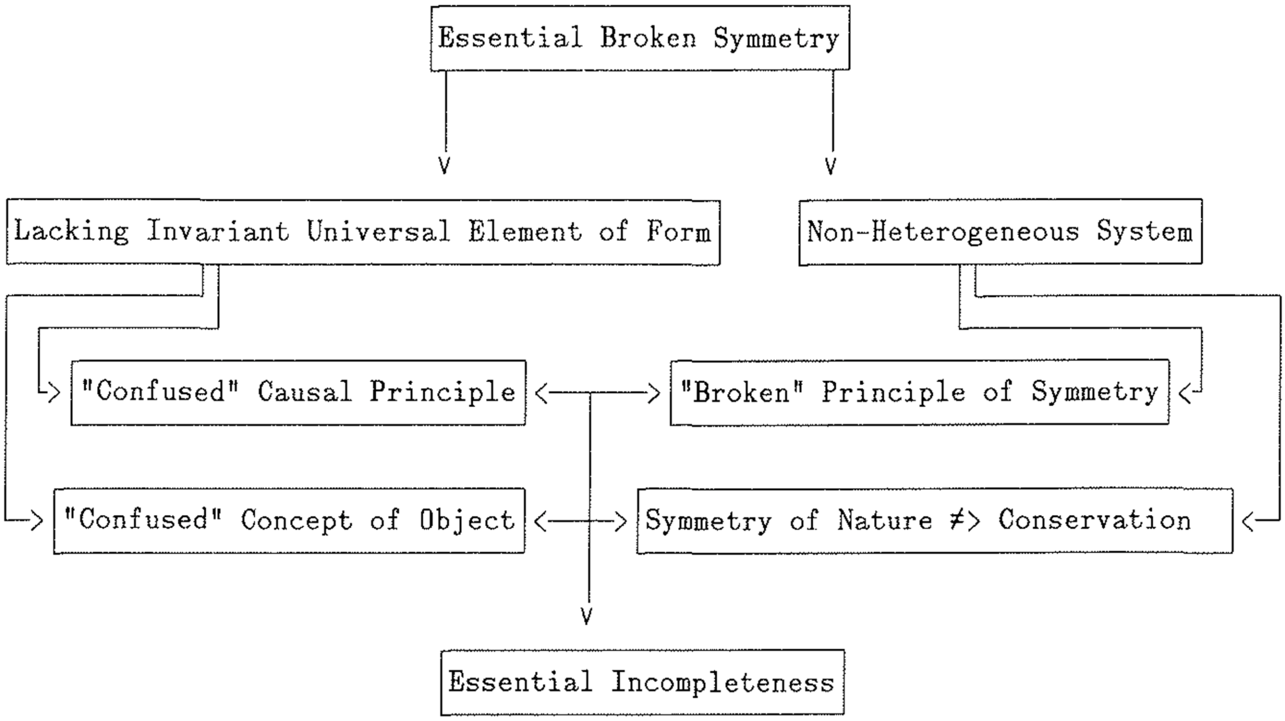
Let us now summarize and formally propose what we will call a broken symmetry ontology. We have established symmetry as a useful fundamental and primitive notion in physics. We propose, however, that in certain ontological domains the fundamental symmetry present is a broken symmetry. This broken symmetry is just as fundamental and primitive a notion as that of symmetry. There must be definite epistemological consequences for grounding our physics in these notions, and we have in fact taken note of these in Chapter II. The existence of a fundamental symmetry ("an invariant universal element of form") underlies our consistent application of the causal principle and our concepts of scientific objects. A broken symmetry ("a lacking invariant universal element of form") leads to an impaired use of the causal principle and a confused concept of object.

In terms of an analysis of symmetry, we found physical principles which parallel our epistemological analysis. In those ontological domains where there exists a fundamental symmetry, physical systems are analyzable in a heterogeneous manner. This leads to the general principle of symmetry and a generalized symmetry-conservation theorem. The principle of symmetry, which states that the symmetry of an isolated physical system either increases or remains the same, as it evolves according to the laws of Nature, was shown to be essentially a translation of the causal principle in terms of the notion of symmetry. The generalized symmetry conservation theorem, which states that for every symmetry of Nature there is a corresponding conserved quantity, is directly related to the consistency of the concept of a scientific object, since conserved quantities are intimately related to the consistent definition of objects.

In an ontological domain where there exists a broken symmetry we are forced to analyze physical systems in a non-heterogeneous manner. In non-heterogeneous systems, the principle of symmetry is inapplicable and the general symmetry-conservation theorem is invalid.

From either an epistemological or a physical symmetry based analysis, we see, then, that in an ontological domain founded on a broken symmetry, certain elements "normally" present in a description of our physics are lacking. We can say that our description in this domain is incomplete, or, to be more precise, it is essentially incomplete, since this description cannot be "completed." In particular, since a physical description is given in terms of physical states which are in turn provided by physical theory, we can say that theory in this domain admits of incomplete state descriptions. This broken symmetry ontology is diagrammed in the figure below.

Figure 1: *


Essential Broken Symmetry
Lacking Invariant Universal Element of Form
Non-Heterogeneous System
"Confused" Causal Principle
"Broken" Principle of Symmetry
"Confused" Concept of Object
Symmetry of Nature ≠> Conservation
Essential Incompleteness

V.2 Separability

We have seen from Jarrett's work that quantum mechanics is essentially incomplete. More specifically, we can say that its state descriptions are incomplete. We can go further and ask in what way these states are incomplete.

Howard (1985)[69] has asked this question and has come to the conclusion that quantum states are incomplete because they do not possess the property of "separability." Separability, or the separability principle, states that spatially separated systems possess separate real states, which is to say that there always exists separate probability measures for spatially separated systems.

Howard showed that he could decompose Jarrett's strong locality (the condition used in deriving the general Bell inequalities) into Einstein locality and separability, thereby demonstrating an equivalence between Jarrett completeness and separability. Given this demonstrated equivalence between separability and Jarrett completeness, we can then say that non-separability is a feature of quantum mechanics, which is the specific way in which quantum mechanics is incomplete. This means that for some previously interacting systems, such as in the EPR-Bohm setup, there are distinguishable (separated in space or time) systems (the two electrons) which can not be characterized as having their own set of intrinsic properties—a characteristic of spatially or temporally separated systems that we take for granted in classical mechanics.

Howard also pointed out that classical field theories, in particular general relativity, are explicitly separable, since their fundamental structure is well-defined at every point in the space-time manifold; that is,

[69] See also Howard (1985b).

they define a separate real state for every space-time point. This is only true, however, for the free-field versions of such theories. In particular, in general relativity the field is not well-defined at those points where there exists massive particles. Similarly, the electromagnetic field is well-defined at every space-time point for the free-field, but not at those points where there exists its sources, namely charges.

Also, we should note, quantum field theories, which are expressed in terms of creation and annihilation operators over the space-time manifold, clearly do not yield a field description well-defined at individual space-time points. The most successful formalism for describing quantum field theories, namely S-matrix theory, even more explicitly gives up the notion of a field defined at individual space-time points. Here, such a description is only available in the remote past or future with respect to some interaction.

V.3 Broken Dilation Invariance

We want, now, to fit quantum mechanics into our broken symmetry ontology. Our motivation for this is that we have found that quantum mechanics is an essentially incomplete theory, and that there is an impaired use of the causal principle (due to the unobservability of ψ) and a confused concept of object (the wave-particle dualism.) There obviously must be a fundamental broken symmetry upon which the quantum ontology is based. This broken symmetry must explain the strange sort of non-locality that is found in the EPR argument, and, in particular, must be related to the principle of separability. Hopefully, also this broken symmetry could be related to the formalism of quantum mechanics, most likely in one of its semiclassical formulations.

V QUANTUM MECHANICS AS A BROKEN SYMMETRY

We propose that broken dilation invariance is at the foundations of the quantum domain. As we pointed out in the last chapter, Weyl's original attempt to give a fundamental role to dilation invariance was rejected because it was seen not to be a fundamental symmetry of Nature. This was mainly because in the quantum domain there is a preferred scale determined by the fundamental constant $\hbar$. Moreover, this fundamental scale results from the basic feature of the quantum domain, namely quantization. In particular, massive particles have associated a de-Broglie wavelength. To be sure, the quantum aspect of Nature is to be found in the ontologically superior theory of quantum field theory. In fact, there, any perfect or approximate dilation symmetry present in the classical Lagrangian is always broken when the field is quantized and renormalized, since, once again, this fundamental feature always sets a preferred scale. We also found that, at the classical level, dilation invariance was associated with masslessness (the absence of dimensionful parameters.) The ultimate origin of mass is to be found most likely in the fundamental aspects of quantum effects (such as the phenomena of dimensional transmutation, where quantum effects explicitly provide a mass scale, and which is believed to be responsible for the mass of quarks.) However the absence of dilation invariance is to be expressed, the fact that this is a basic aspect of our world is to be directly attributed to the basic quantum aspect of physics. This aspect is, naturally, to be found most clearly in one-particle quantum mechanics where we study individual quantum systems. However, being a fundamental feature of the world itself, it expresses itself at the classical level by the existence of objects of definite mass and states of definite charge. Hence, we expect dilation invariance to be explicit in classical physics only where these effects can be avoided, in particular, in the free-field versions of classical field theories. In fact, these theories, such as free electromagnetism, are expressly dilation symmetric.

Now we can make direct connection with our observations about separability in physics. We found that free classical field theories are explicitly separable because they are well-defined at every space-time point. We see that separability is a property to be ascribed to a theory if it possesses dilation invariance, since if it assigns definite properties to every point in the space-time manifold, these properties cannot be affected by a relative change of scale, which keeps each point (each system) in the same relation to every other.

Conversely, a theory which is manifestly not dilation symmetric, such as quantum mechanics, will yield state descriptions which are not separable. In this case systems individuated only by a space-time separation do not possess individual sets of properties, so a dilation transformation is not well-defined.

V.4 Zeeman Causality and Non-Heterogeneity

To understand the strange non-locality found in the Bell-EPR analysis, in which there is not a violation of Einstein locality (i.e., signals are not transmitted superluminally,) but, instead, there is some sort of "spooky action at a distance,"[70] consider the analysis of causality presented by Zeeman (1964).

Zeeman defined the group of causal automorphisms, or, more simply, the causal group, as those one-to-one mappings on Minkowski space, M, which preserve the partial ordering on M, $x < y$, where an event at x

[70] Coined by Einstein. From Born (1971).

can influence an event at y (i.e., x are those points in the backward light cone of y.) He then showed that this causal group is equivalent to the group consisting of the Poincaré transformations and the dilations.[71]

Other researchers have shown that by replacing Zeeman's condition of maintaining the partial ordering with a condition for preserving ~~the magnitude of the velocity of light~~, and thereby relaxing Zeeman's demand of causality, one obtains an equivalence just with the Poincaré group.[72] And, of course, the Poincaré group is usually obtained by physicists by instituting the principle of relativity.

Correction: “the norms of timelike vectors,”

We see, then, that a general condition of causality—which we shall call “Zeeman causality”—implies a larger space-time symmetry group than that required by simple Einstein causality. This larger space-time symmetry group includes the dilations. Given that the quantum mechanical EPR states manifest the broken dilation invariance found in quantum mechanics, we expect them to manifestly violate Zeeman causality. Consequently, we can identify that property of a theory, or equivalently of the states descriptions it yields, such that it is Zeeman causal, with the property called strong locality (so called by Jarrett.) To be clear, let us call this property “Zeeman locality” and a theory which possesses Zeeman locality “Zeeman local.” The type of non-locality found in quantum mechanics is then a violation of Zeeman locality but not of Einstein locality.

Williams further demonstrated that the Poincaré group and the dilations preserves the norms of null cones (which is equivalent to preserving the speed of light), and that the timelike norm preserving group is a proper subgroup of this latter group.

This violation of Zeeman causality does not entail superluminal transmission of information, nor does it violate our fundamental notion of causality (as it cannot, by our discussion of Chapter II) in such a way to create causal paradoxes. Zeeman causality is, however, not a generalized

[71] It also contains the discrete symmetry of spatial inversion.

[72] See, for instance, Williams (1973).

form of causality, but the actual form of causality found in a strictly classical (no quantum effects) world. The way in which the causal principle is restricted in the quantum domain is by its restriction to Einstein causality.

Our broken symmetry ontology predicts that quantum mechanics will describe physical systems with a non-heterogeneous logical structure. We now examine this possibility in the context of the Bohm-EPR experiment.

Recall that non-heterogeneous systems contain equivalence relations connecting the cause and effect subsystems. In a consideration of the Bohm-EPR setup as a means of measuring the spin of a remote electron (i.e., one electron's spin is measured and the other's is inferred), we can consider the measured electron as being the cause subsystem and the remote electron as the effect subsystem. We know from our previous analysis that this causal connection cannot be an Einstein causal one; that is, there is no causal relation between the two systems as usually derived from special relativistic considerations. Rather, we can interpret the correlations (as they are usually referred to) between measurements made simultaneously on both electrons, as due to these connecting equivalence relations. These connecting equivalence relations, then, account for the stochastic interdependence of the outcomes of measurements made on the two electrons.

The "confused" concept of object offered in quantum mechanics can also be understood in terms of the EPR setup. In particular, we can discuss the wave-particle dualism by examining the original EPR argument, in which position and momentum measurements on two previously interacting particles prepared in a zero momentum state are considered. Here, there are the same correlations, and correspondingly the same connecting equivalence relations, this time connecting the properties of position and momentum (instead of spin.) We can, again, consider making a measurement on one particle to "determine" the properties of the other remote particle. Making a position measurement on one particle then determines the position of the other particle, and, hence, localizes it and determines it as having the property of being a particle. Making a momentum measurement, instead, on the first particle would determine the momentum of the other particle but not its position, and thereby determine it as having a wave property. Now, the absence of dilation invariance as a symmetry of these physical states (as required by our broken symmetry ontology) prohibits the assigning of certain unique conserved properties to the second particle. If one considers the solutions to the equation of motion as saying, for instance, that particle one will always have a momentum opposite to particle two, or, that particle one will always have a position the same distance from some origin as particle two, then these solutions are related by a symmetry (just as the solutions to the equations of motion are still gauge-symmetric in the Weinberg-Salam model after spontaneous symmetry breaking.) We can use a dilation transformation to reduce a wave solution to a point or the opposite transformation to expand a particle solution into a wave solution. Of course, these transformations are physically non-implementable, since dilation invariance is not a symmetry of the physical states.

More generally, since in an analysis of measurement in quantum mechanics one can take the EPR setup as the simplest type of measuring device—using one quantum system to measure another—a complete theory of measurement, in which a quantum mechanical description would be given for a classical measuring device, should be expressible similarly in terms of these connecting equivalence relations. We will not pursue such a program here.

V.5 Reevaluation of the Quantum Potential Approach

An equivalence between the eikonal equation of geometrical optics and the Hamilton-Jacobi equation, where the phase of the wave motion is equivalent to S, was first realized by Hamilton. In fact, it was this optical-mechanical analogy which lead Schrödinger to his equation.[73]

This connection can be demonstrated by showing that the Schrödinger equation reduces to the Hamilton-Jacobi equation in the short-wavelength limit, thereby running Schrödinger's argument backwards. The basis of the optical-mechanical analogy is the equating of the phase to $S/\hbar$, which leads to

$$\psi = e^{iS/\hbar}. \tag{V.5.1}$$

Substituting this into Schrödinger's equation, we find

$$\partial S/\partial t + [(\nabla S)^2/2m] + V = [(i\hbar)/(2m)]\nabla^2 S. \tag{V.5.2}$$

This is the Hamilton-Jacobi equation if the term on the right hand side can be neglected. This is true if

$$\hbar\nabla^2 S << (\nabla S)^2 \tag{V.5.3}$$

[73] See Lanczos (1986), Section 8.8, and Goldstein (1980), Section 10-8, for a discussion.

or

$$\hbar\nabla\cdot p << p^2, \tag{V.5.4}$$

since $p = \nabla S$. Substituting the de-Broglie wavelength $\lambda = h/p$, we find

$$(1/p)\nabla\cdot p << 2\pi/\lambda. \tag{V.5.5}$$

The condition (V.5.5) states that classical mechanics is the geometrical optics limit of quantum mechanics, since the wavelength must be small compared to the change in momentum of the particle, or, in other words, the potential should not vary greatly over a wavelength.

Bohm's and the stochastic approaches, in addition to examining the possibility when no approximation like that indicated by the relation (V.5.5) is made, also assume that the ψ-field should be written in terms of two independent functions. The additional function specifies an independent variation of the norm of ψ.

One way to see this generalization of Bohm is to take equation (V.5.1) and let

$$S \to S - i\hbar \ln R, \tag{V.5.6}$$

so that

$$\psi = Re^{iS/\hbar}. \tag{V.5.7}$$

In fact, if we apply the replacement (V.5.6) to equation (V.5.2) and we assume the continuity equation found by Bohm:

$$(\partial p/\partial t) + \nabla \cdot (p \nabla S/m) = 0, \tag{V.5.8}$$

which can be written as

$$(\partial \ln R/\partial t) + (1/m)\nabla S \cdot \nabla \ln R = -(1/2m)\nabla^2 S, \tag{V.5.9}$$

then we find

$$(\partial S/\partial t) + [(\nabla S)^2/2m] + V - (\hbar^2/2m)(\nabla^2 R/R) = 0, \tag{V.5.10}$$

which is just the modified Hamilton-Jacobi equation found by Bohm.

We see, therefore, that the existence of the quantum potential which is already a function the norm of ψ, can be directly attributed to a modification of the Action, whereby a term proportional to the norm is added to it. This term then sets a preferred scale proportional to $\hbar$, which consequently breaks dilation invariance. Since the norm of ψ can vary locally, this preferred scale can also vary from point to point. We can interpret Q, the quantum potential, as the connection established by these local variations. In fact, if there were no such local variations, Q would be zero; there would then be no locally preferred scale.

We see that the term added to the Action is not added as the usual interaction terms are added to the Lagrangian in field theory. There, such terms indicate an interaction at a space-time point: at the same time, these terms arise naturally out of considerations of maintaining a symmetry of the Action. Consequently, we can attribute the "non-local" aspect of this potential to the explicit symmetry breaking manner in which the replacement (V.5.6) is made.

We offer this example not as support for the quantum potential approach as a preferred interpretation, but, rather, as an example of how we can understand, in terms of our broken symmetry ontology, the at least partial success of describing quantum mechanics in this particular manner.

V.6 Broken Gauge Symmetry and its Relation to Broken Dilation Symmetry

We already indicated how the theory of the weak interactions can be understood as an essentially incomplete theory. Now that we have formally presented our broken-symmetry ontology, we can summarize how this theory fits this ontology.

The symmetry broken in this case is a gauge symmetry. In particular, the broken symmetry on which the theory of the weak interactions is based is a broken $SU(2)$ symmetry. This results in a non-heterogeneous structure in which the symmetry transformations of $SU(2)$ are based on the connecting equivalence relations. The way in which causality is restricted here is that for certain interactions described by the theory one finds infinite cross sections and non-unitary results.[74]

We have not yet clearly understood in what sense one finds a confused concept of object in this domain. We take note, though, that weak

[74] See, for instance, Leader and Predazzi (1982) for a discussion of the problems of an independent theory of the weak interactions.

charge eigenstates are definite states of chirality; hence, there is an intimate connection between the weak interaction and chiral symmetry. We speculate, therefore, that since anomalies (which are still not clearly understood) can easily be interpreted as a confusion in the concept of object in quantum field theory, chiral anomalies are due to the existence of the fundamental broken symmetry underlying the weak interaction.[75]

Finally, one may ask how quantum mechanics and the theory of the weak interactions can be based on similar ontologies and yet be such different theories and with different kinds of causal restrictions. Again, we only speculate that modern-version Kaluza-Klein theories will be successful in describing gauge symmetries as being equivalent to space-time symmetries in compactified dimensions. In this case, dilation and gauge symmetry (or its equivalent replacement) would be put on an equal footing—both would be space-time symmetries. One could further imagine that one could find an expanded group of causal automorphisms over this enlarged space-time, and a correspondingly enlarged generating group of space-time transformations. A broken gauge symmetry could then correspond to a restricted causal structure in the same way that broken dilation symmetry does.

[75] Broken dilation and conformal invariance in quantum field theory also leads to anomalies.

References

[1] Aspect, Alain, J. Dalibard and G. Roger (1982), "Experimental Test of Bell's Inequalities Using Time-Varying Analyzers", Phys. Rev. Let., **49**, 1804.

[2] Ballentine, L. E. (1970), "The Statistical Interpretation of Quantum Mechanics", Rev. Mod. Phys., **42**, 358.

[3] Belinfante, F. J. (1973), A Survey of Hidden-Variables Theories (Pergamon, Oxford).

[4] Bell, J. S. (1965), "On the Einstein Podolsky Rosen Paradox", Physics, **1**, 195.

[5] Birkoff, George David (1950), "The Principle of Sufficient Reason", in Collected Mathematical Papers, Vol. III (American Mathematical Society, New York).

[6] Bohm, D. (1951), Quantum Theory (Prentice Hall, Engelwood Cliffs, N.J.).

[7] — (1952a), "A Suggested Interpretation of the Quantum Theory in Terms of 'Hidden Variables', I", Phys. Rev., **85**, 166.

[8] — (1952b), "A Suggested Interpretation of the Quantum Theory in Terms of 'Hidden Variables', II", Phys. Rev., **85**, 180.

[9] — and Hiley, B. J. (1975), "On the Intuitive Understanding of Non-Locality as Implied by Quantum Theory", Found. of Phys., **5**, 93.

[10] — and — (1984), "Measurement Understood through the Quantum Potential Approach", Found. of Phys., **14**, 255.

[11] — and — (1985), “Unbroken Quantum Realism, from Microscopic to Macroscopic Levels”, Phys. Rev. Lett., **55**, 2511.

[12] Bohr, N. (1935), “Can The Quantum-Mechanical Description of Physical Reality Be Considered Complete?”, Phys. Rev., **48**, 696-702.

[13] Born, M. (1971), The Born-Einstein Letters (Walker. New York), parts quoted in N. David Mermin, “Is the Moon There When Nobody Looks? Reality and the Quantum Theory”, Phys. Tod., April 1985. p. 46.

[14] Boyer, Timothy H. (1966), “Derivation of Conserved Quantities From Symmetries of the Lagrangian in Field Theory”, Am. J. Phys., **34**, 475.

[15] — (1967), “Continuous Symmetries and Conserved Currents”, Ann. Phys., **42**, 445.

[16] Candotti, E., C. Palmieri and B. Vitale (1970), “On the Inversion of Noether’s Theorem in the Lagrangian Formalism. Il. – Classical Field Theory”, Nuovo Cimento, **70A**, 233.

[17] —, —, and — (1972), “Universal Noether’s Nature of Infinitesimal Transformations in Lorentz-Covariant Field Theories”, Nuovo Cimento, **7A**, 271.

[18] Cassirer, Ernst (1953), Substance and Function and Einstein’s Theory of Relativity (Dover, n.p.).

[19] — (1956), Determinism and Indeterminism in Modern Physics (Yale U.P., New Haven).

[20] Clauser, J. F. and A. Shimony (1978), "Bell's Theorem: Experimental Tests and Implications", Rep. Prog. Phys, **41**, 1881.

[21] Coleman, Sidney (1973), "Secret Symmetry: An Introduction to Spontaneous Symmetry Breakdown and Gauge Fields", in Aspects of Symmetry by Sidney Coleman (Cambridge U.P., New York, 1985).

[22] Courant, B. and D. Hilbert (1953), Methods of Mathematical Physics, vol. I (Interscience, New York).

[23] DeWitt, Bryce S. and Neill Graham, eds. (1973), The Many Worlds Interpretation of Quantum Mechanics (Princeton U.P., Princeton).

[24] Einstein, A., B. Podolsky and N. Rosen (1935), "Can The Quantum-Mechanical Description of Physical Reality Be Considered Complete?", Phys. Rev., **47**, 777.

[25] Everett, H. (1957), "'Relative State' Formulation of Quantum Mechanics", Rev. Mod. Phys., **29**, 454.

[26] — (1973), "The Theory of the Universal Wave Function", in The Many-Worlds Interpretation of Quantum Mechanics, B. S. DeWitt and N. Graham, eds. (Princeton U.P., Princeton, 1973).

[27] Frampton, Paul H (1987), Gauge Field Theories (Benjamin/Cummings, Reading, Ma).

[28] Furry, W. H. (1936), "Note on the Quantum-Mechanical Theory of Measurement", Phys. Rev., **49**, 393.

[29] Goldstein, Herbert (1980), Classical Mechanics, 2nd ed. (Addison-Wesley, Reading, Ma).

[30] Herstein, I. N. (1975), Topics in Algebra, 2nd ed. (Wiley, New York).

[31] Howard, Don (1985), “Einstein on Locality and Separability”, Stud. Hist. Phil. Sci., **16**, 171.

[32] — (1985b), “Locality, Separability, and the Physical Implications of the Bell Experiments: A New Interpretation”, Preprint.

[33] — (n.d.) “What Makes a Classical Concept Classical? Toward a Reconstruction of Niels Bohr’s Philosophy of Physics”, Preprint.

[34] Huang, Kerson (1982), Quarks Leptons and Gauge Fields (World Scientific, Singapore).

[35] Jackiw, Roman (1972), “Field Theoretic Investigations in Current Algebra”, in Lectures on Current Algebra and Its Applications by Sam B. Freiman, Roman Jackiw and David J. Gross (Princeton U.P., Princeton, 1972).

[36] Jammer, Max (1974), The Philosophy of Quantum Mechanics (Wiley, New York).

[37] Jarrett, Jon P. (1984), “On The Physical Significance of the Locality Conditions in the Bell Arguments”, Nous, **18**, 569.

[38] Lanczos, Cornelius (1986), The Variational Principles of Mechanics (Dover, New York).

[39] Leader, Elliot and Enrico Predazzi (1982), An Introduction To Gauge Theories and the New Physics (Cambridge U.P., Cambridge).

[40] Moriyasu, K. (1982), “The Renaissance of Gauge Theory”, Contemp. Phys., **2**3, 553.

[41] O’Raifeartaigh, L. (1979), “Hidden Gauge Symmetry”, Rep. Prog. Phys., **42**, 159.

[42] Palmieri, C. and B. Vitale (1970), "On the Inversion of Noether's Theorem in the Lagrangian Formalism", Nuovo Cimento, **66a**, 299.

[43] Pathria, P. K. (1978), Statistical Mechanics (Pergamon. New York).

[44] Pena-Auerbach, L. de la (1967), "A Simple Derivation of the Schrödinger Equation from the Theory of Markov Processes", Phys. Lett., **24A**, 603.

[45] — and Leopoldo S. Garcia-Colin (1968a), "Quantum-Mechanical Description of a Brownian Particle", J. Math. Phys., **9**, 668.

[46] — and — (1968b), "Possible Interpretation of Quantum Mechanics", J. Math. Phys., **9**, 916.

[47] Quigg, Chris (1983), Gauge Theories of the Strong, Weak, and Electromagnetic Interactions (Benjamin/Cummings, Reading. Ma).

[48] Ramond, Pierre (1981), Field Theory, A Modern Primer (Benjamin /Cummings, Reading, Ma).

[49] Reif, F. (1965), Fundamentals of Statistical and Thermal Physics (McGraw-Hill, New York).

[50] Rosen, Joe (1972), "Noether's Theorem in Classical Field Theory", Ann. Phys., **69**, 349.

[51] — (1974a), "Generalized Noether's Theorem. I. Theory", Ann. Phys., **82**, 54.

[52] — (1974b), "Generalized Noether's Theorem. I. Application", Ann. Phys., **82**, 70.

[53] — and Yehudah Freundlich (1978), "Symmetry and Conservation", Am. J. Phys., **46**, 1030.

[54] — (1980), “Symmetry and Conservation: Inverse Noether’s Theorem and General Formalism”, J. Phys., **A13**, 803.

[55] — (1983), A Symmetry Primer For Scientists (Wiley, New York).

[56] Schrödinger E. (1935), “Discussion of Probability Relations between Separated Systems”, Proc. Camb. Phil. Soc., **31**, 555.

[57] Shubnikov, A. V. and V. A. Koptsik (1974), Symmetry in Science and Art (Plenum, New York).

[58] Stapp, Henry Pierce (1972), “The Copenhagen Interpretation”, Am. J. Phys., **40**, 1098.

[59] — (1973), “S-Matrix Interpretation of Quantum Theory”, Phys. Rev., **D8**, 1303.

[60] von Neumann, J. (1955), Mathematical Foundations of Quantum Mechanics (Princeton U.P., Princeton).

[61] Wang and Uhlenbeck (1945), “On the Theory of the Brownian Motion II”, Rev. Mod. Phys., **17**, 323, reprinted in Noise and Stochastic Processes, N. Wax, ed. (Dover, New York, 1954).

[62] Weyl, Hermann (1952), Symmetry (Princeton U.P., Princeton).

[63] Williams, Gareth (1973), “A Discussion of Causality and the Lorentz Group”, Int. J. Th. Phys., **1**, 415.

[64] Yang, C. N. and R. C. Mills (1954), “Conservation of Isotopic Spin and Isotopic Gauge Invariance”, Phys. Rev., **96**, 191.

[65] Zeeman, E. C. (1964), “Causality Implies the Lorentz Group”, J. Math. Phys., **5**, 490.